Interactive Model Learning from High-Dimensional Data: A Visual Analytics Approach

Klaus Mueller

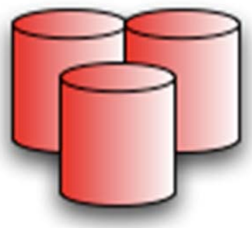
Computer Science

Lab for Visual Analytics and Imaging (VAI)

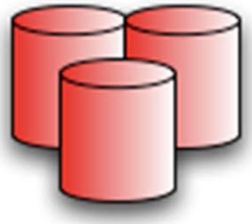
Stony Brook University

Visual Analytics

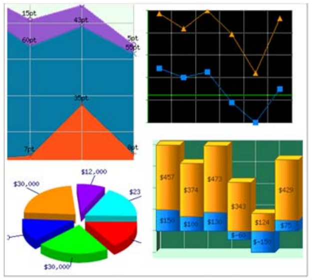
Visual Analytics (Layman's View)

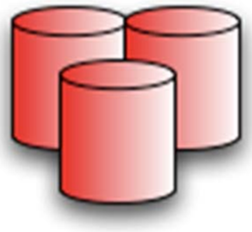


Visual Analytics (Layman's View)

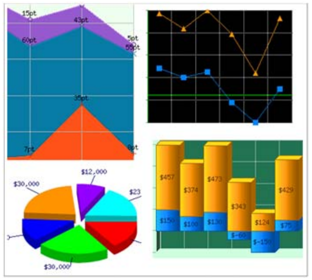


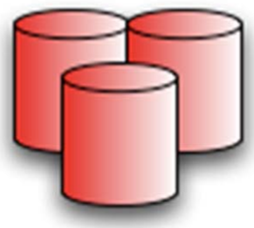
Visual Analytics (Layman's View)



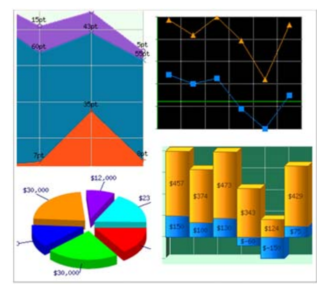


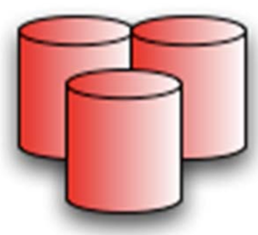
Visual Analytics (Layman's View)



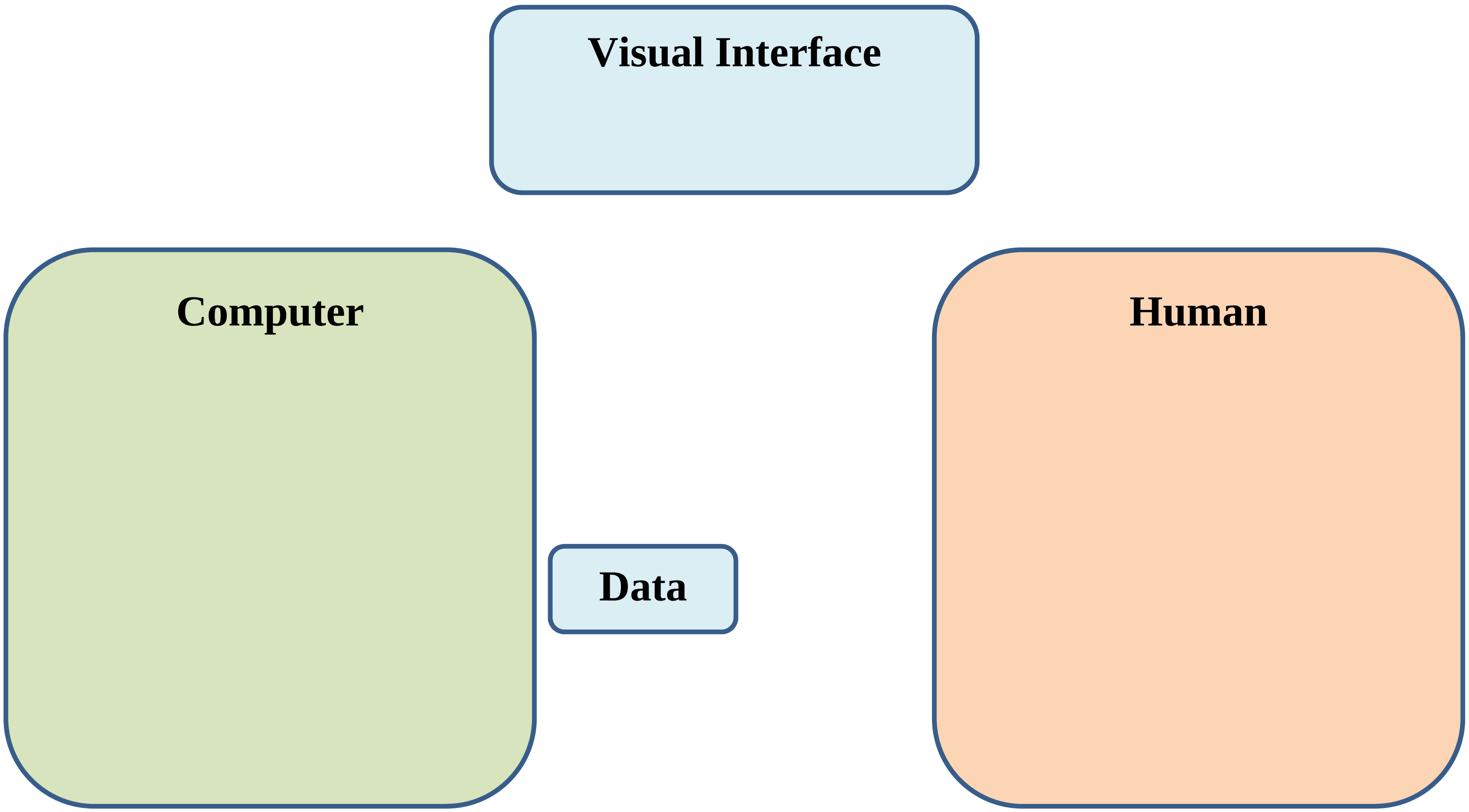


Visual Analytics (Layman's View)

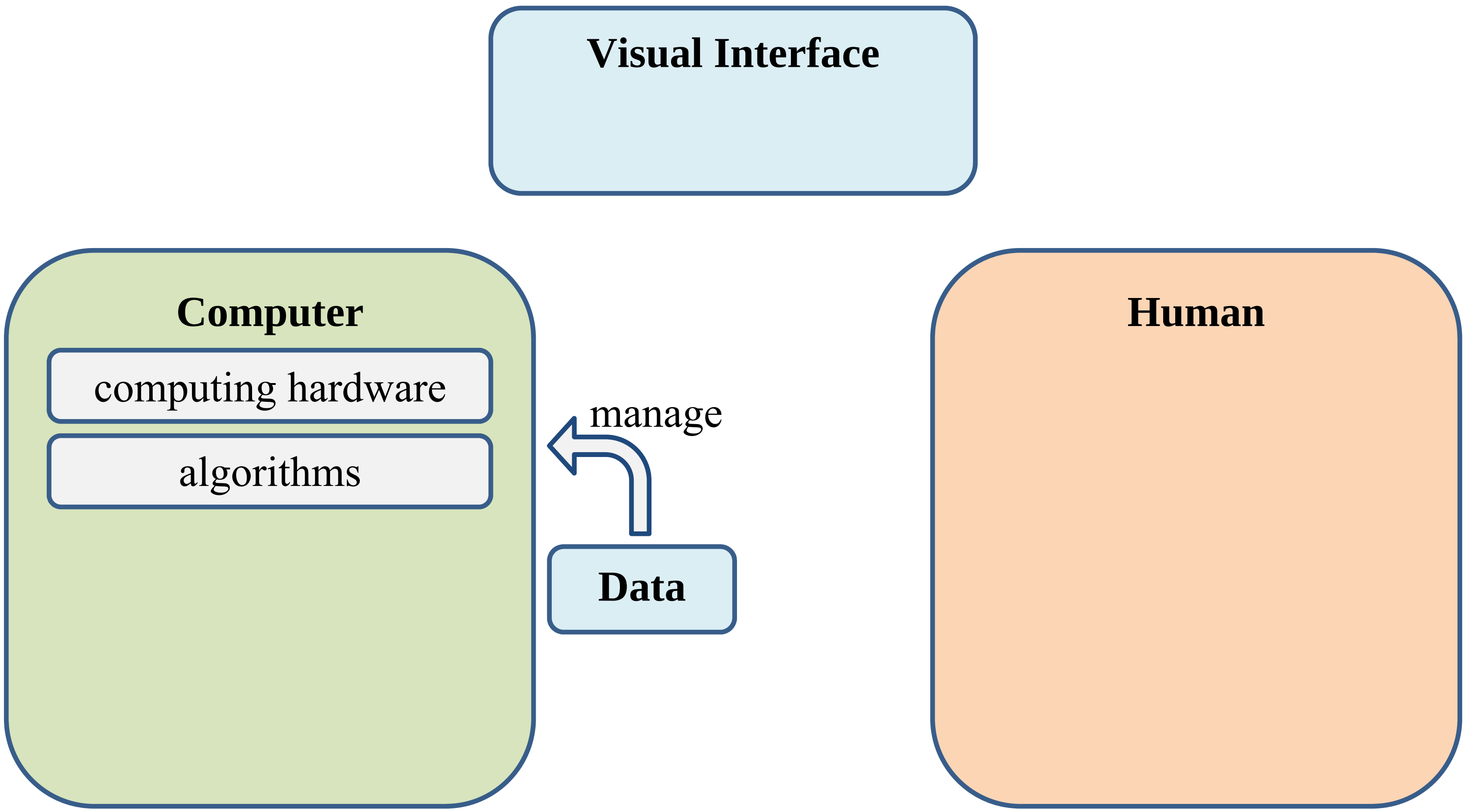




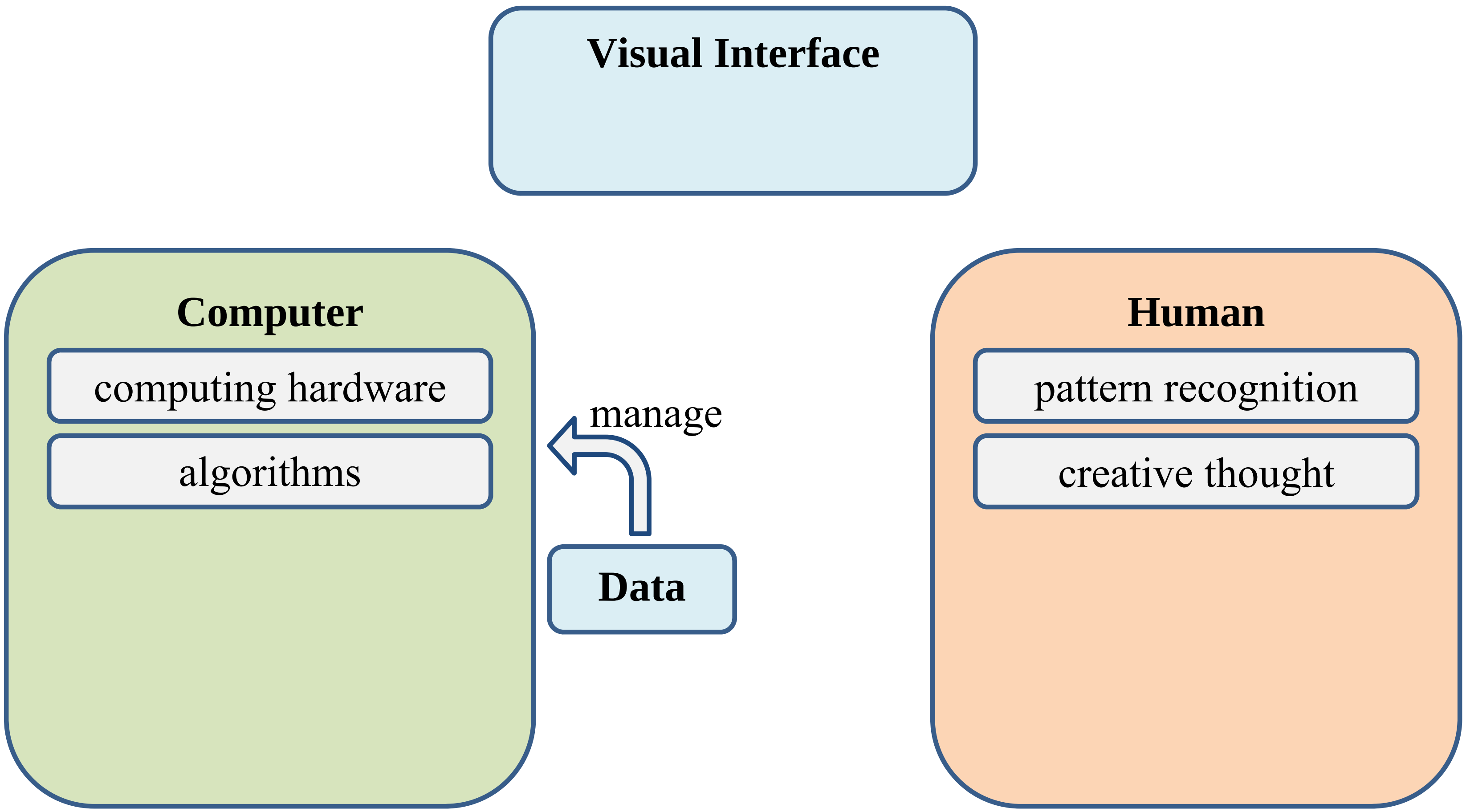
Visual Analytics (Expert View)



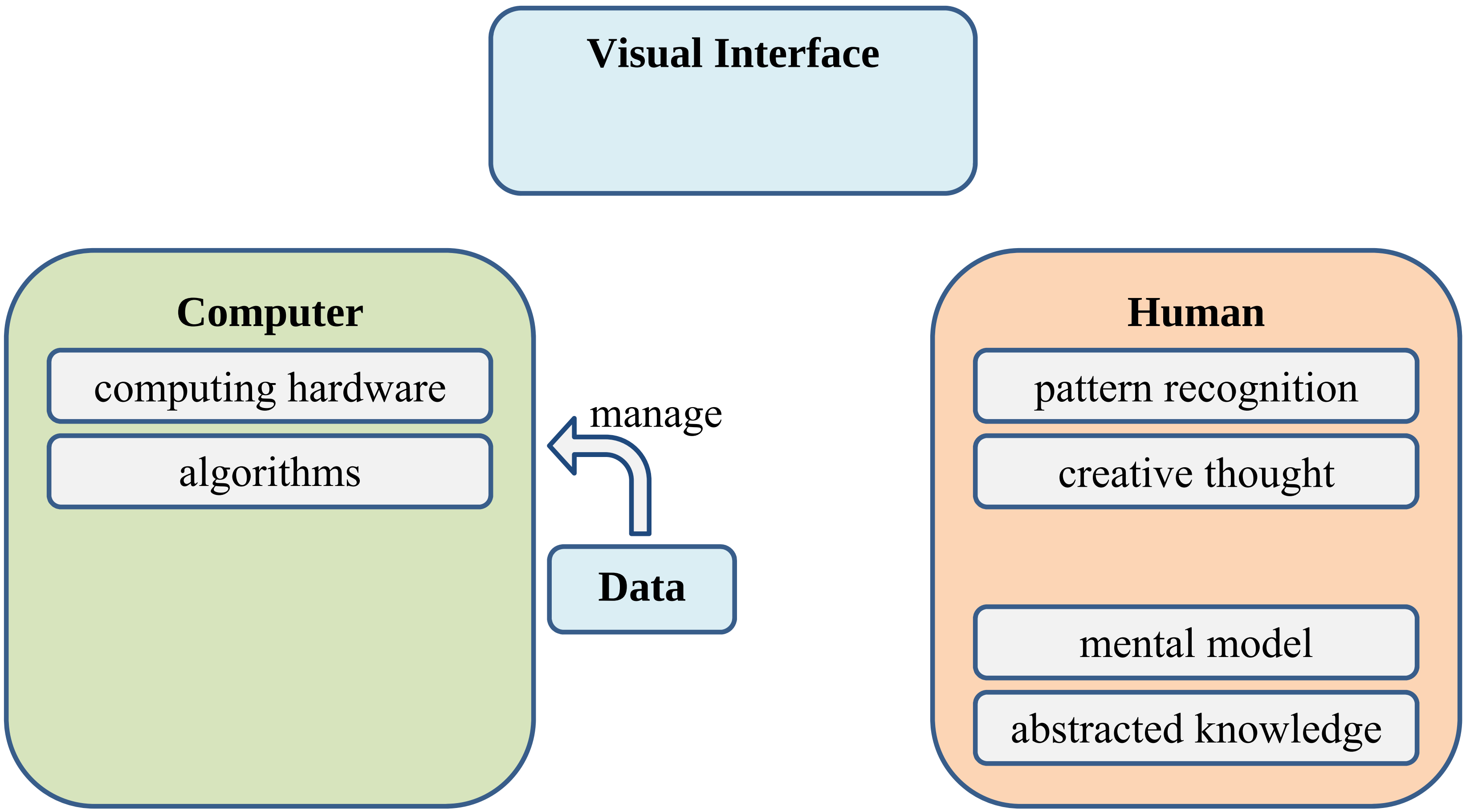
Visual Analytics (Expert View)



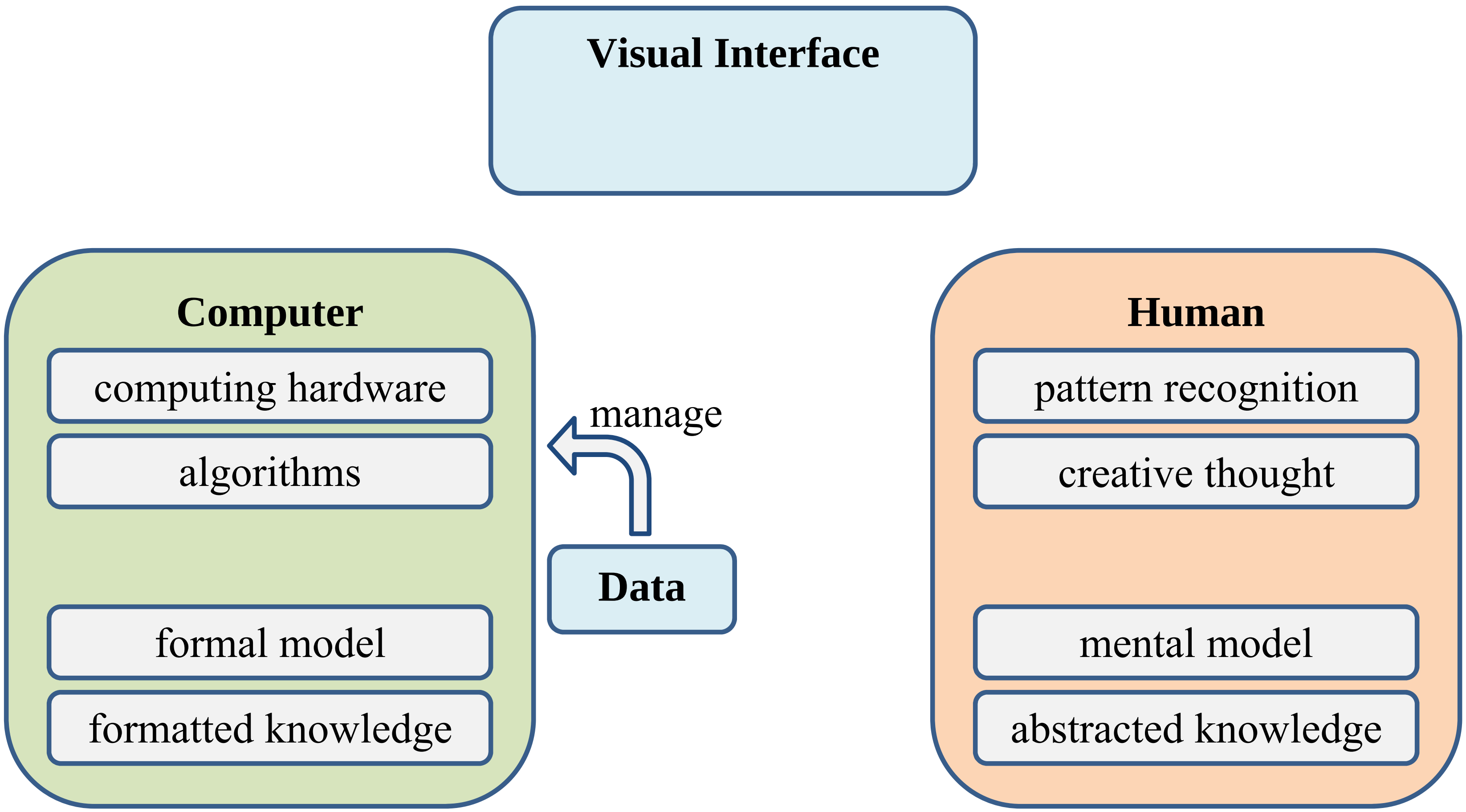
Visual Analytics (Expert View)



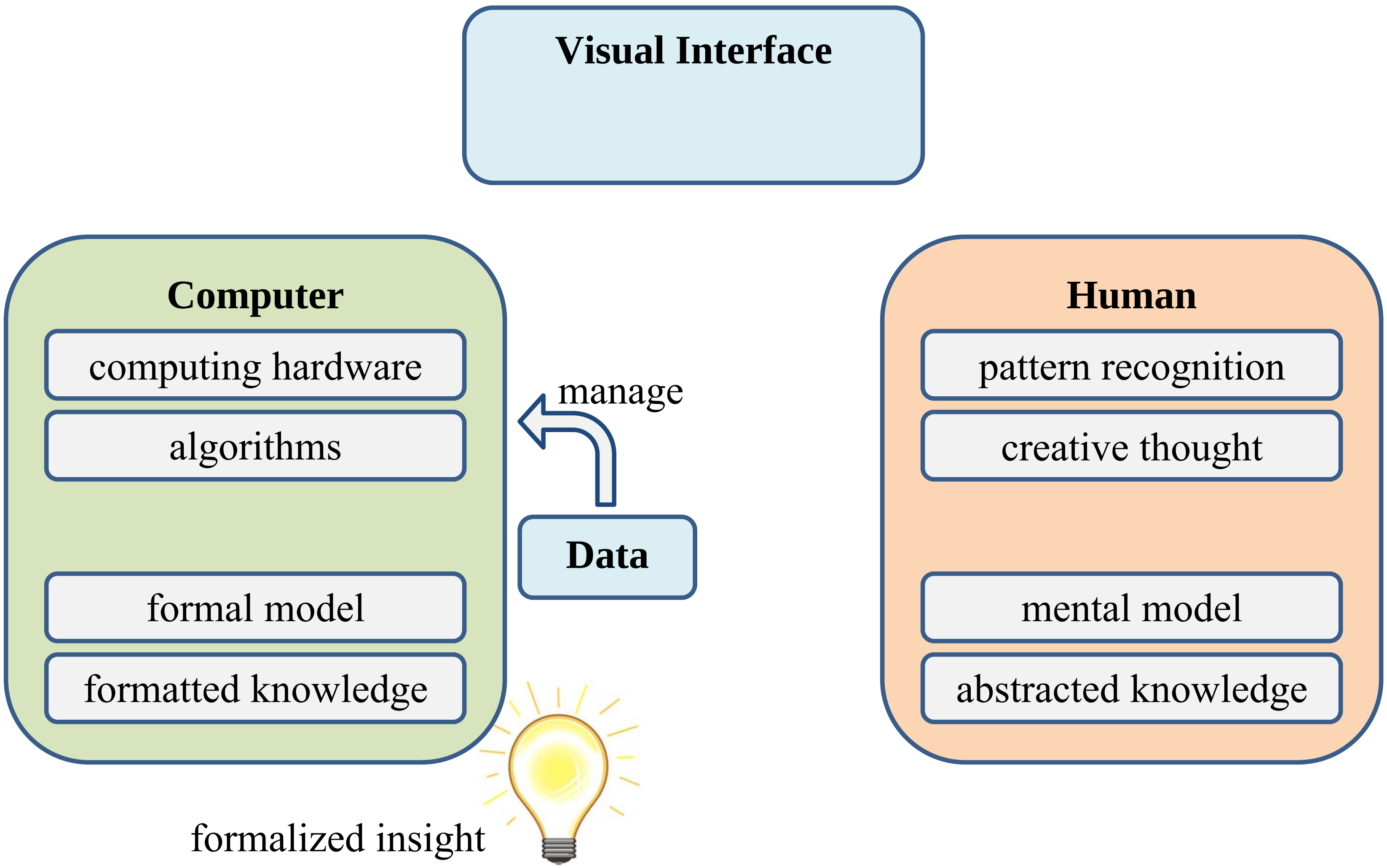
Visual Analytics (Expert View)



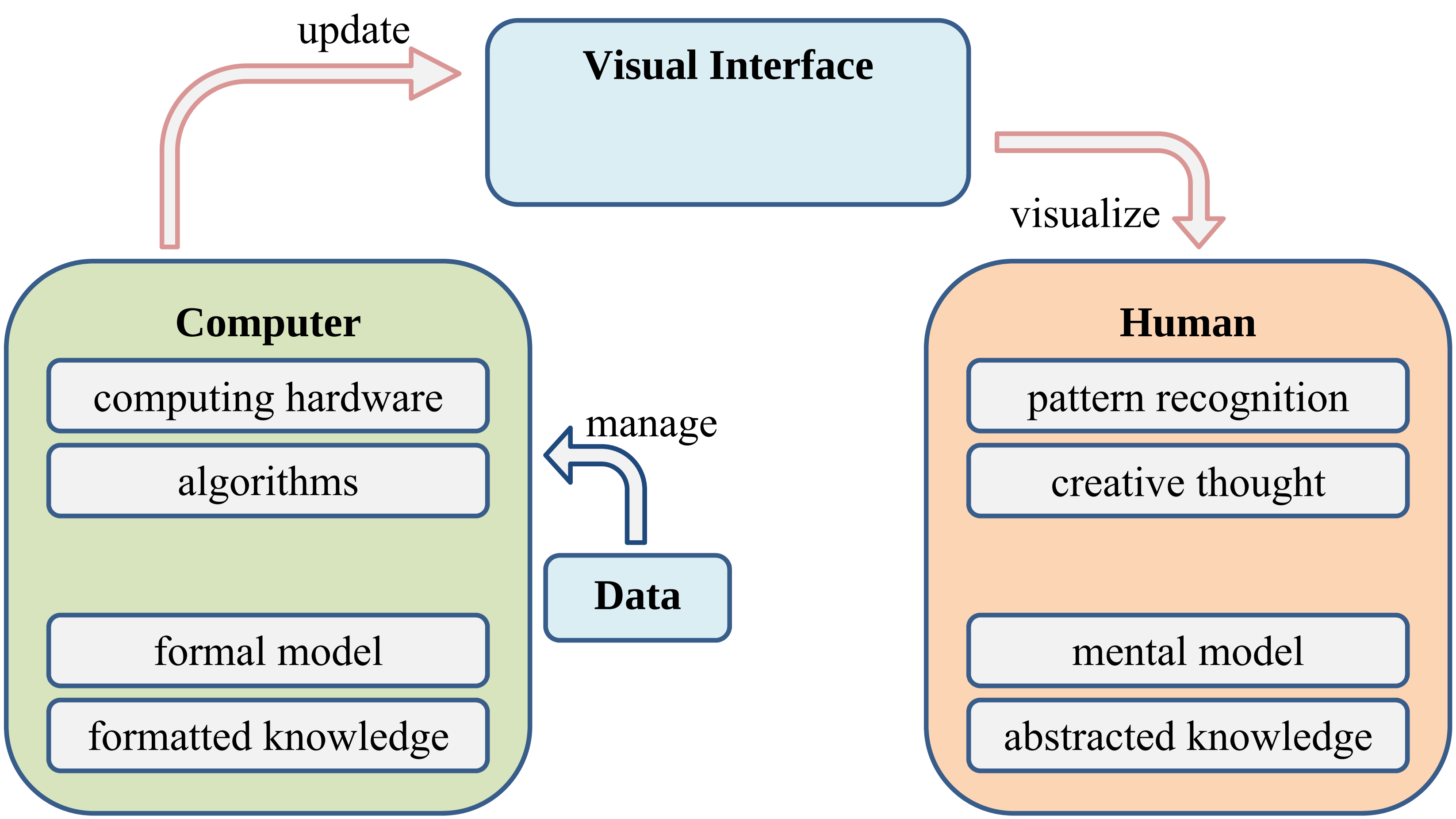
Visual Analytics (Expert View)



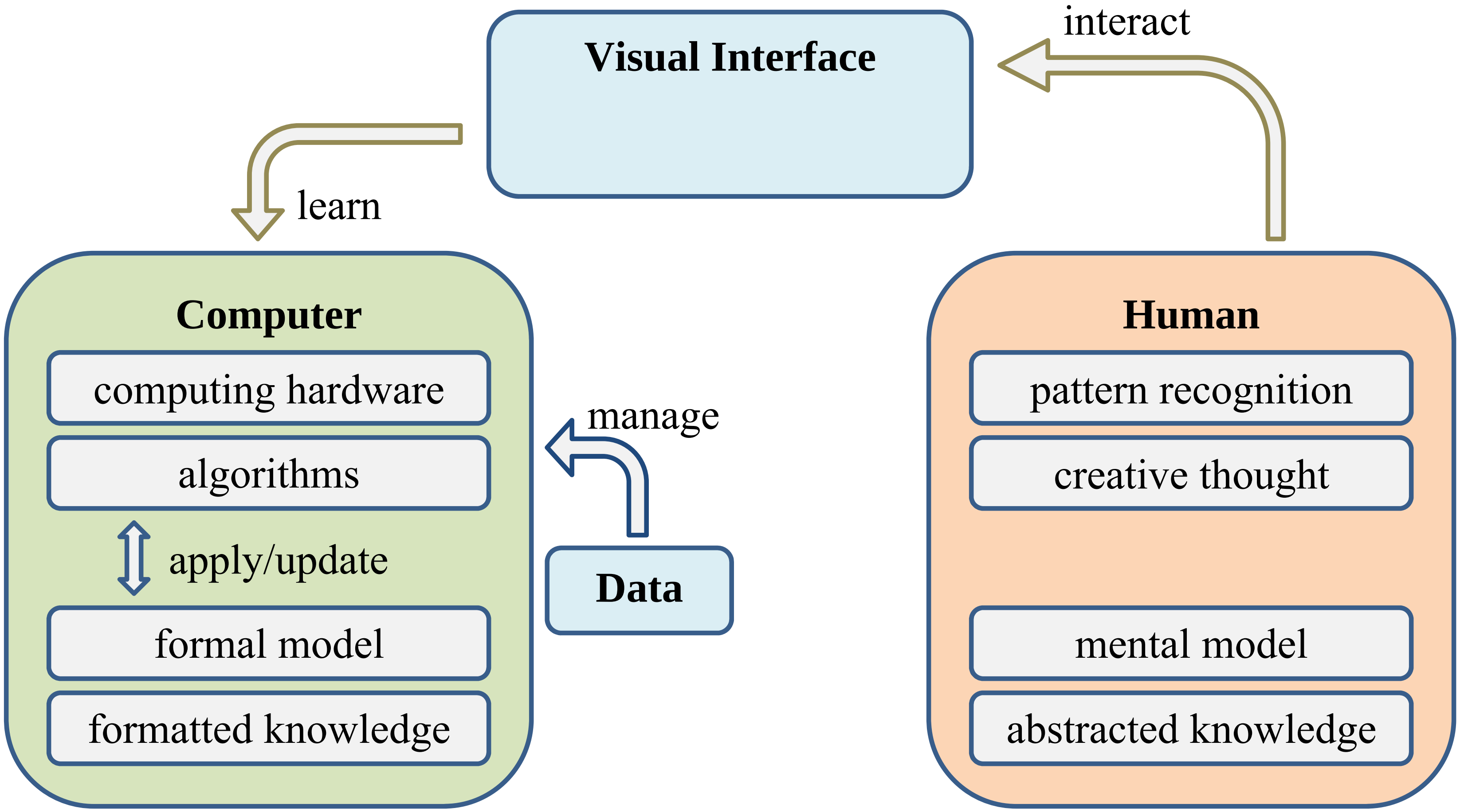
Visual Analytics (Expert View)



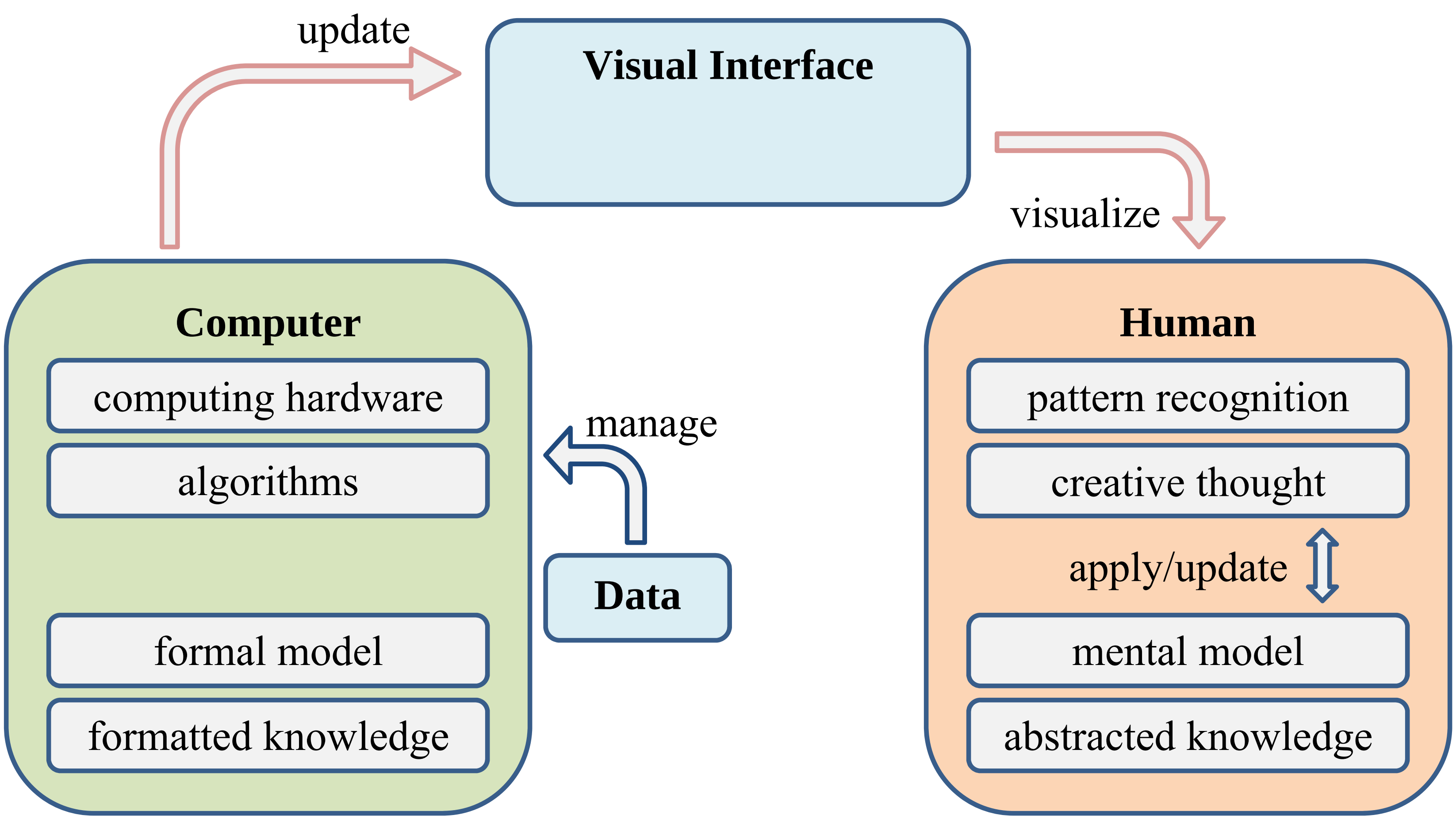
Visual Analytics (Expert View)



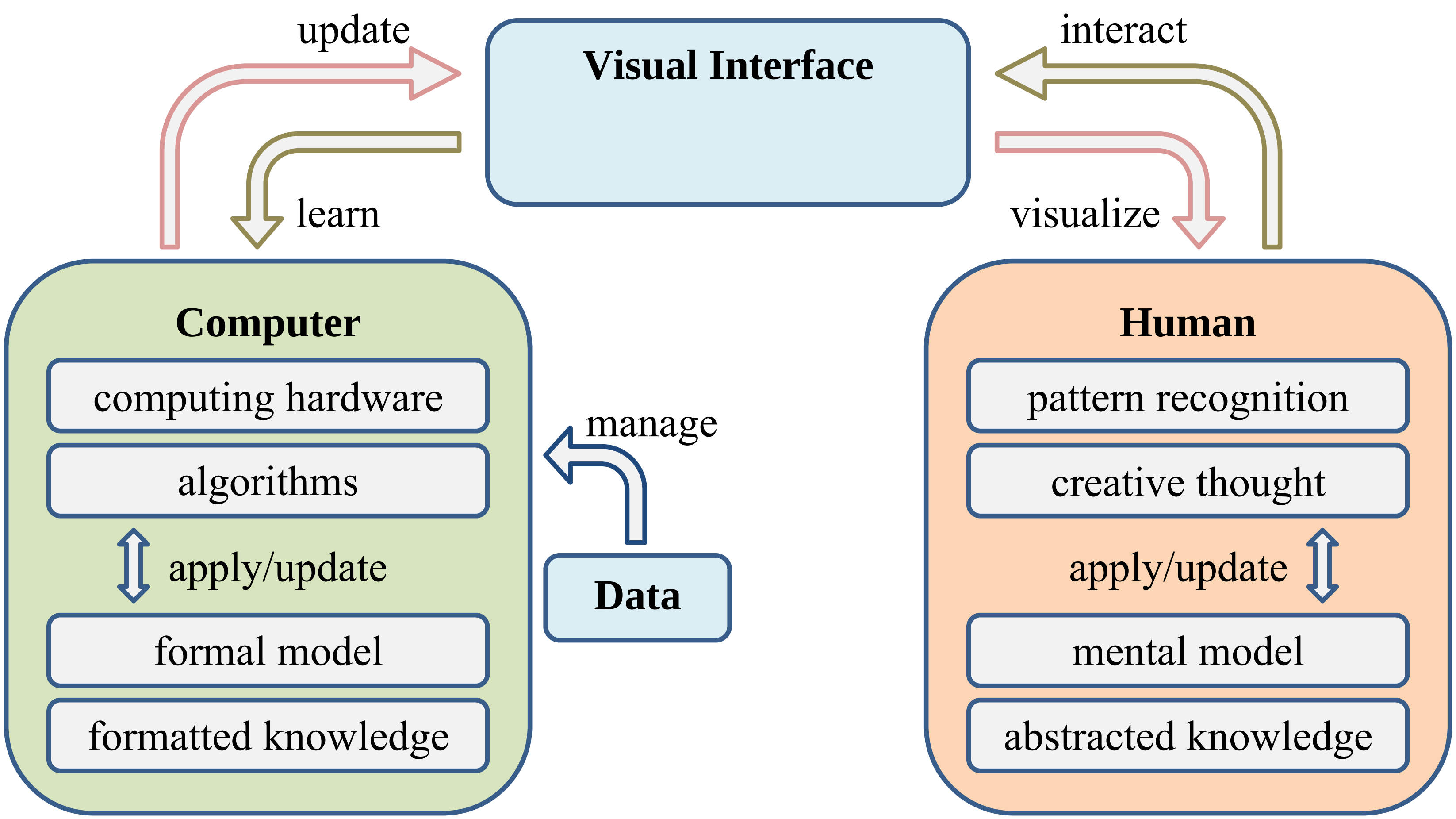
Visual Analytics (Expert View)



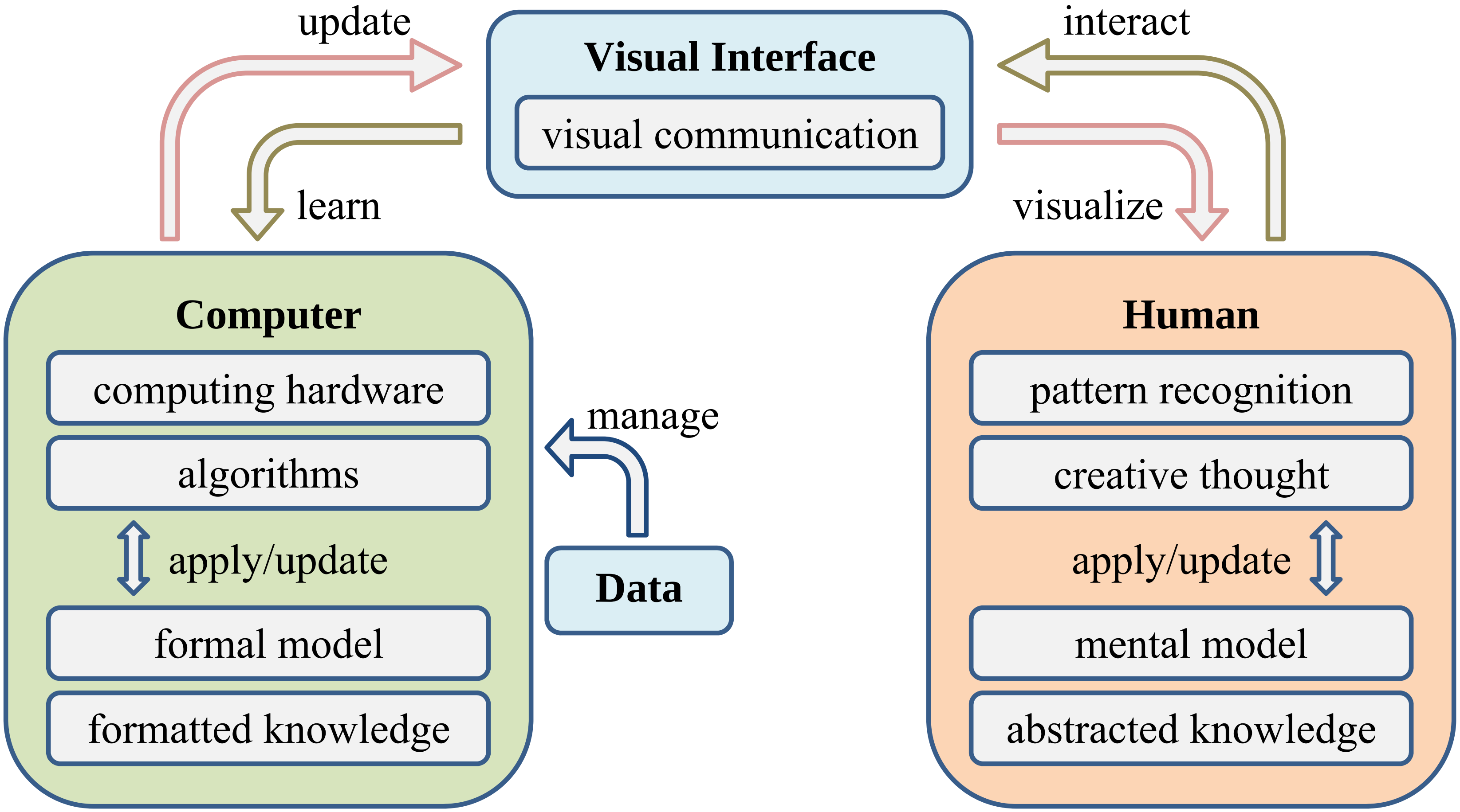
Visual Analytics (Expert View)



Visual Analytics (Expert View)



Visual Analytics (Expert View)



Mueller, et al. IEEE CG&A, 2011

Visual Communication

Obviously, the better a communicator the computer is, the better the learnt model

- computer communicates its current model via visualizations
- analyst critiques it via visual interactions
- computer learns a better model
- and so on…

Visual Communication

Obviously, the better a communicator the computer is, the better the learnt model

- computer communicates its current model via visualizations
- analyst critiques it via visual interactions
- computer learns a better model
- and so on…

A key question is thus:

- can computers master the art of communication?

Visual Communication

Obviously, the better a communicator the computer is, the better the learnt model

- computer communicates its current model via visualizations
- analyst critiques it via visual interactions
- computer learns a better model
- and so on…

A key question is thus:

- can computers master the art of communication?

Good visual design and interaction is important

Mueller, et al. IEEE CG&A, 2011

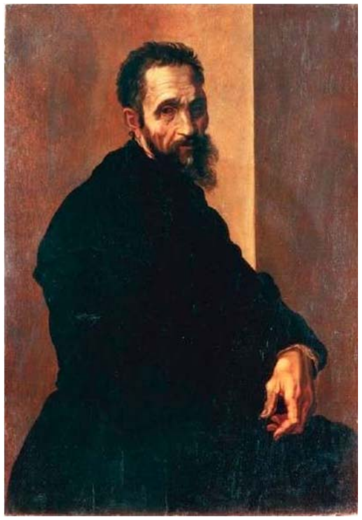
Visual Model Sculpting

Some motivating quotes from Michelangelo:

I saw the angel in the marble and carved until I set him free.

Every block of stone has a statue inside it and it is the task of the sculptor to discover it.

The marble not yet carved can hold the form of every thought the greatest artist has.



Visual Model Sculpting

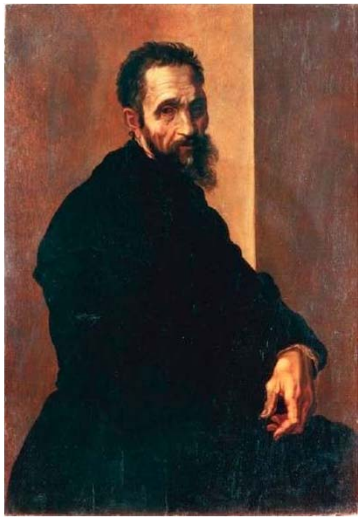
Some motivating quotes from Michelangelo:

I saw the angel in the marble and carved until I set him free.

Every block of stone has a statue inside it and it is the task of the sculptor to discover it.

The marble not yet carved can hold the form of every thought the greatest artist has.

Exchange 'angel' or 'statue' by 'model' and you can be the Michelangelo of Visual Analytics ☺



Differences

Michelangelo's 'data' were 3-D blocks of marble

- ours are N-D blocks of bytes

Michelangelo's tools were chisels, etc.

- ours are mouse, multi-touch devices, etc

Michelangelo would say things like this:

- "It is well with me only when I have a chisel in my hand. "

High-D Visualization

Problems

- comprehensive high-D visualizations can be very confusing
- need to make high-D visualization user friendly and intuitive

High-D Visualization

Problems

- comprehensive high-D visualizations can be very confusing
- need to make high-D visualization user friendly and intuitive

Key elements towards these goals

- interactive: allow users to playfully sculpt the knowledge
- communicative: let the data tell their story
- illustrative: abstract away irrelevant detail
- grounded: maintain a reference to native data space

High-D Visualization

Problems

- comprehensive high-D visualizations can be very confusing
- need to make high-D visualization user friendly and intuitive

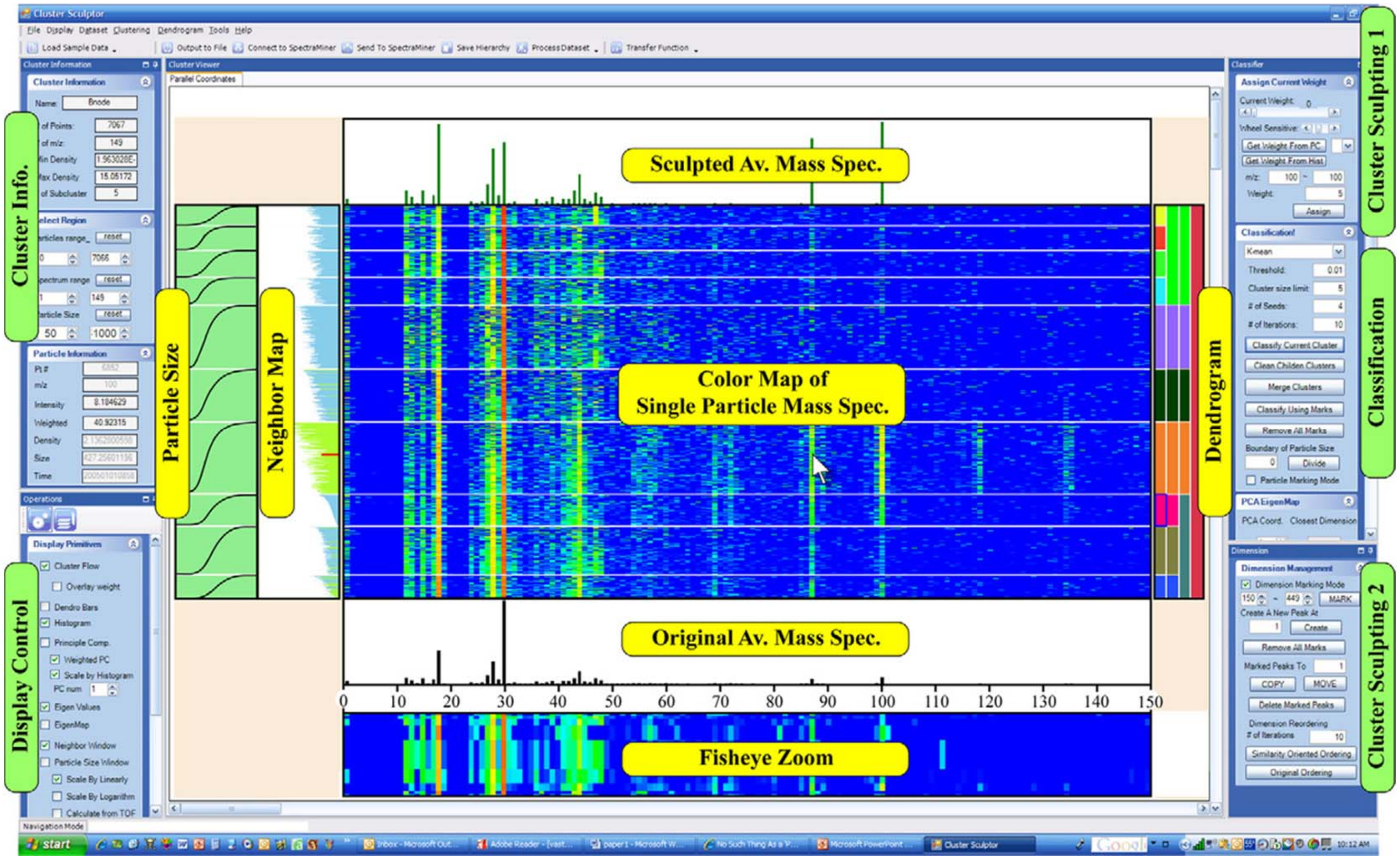
Key elements towards these goals

- interactive: allow users to playfully sculpt the knowledge
- communicative: let the data tell their story
- illustrative: abstract away irrelevant detail
- grounded: maintain a reference to native data space

Four (somewhat) complementary paradigms

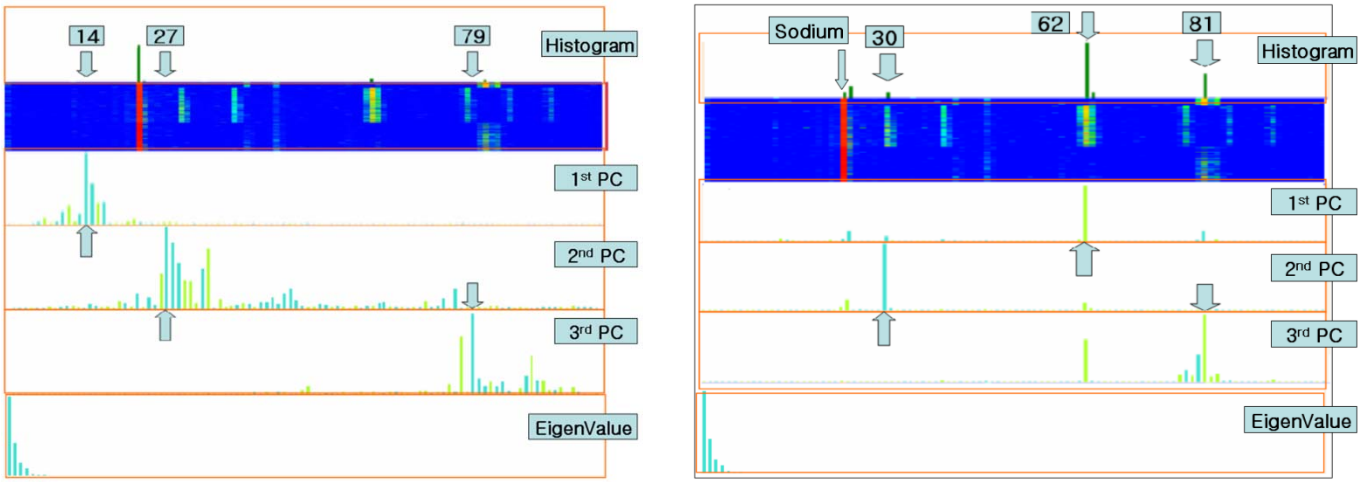
- spectral plots → see high-D hierarchies
- *dynamic* scatterplots → see high-D shapes
- parallel coordinates → see high-D cause + effect
- space embeddings → see high-D relationships

Spectral Plots (SpectrumMiner)



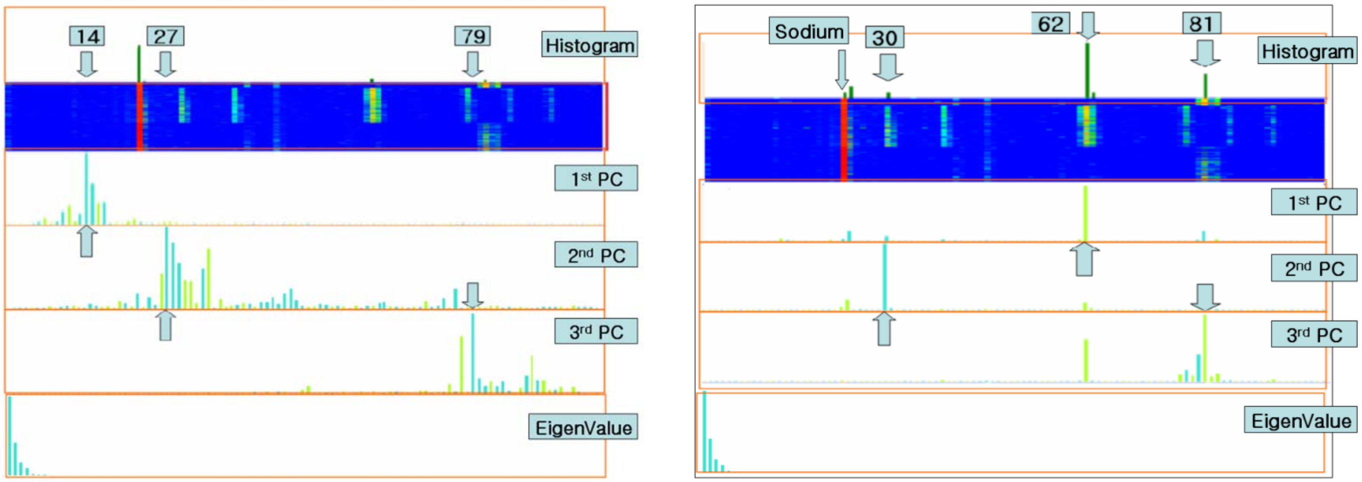
shown: 7076 particles of 450-D mass spectra acquired with single particle mass spectrometer (SPLAT)

N-D Sculpting w/SpectrumMiner

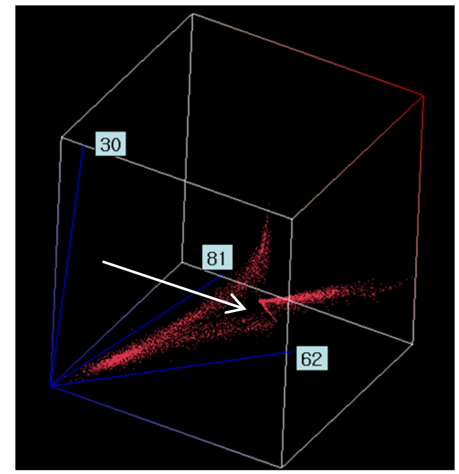


reducing the effect of sodium (set weight = 0.1)

N-D Sculpting w/SpectrumMiner



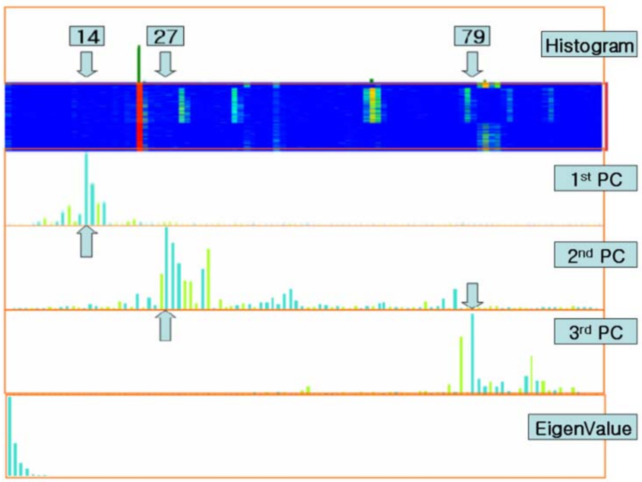
reducing the effect of sodium (set weight = 0.1)

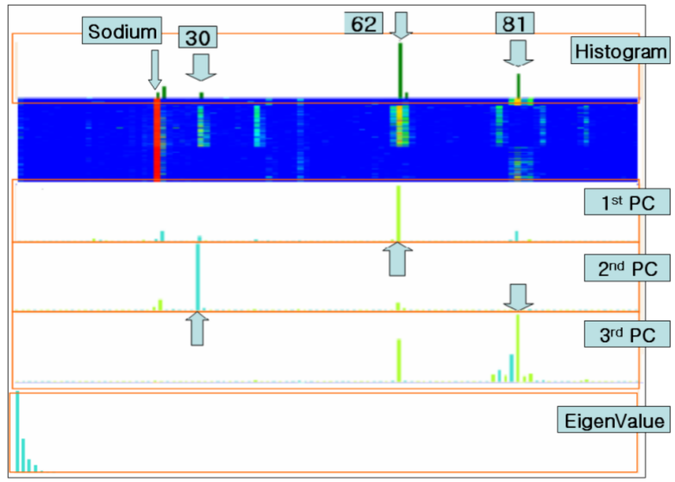


3D PCA view

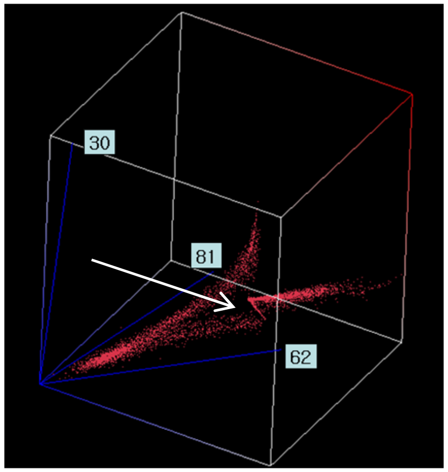
Garg, Nam, Ramakrishnan, Mueller, IEEE VAST 2008

N-D Sculpting w/SpectrumMiner

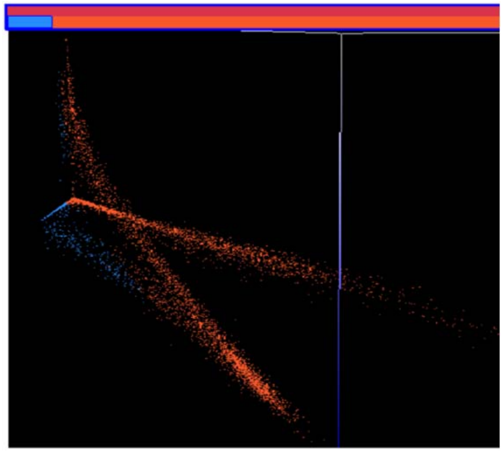




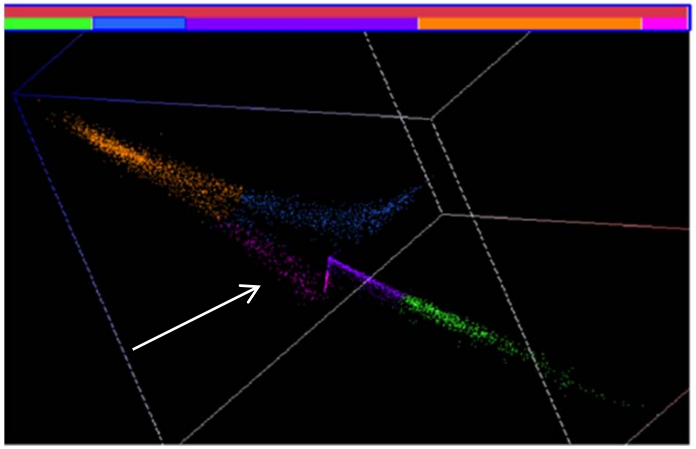
reducing the effect of sodium (set weight = 0.1)



3D PCA view

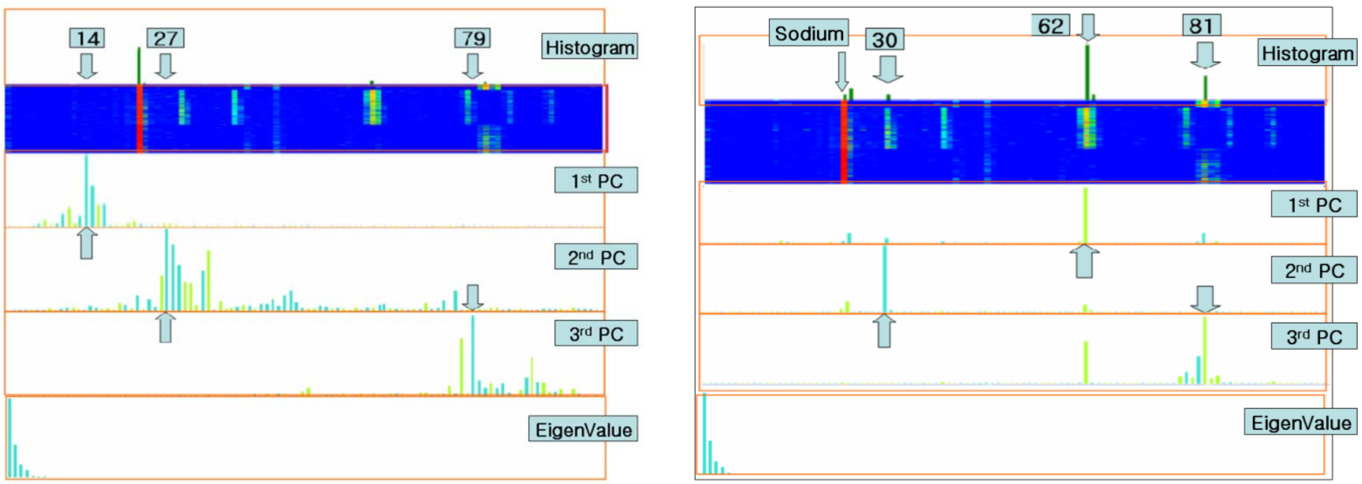


automated k-means

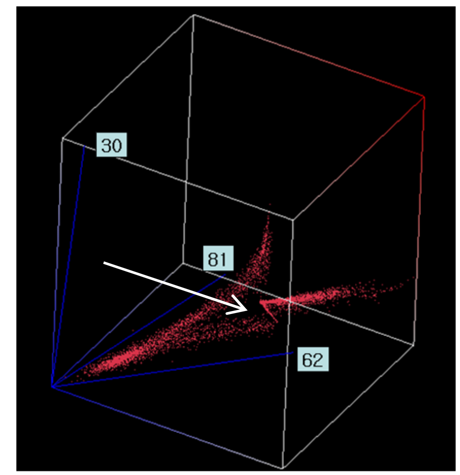


user chooses k=5

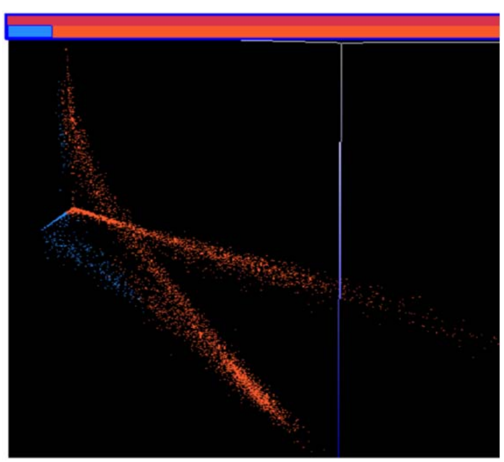
N-D Sculpting w/SpectrumMiner



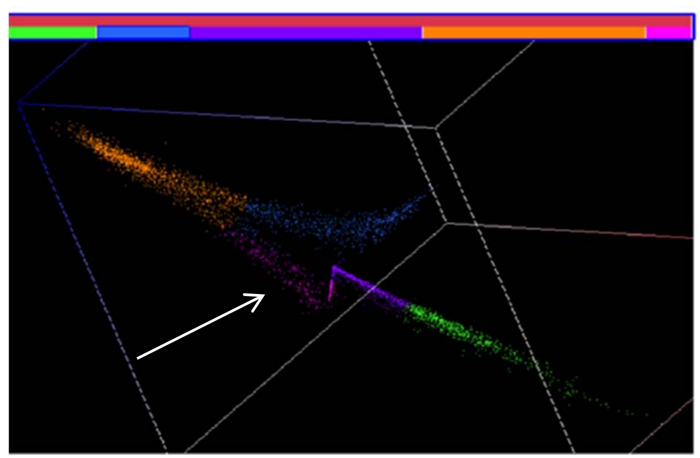
reducing the effect of sodium (set weight = 0.1)



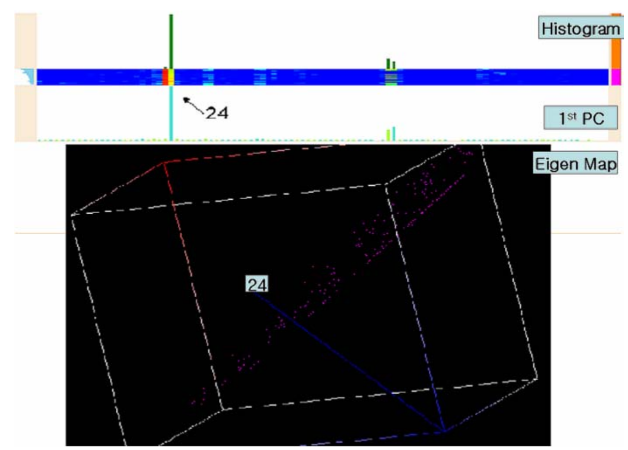
3D PCA view



automated k-means

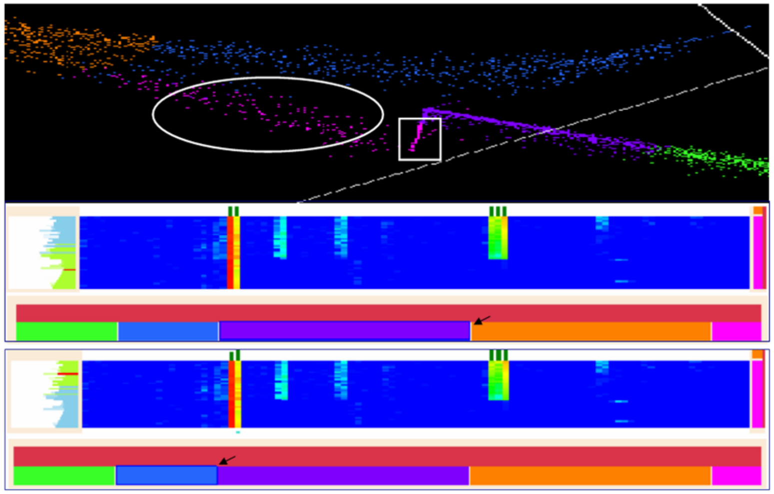


user chooses k=5



inspect more closely

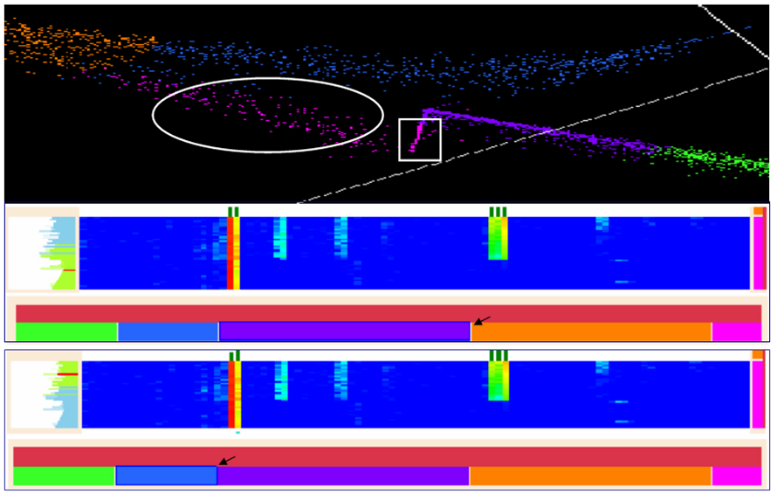
N-D Sculpting w/SpectrumMiner



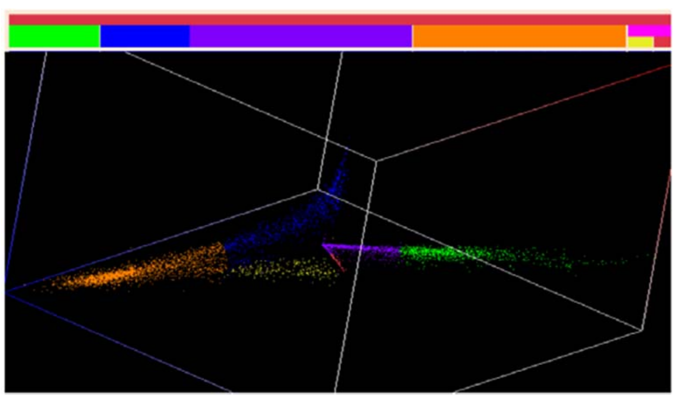
show dimension interactions in neighborhood map

Nam, Zelenyuk, Imre, Mueller, IEEE VAST 2007

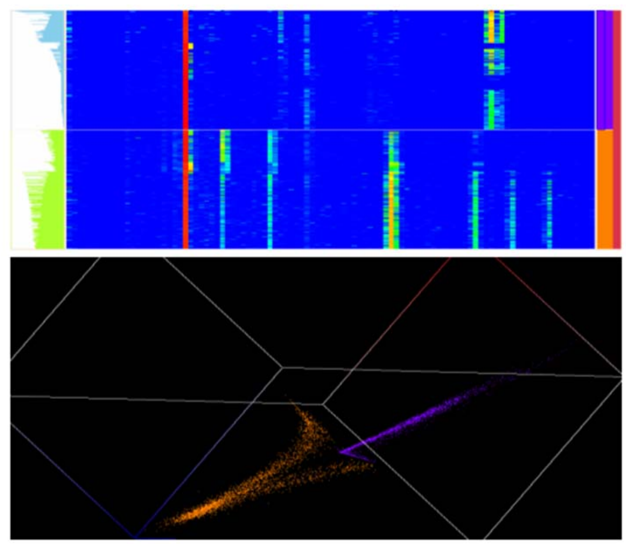
N-D Sculpting w/SpectrumMiner



show dimension interactions in neighborhood map

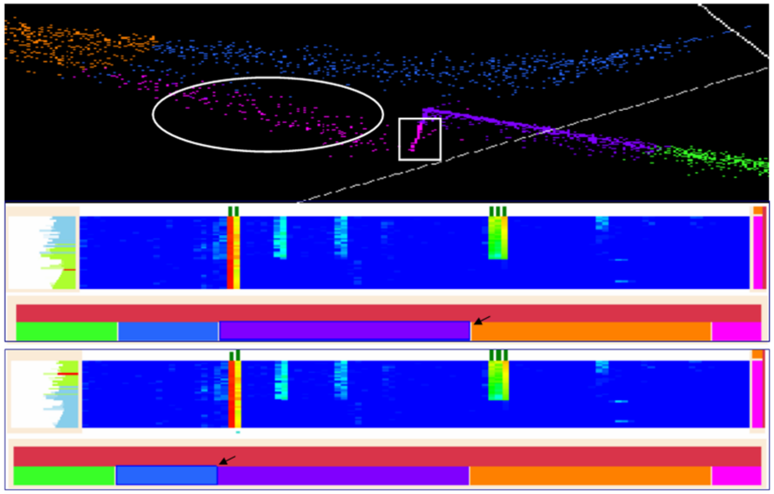


before merge

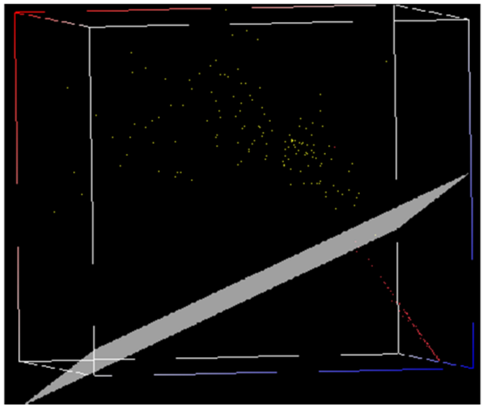


after merge

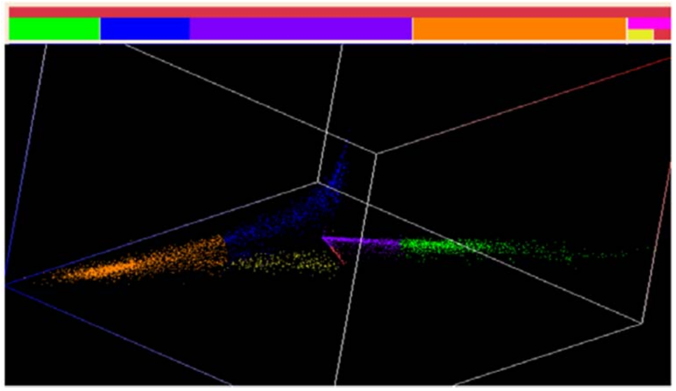
N-D Sculpting w/SpectrumMiner



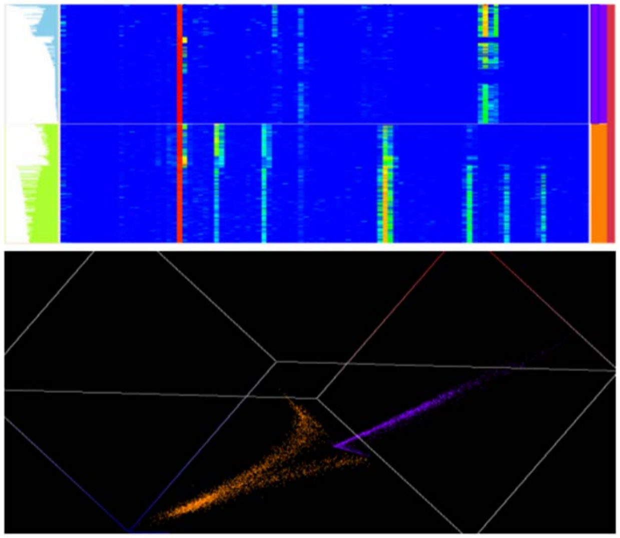
show dimension interactions in neighborhood map



Support Vector Machine (SVM) Model encodes this knowledge



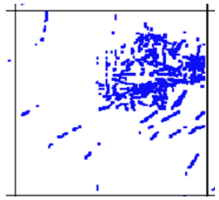
before merge



after merge

Scatterplots

Familiar for the display of bi-variate relationships

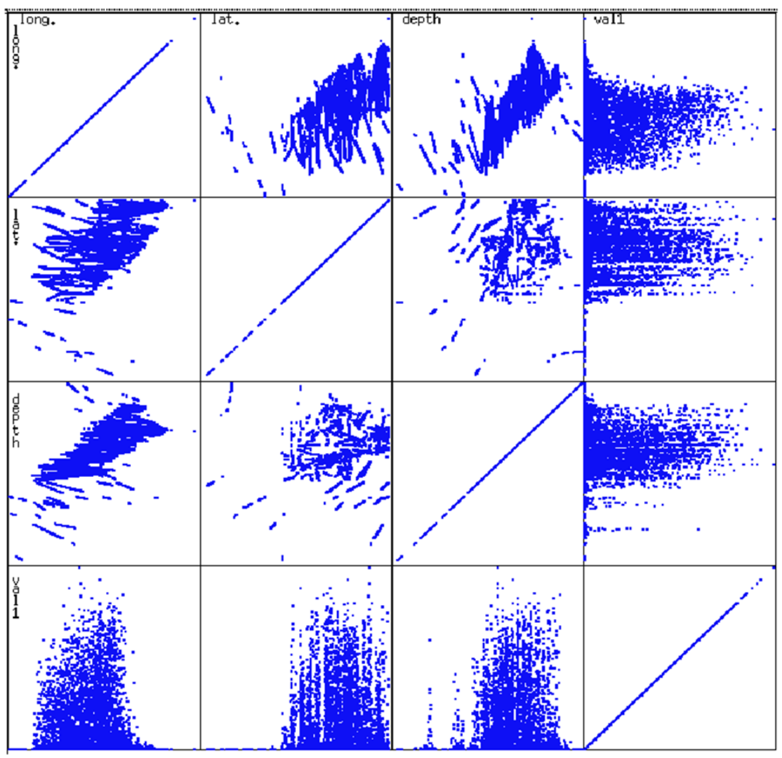


Scatterplots

Familiar for the display of bi-variate relationships

Multivariate relationships arranged in scatterplot matrices

- not overly intuitive to perceive multivariate relationships



Dynamic Scatterplots

Interaction to help 'see' N-D

- user interface is key → N-D Navigator™

Dynamic Scatterplots

Interaction to help 'see' N-D

- user interface is key → N-D Navigator™

Motion parallax beats stereo for 3D shape perception

- the same is true for N-D shape perception
- help perception by illustrative motion blur

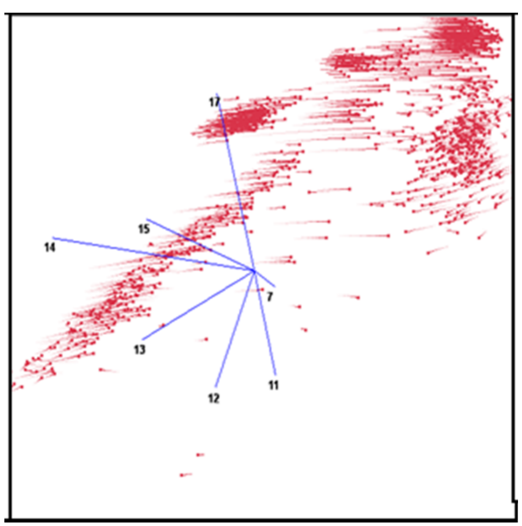
Dynamic Scatterplots

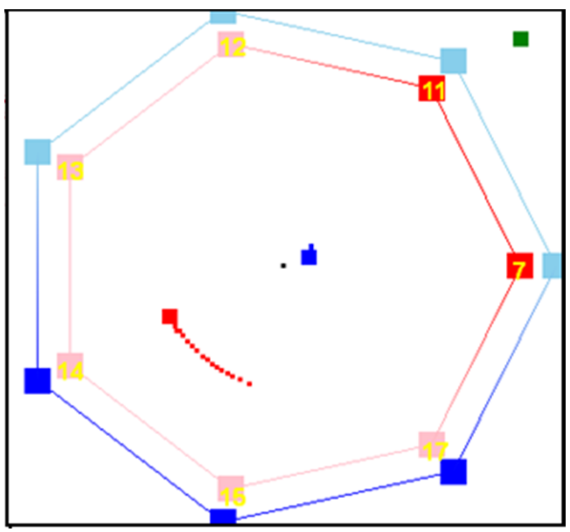
Interaction to help 'see' N-D

- user interface is key → N-D Navigator™

Motion parallax beats stereo for 3D shape perception

- the same is true for N-D shape perception
- help perception by illustrative motion blur





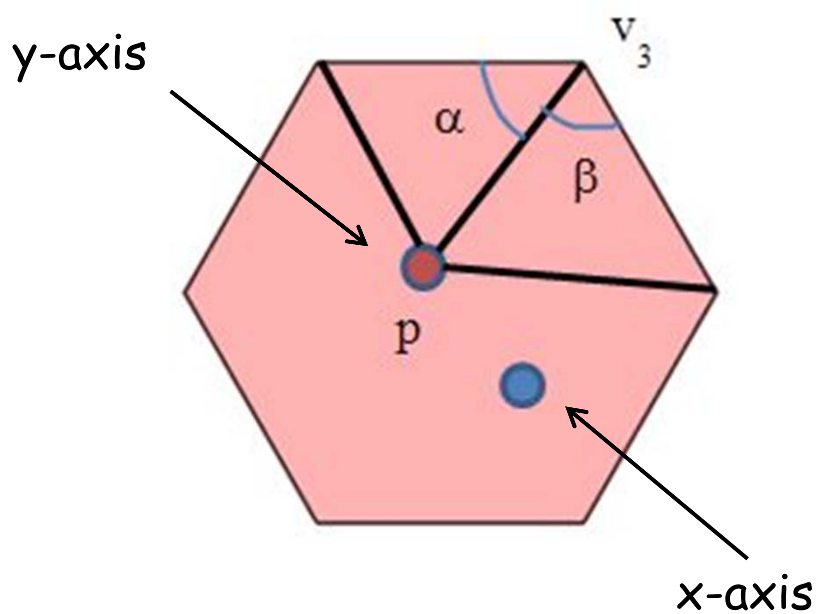
Dynamic Scatterplots

Elemental component is the polygonal touchpad

- allows navigation of projection plane in N-D space
- get axis vectors using generalized barycentric interpolation

$$\overrightarrow{w_3} = \frac{\cot(\alpha) + \cot(\beta)}{\| \vec{p} - \overrightarrow{v_3} \|^2}$$

$$p = \sum_{i=1}^{N} a_i v_i \quad \text{where } a_i = w_i \Big/ \sum_{k=1}^{N} w_k$$



Garg, Nam, Ramakrishnan, Mueller, IEEE VAST 2008

Application: Cluster Analysis

Step 1:

- dimension reduction using subspace clustering

Step 2:

- visit each subspace
- initialize projective view using projection pursuit
- set up touchpad

Step 3:

- lift-off…

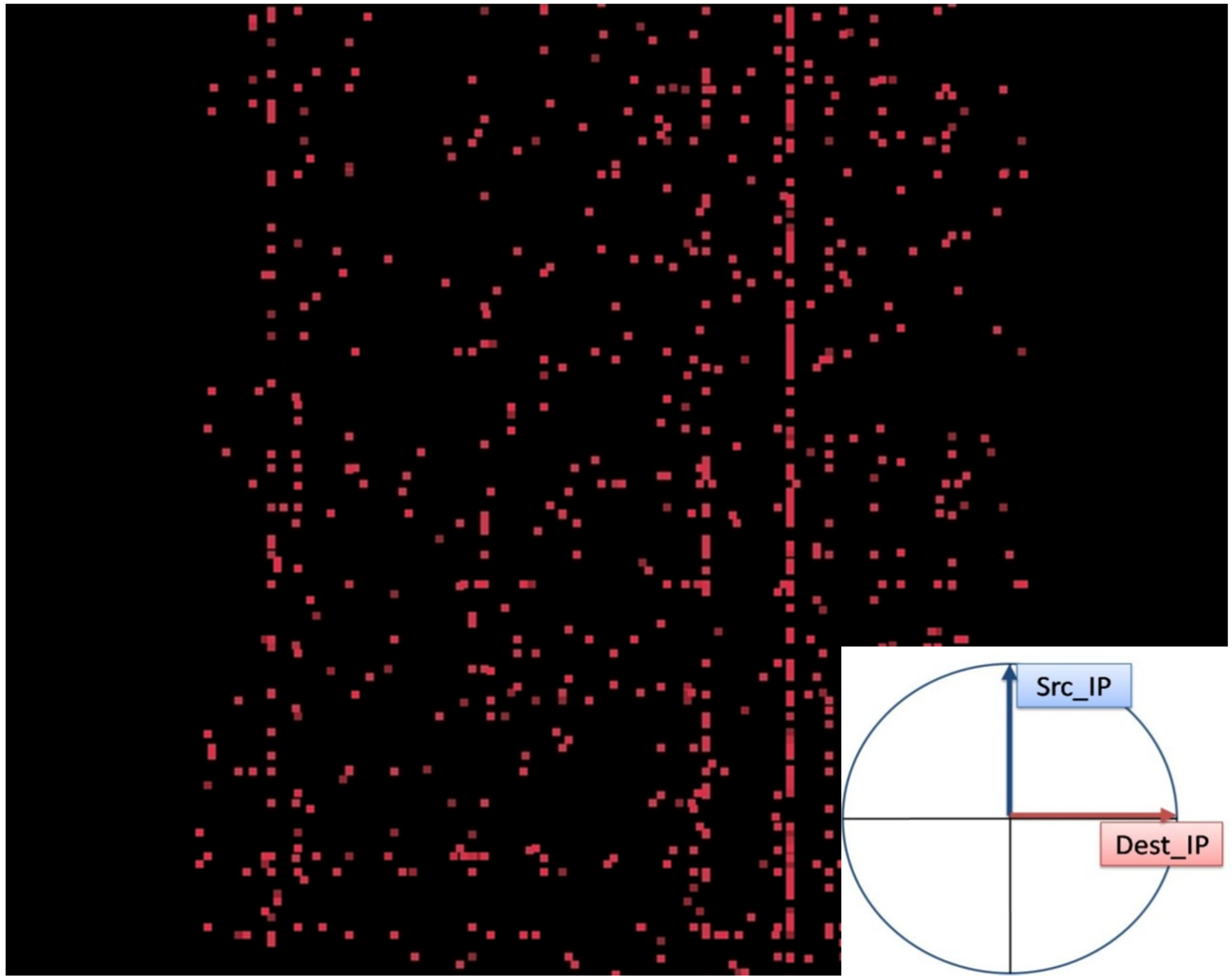
Nam, Mueller, (submitted) IEEE TVCG, 2010

Video

Locating Interesting Patterns - Dynamic Display

Initial view

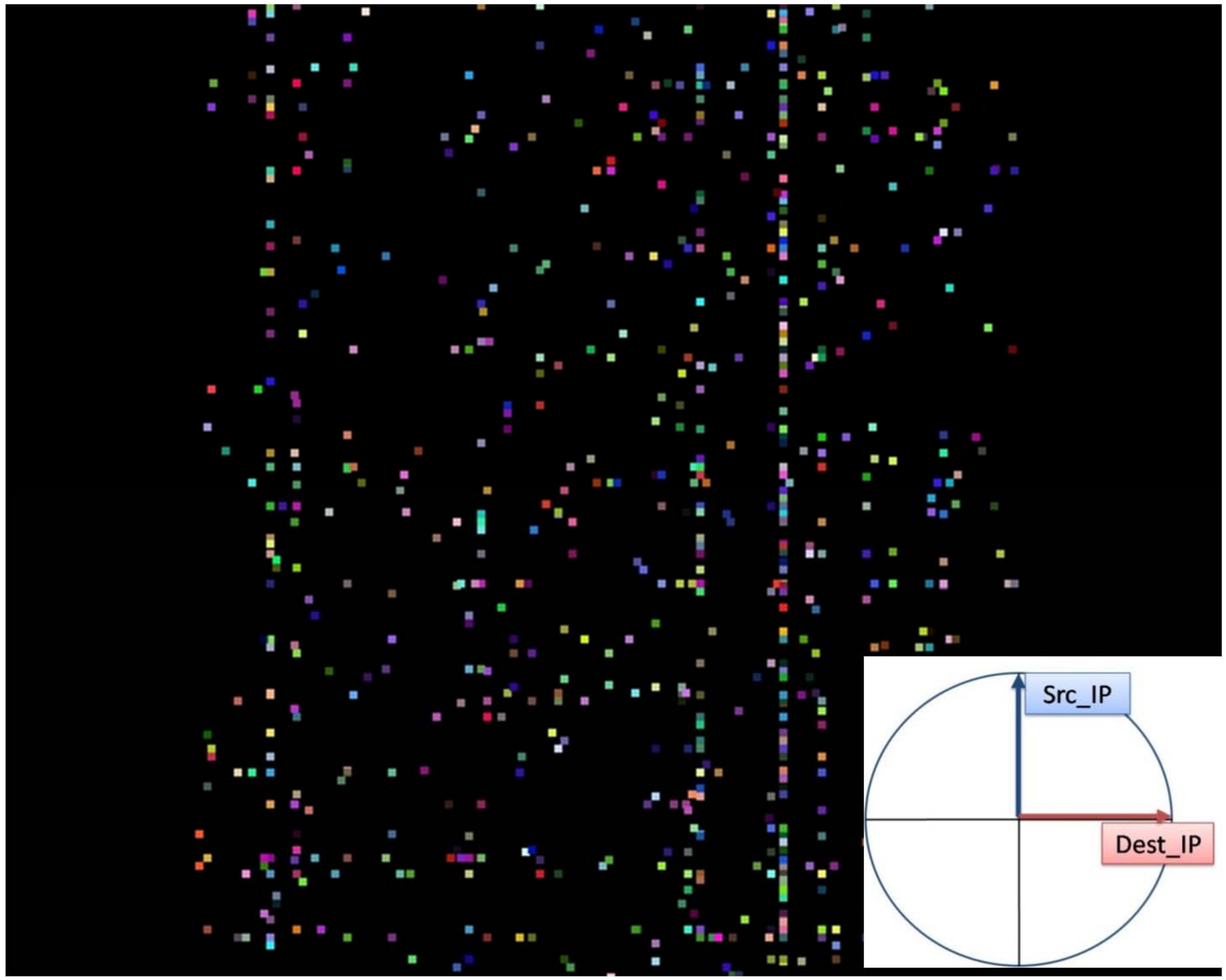
All packets have source port 80.



Garg, Nam, Ramakrishnan, Mueller, VAST 2008

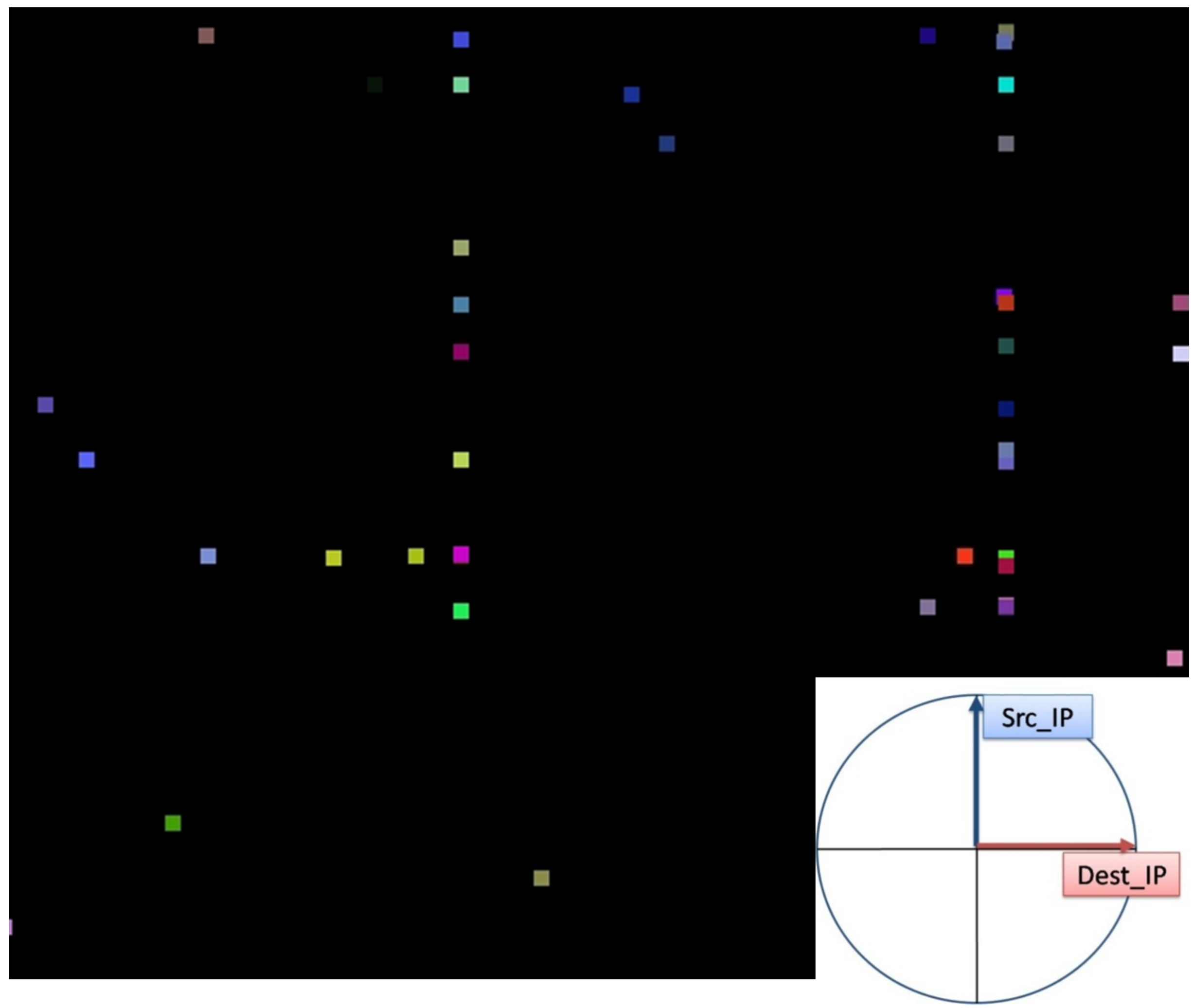
Locating Interesting Patterns - Dynamic Display

Random Coloring



Locating Interesting Patterns - Dynamic Display

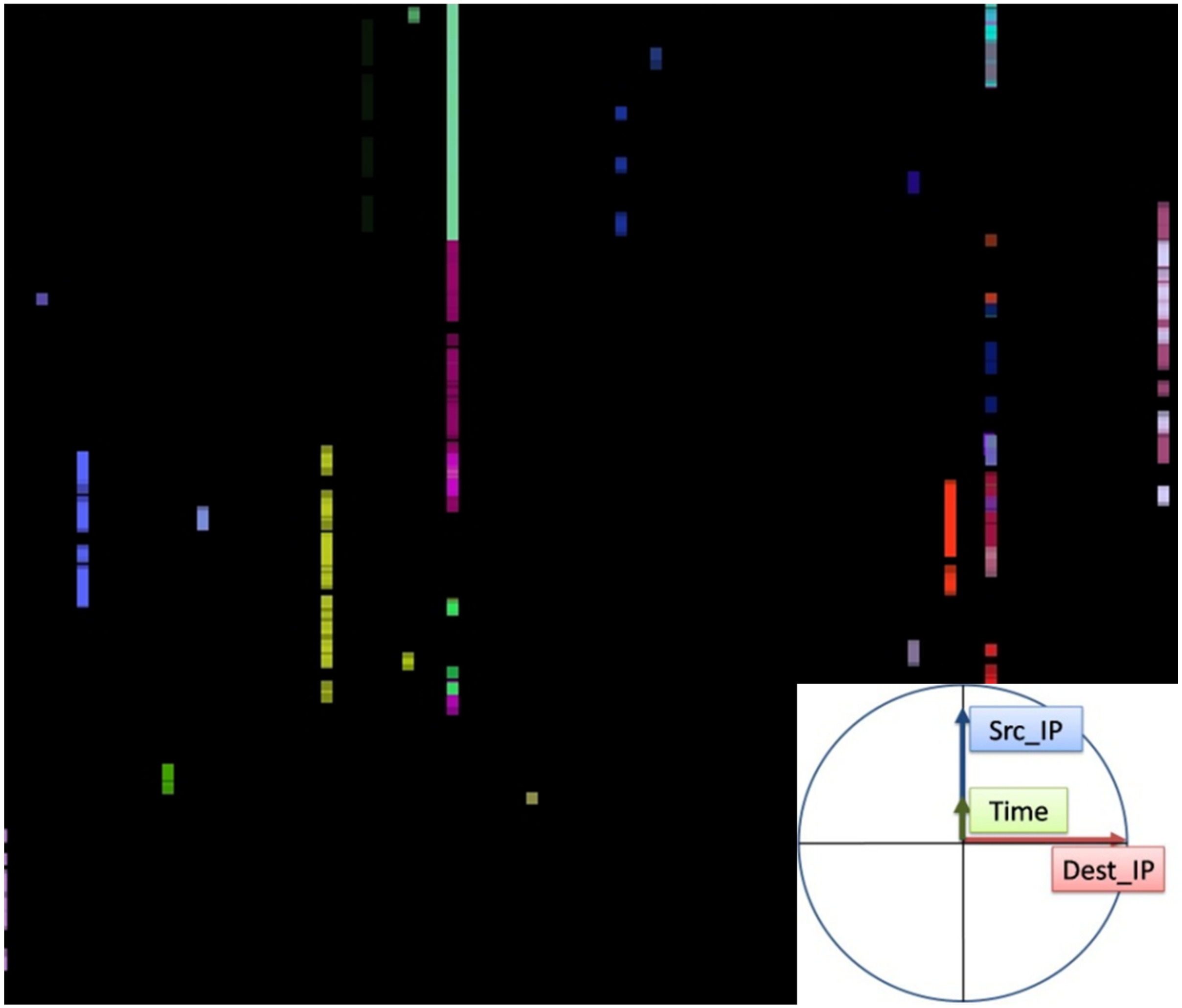
Zooming



Locating Interesting Patterns - Dynamic Display

Moving the Y Axis between Src_IP and Time dimension

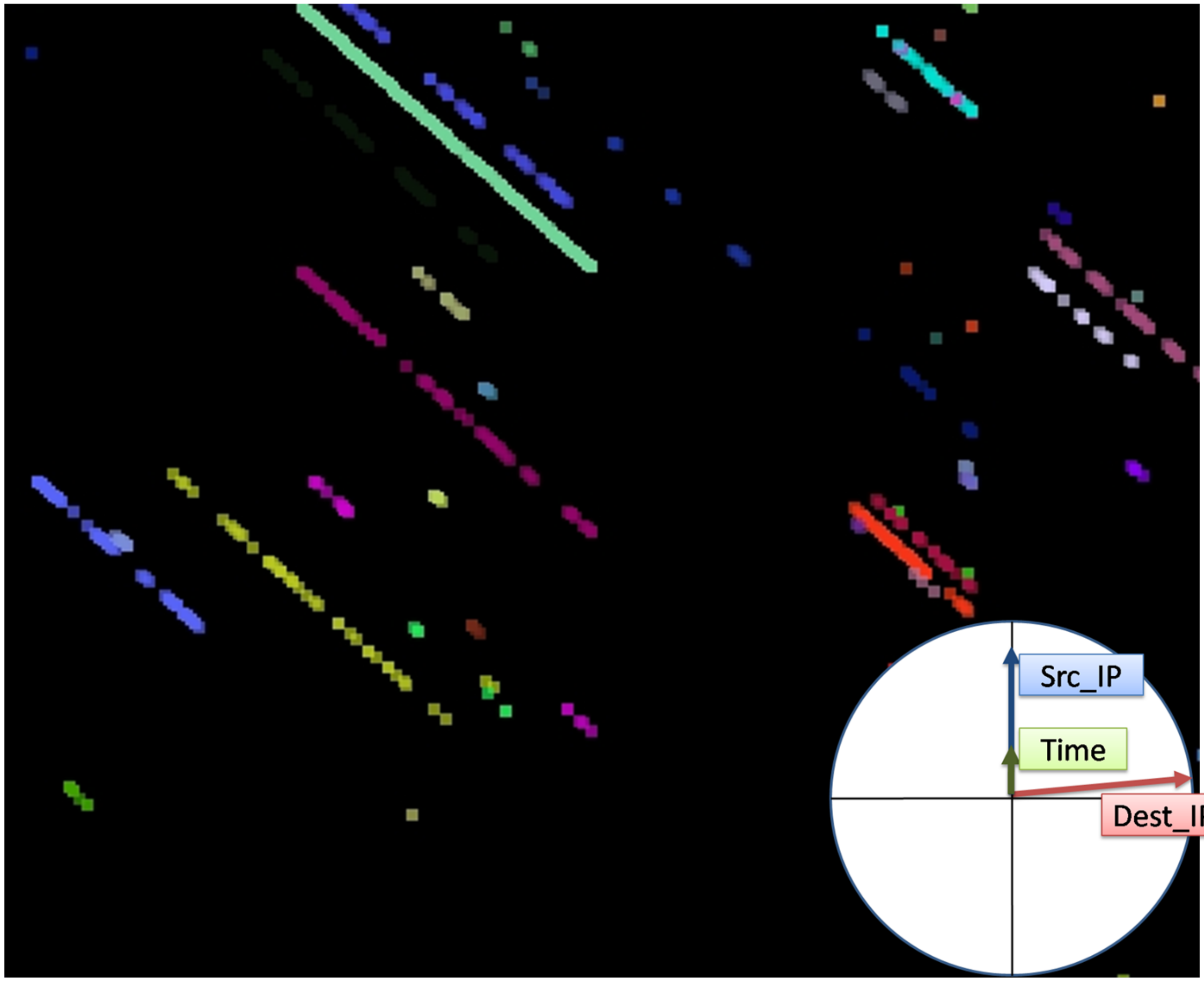
Same Color: Same Src_IP and Dest_IP



Locating Interesting Patterns - Dynamic Display

To overcome the overlap, twist the X-axis a bit.

Separate different packet groups.

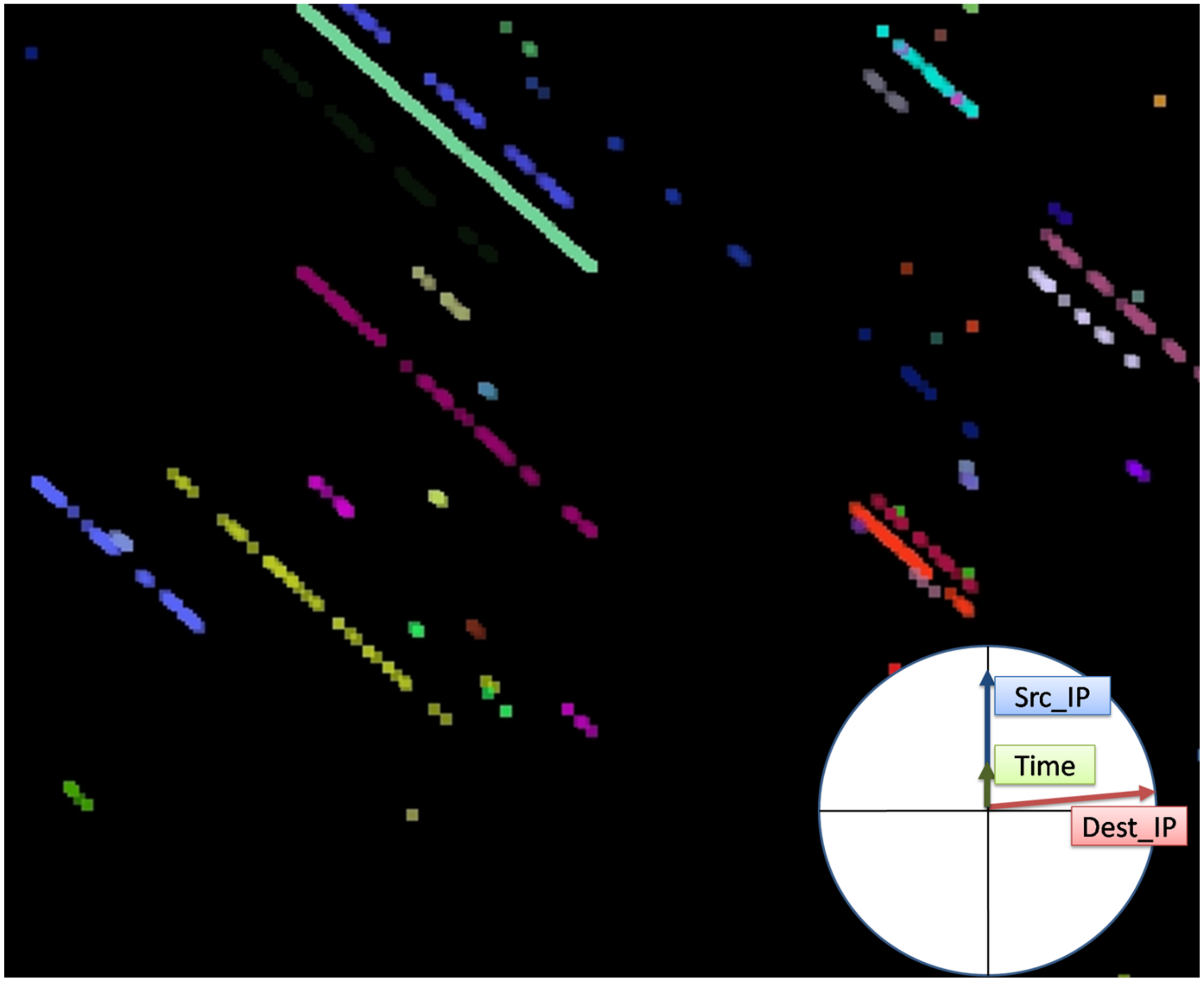


Locating Interesting Patterns - Dynamic Display

What are we looking for?

Patterns for Webpage loading

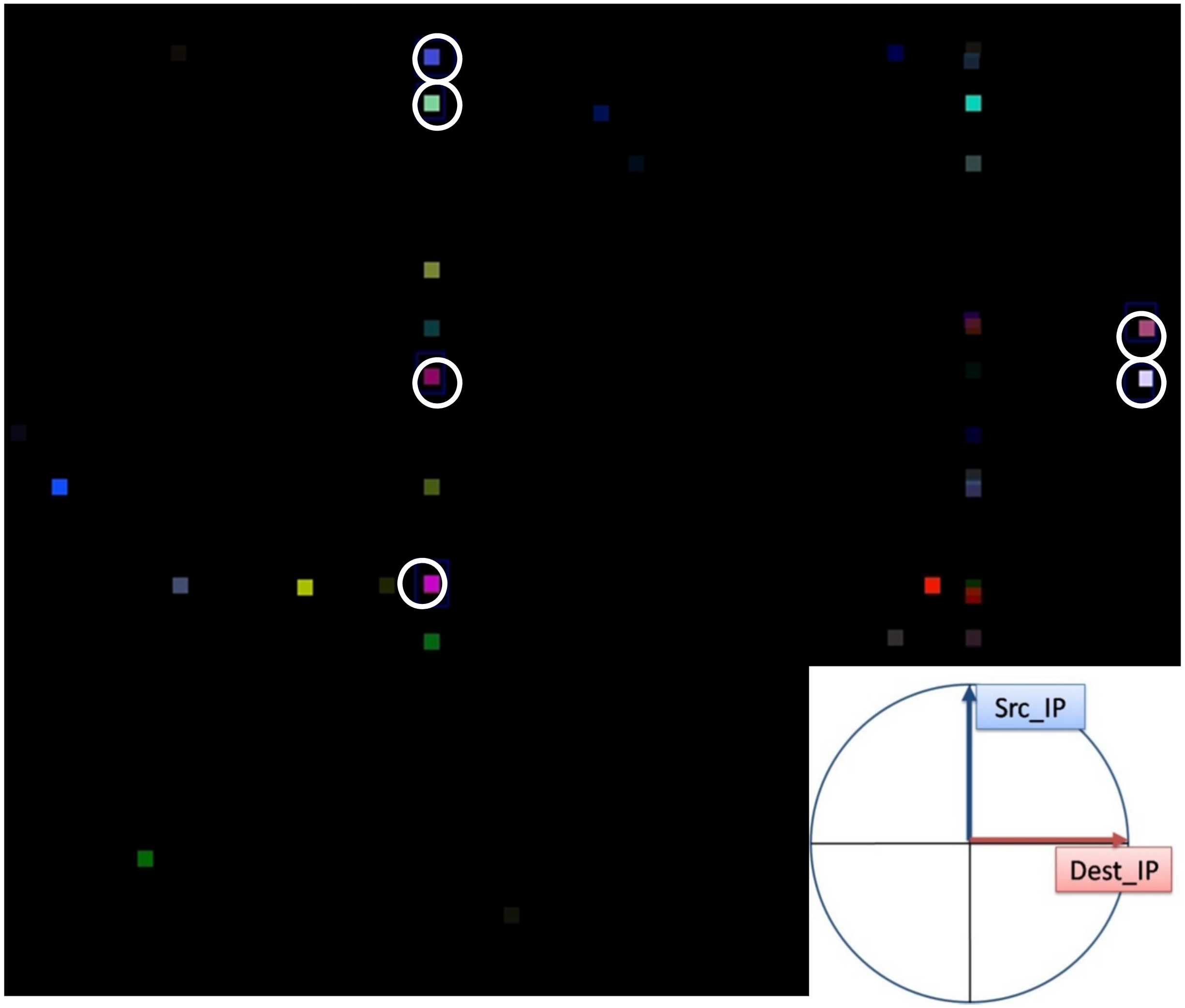
Exchanged packets between same Src IP and Dest IP in a short time period



Locating Interesting Patterns - Dynamic Display

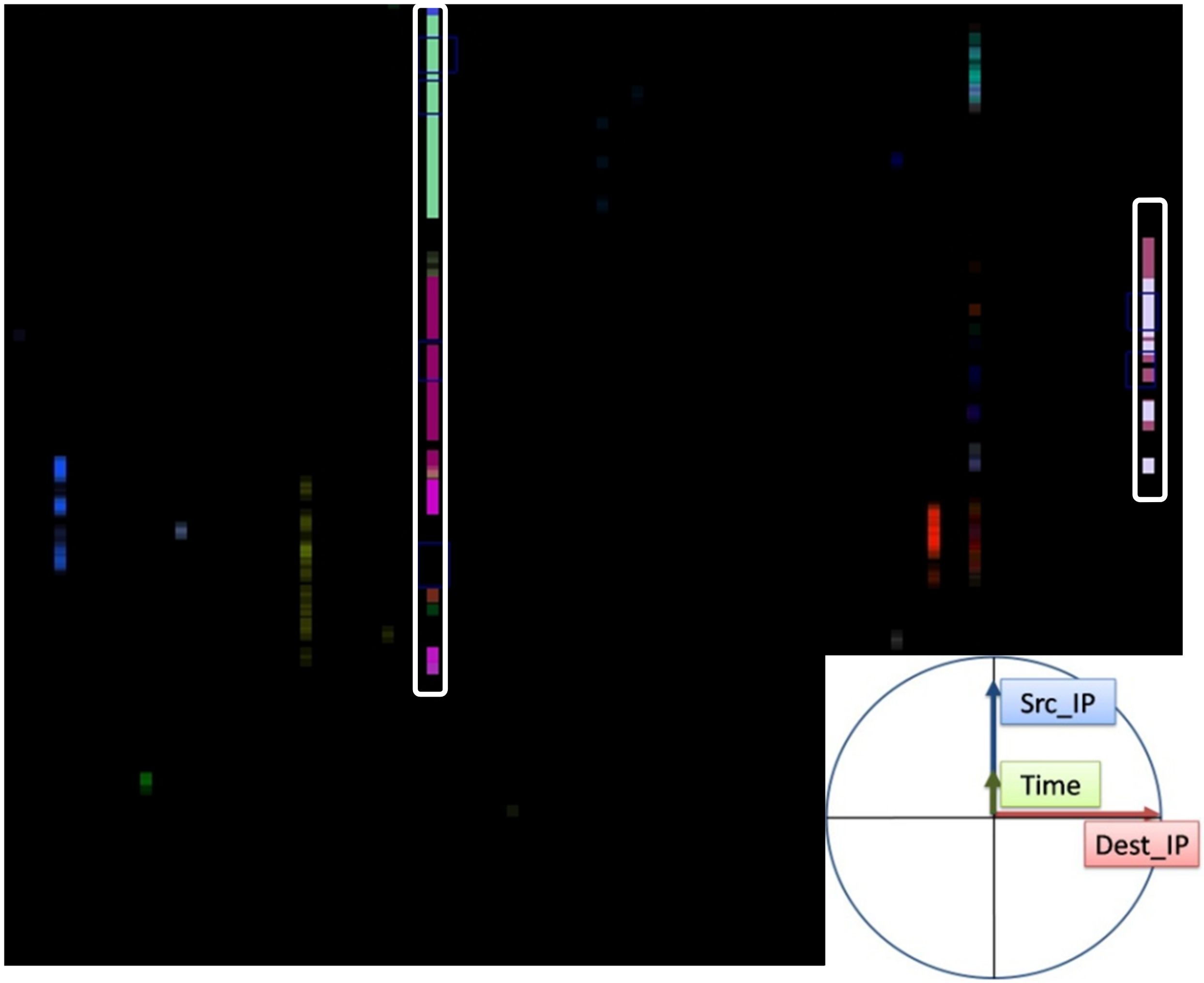
Select interesting packets

Highlight them



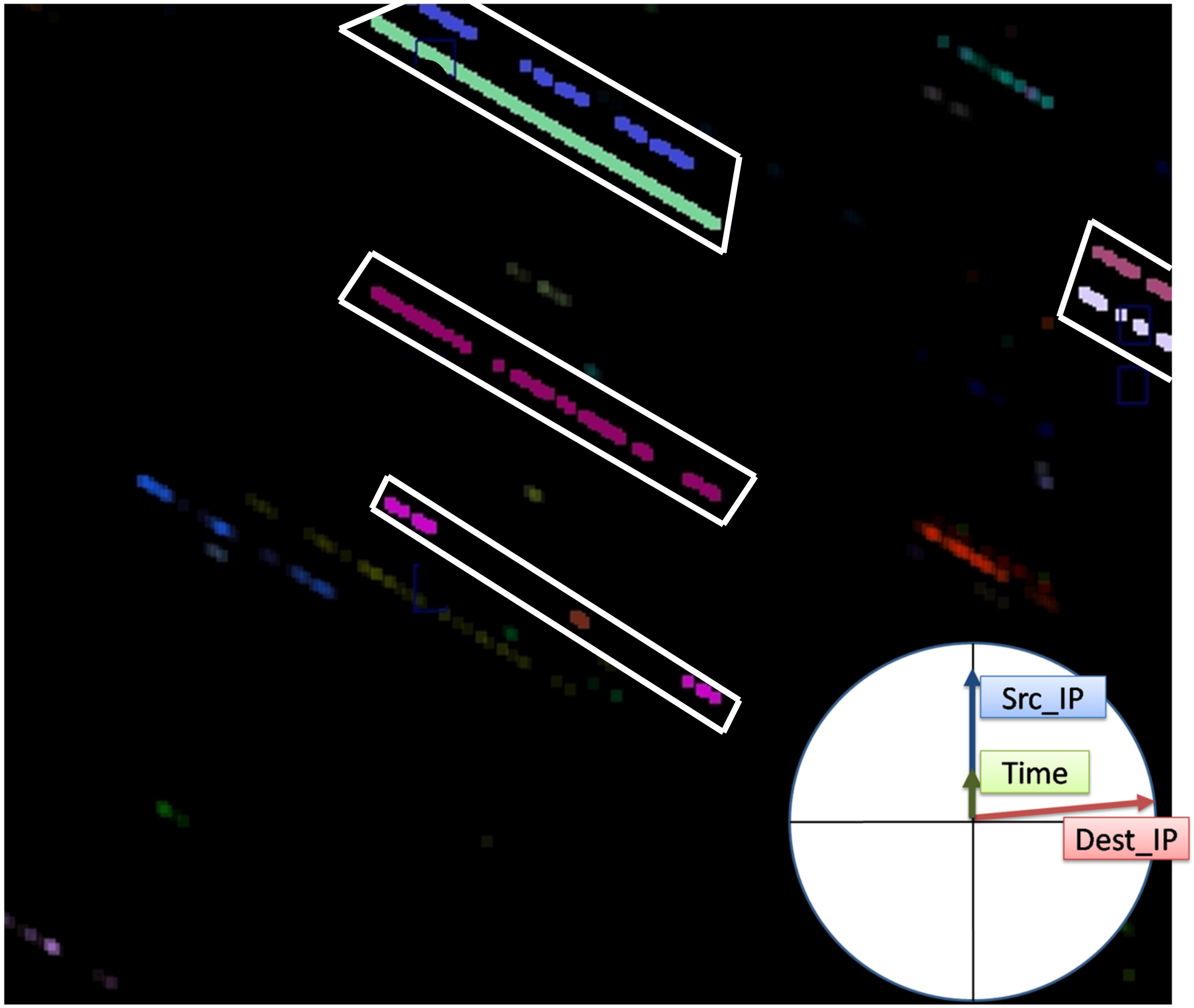
Locating Interesting Patterns - Dynamic Display

Confirm that selected packets are spreading over time

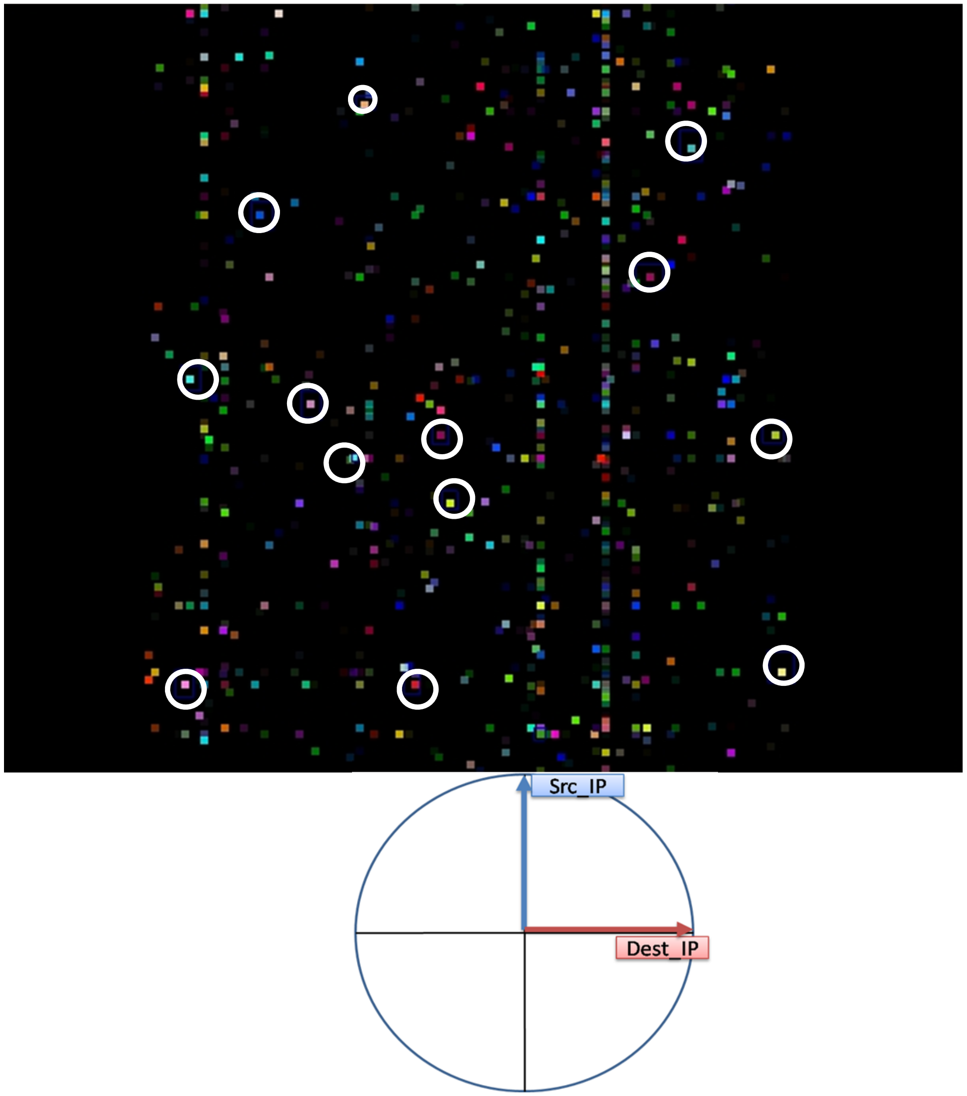


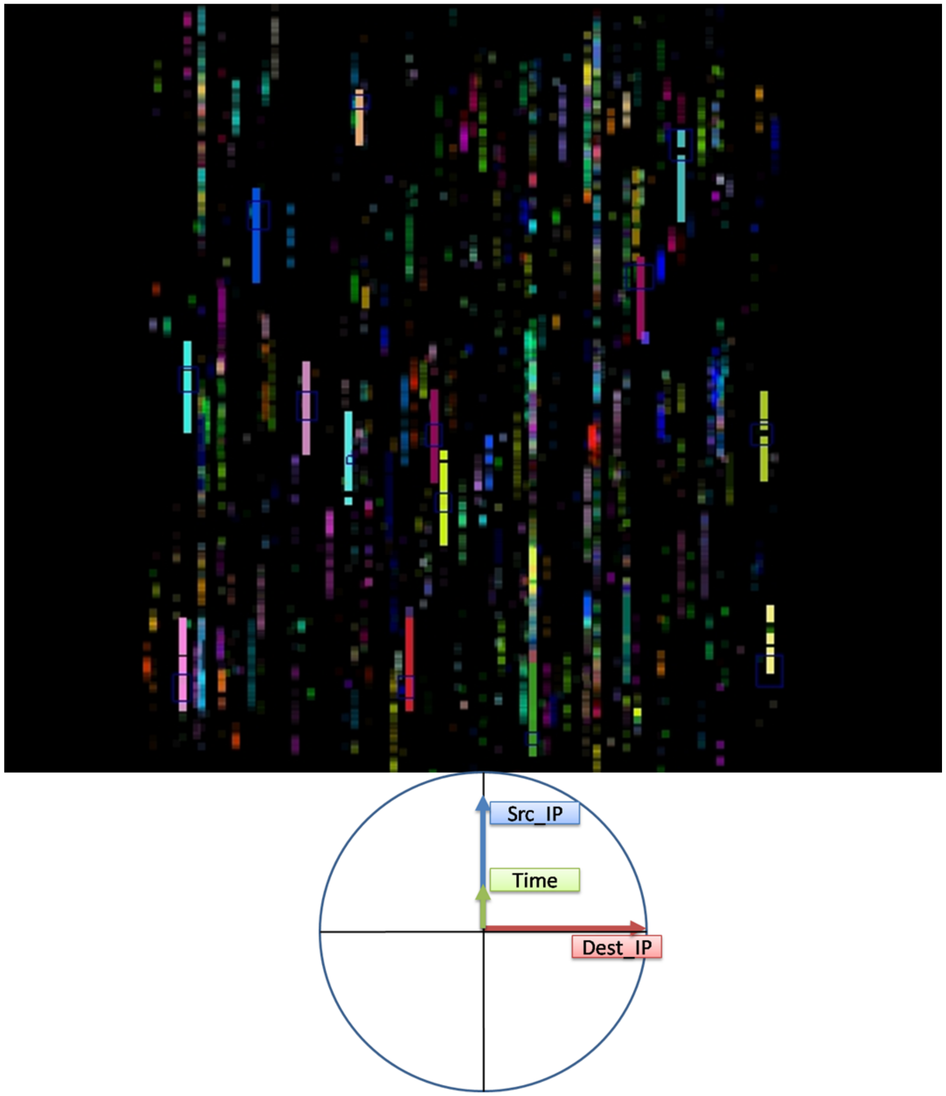
Locating Interesting Patterns - Dynamic Display

- Twist the view to separate overlapped packets



Locating Interesting Patterns - Full View





Learn the Model

Use Inductive Logic Programming (Prolog) to formulate initial model (rule):

```
webpage_load(X) :-
  same_src_ips(X),same_dest_ips(X),same_src_port(X,80),
  timeframe_upper(X,10).
```

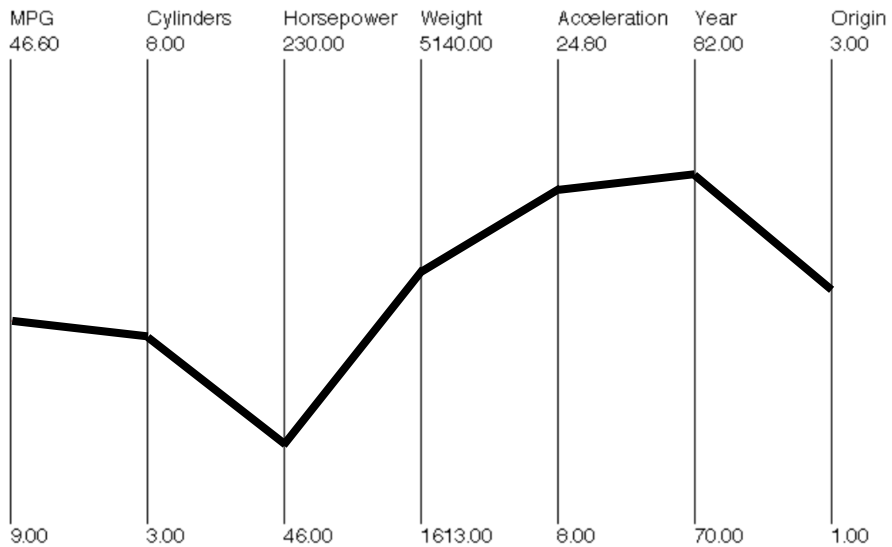
Classify other data points with this rule and visualize

Marking negative examples yields updated/refined rule:

```
webpage_load(X) :-
  same_src_ips(X),same_dest_ips(X),same_src_port(X,80),
  timeframe_upper(X,10),length(X,L),greaterthan(L,8).
```

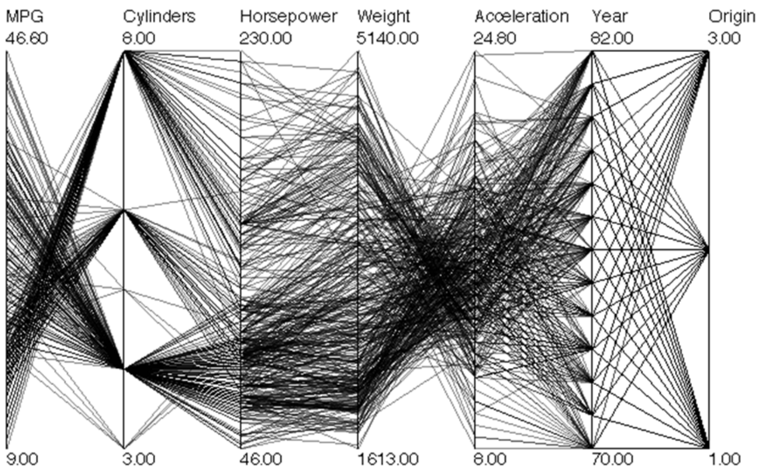
Garg, Nam, Ramakrishnan, Mueller, VAST 2008

Parallel Coordinates



a car as a 7-dimensional data point

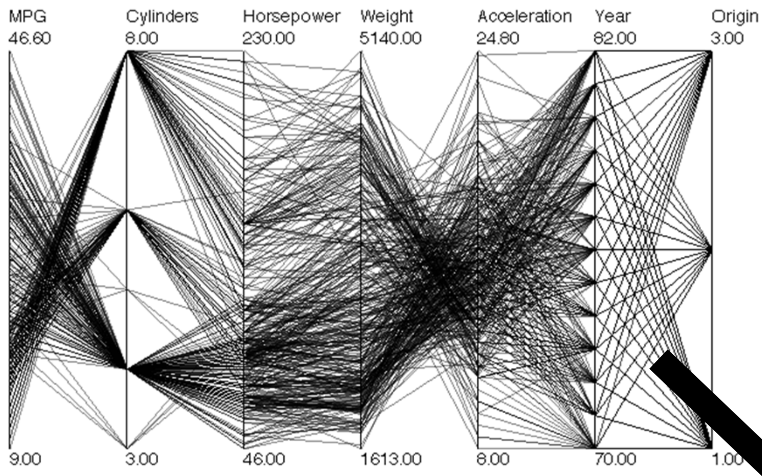
Illustrative Parallel Coordinates



Traditional parallel coordinates plot

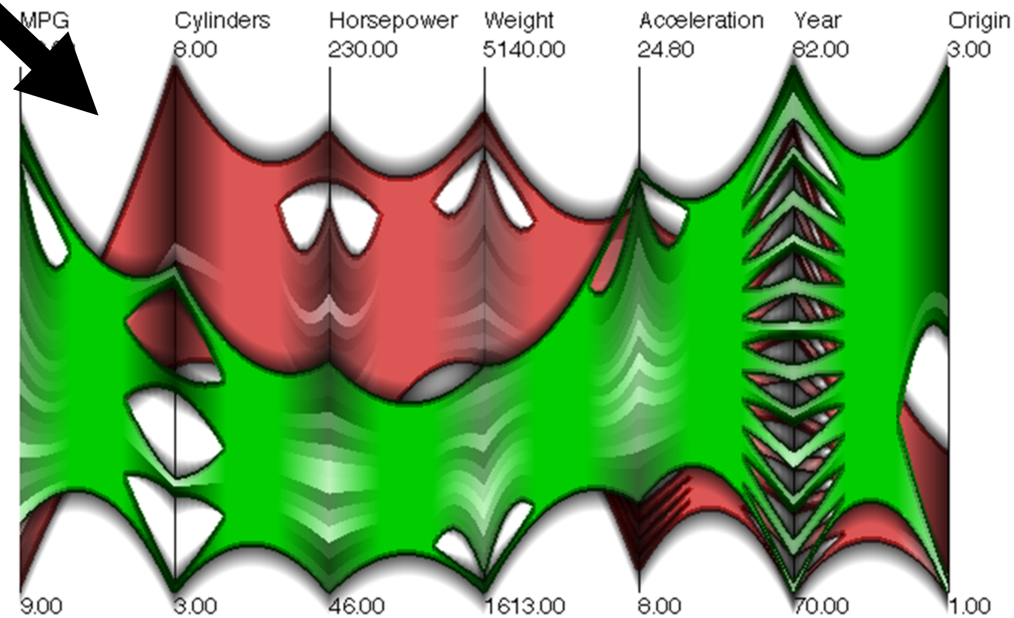
Illustrative Parallel Coordinates

STONY BROOK
Center for Visual Computing



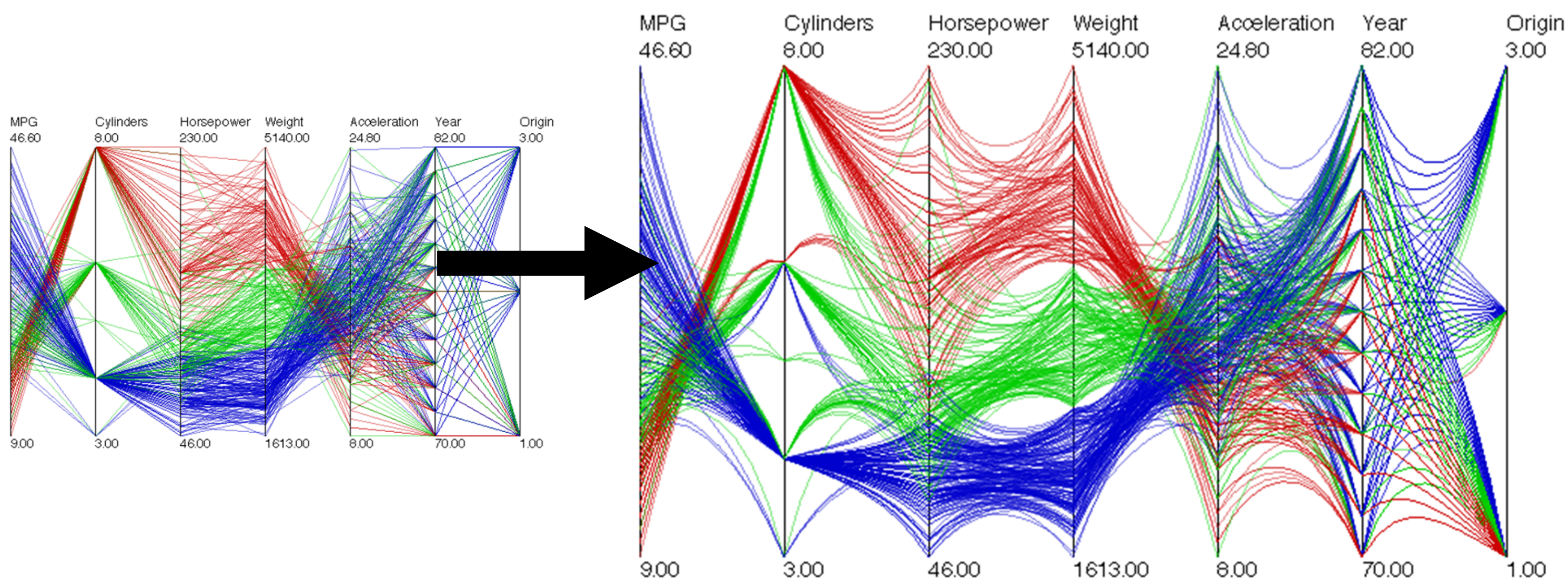
Traditional parallel coordinates plot

Illustrative parallel coordinates plot



Technique 1: Edge Bundling

Reduced clutter by replace poly-lines with poly-curves (color indicates cluster membership):

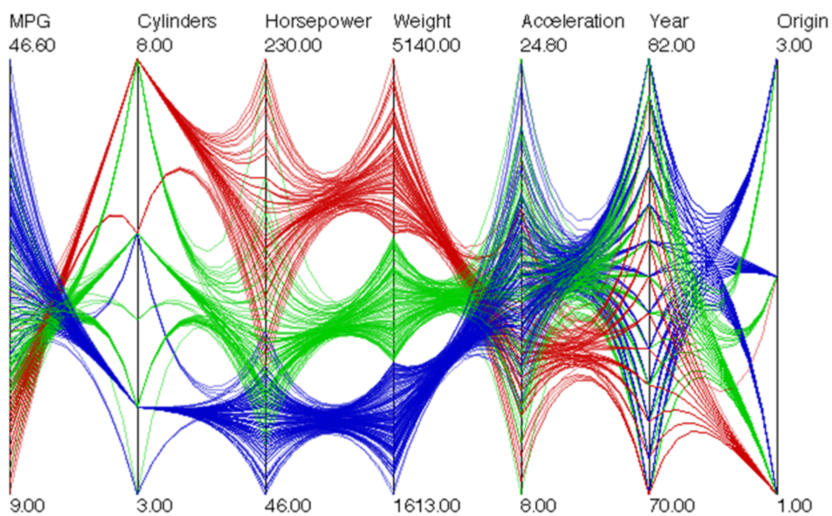


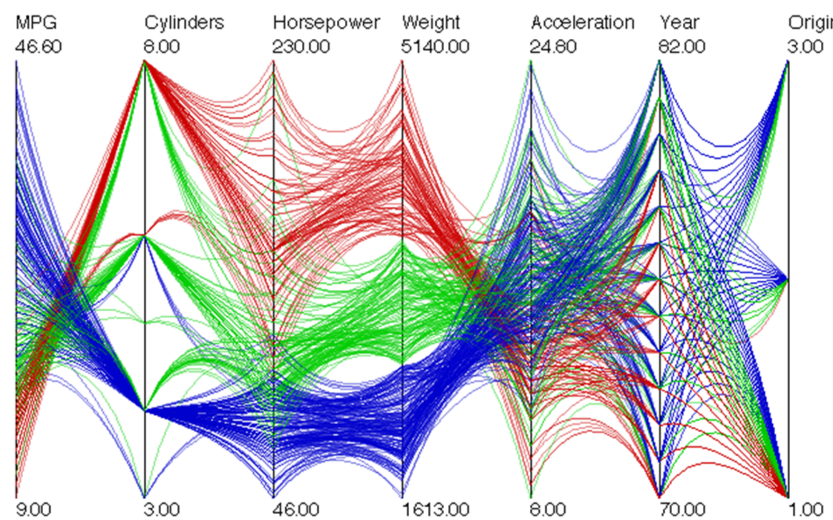
McDonnell, Mueller, Computer Graphics Forum. 2008

Edge Bundling (cont.)

Center for Visual Computing

- The user can change the tension to control the amount of clutter reduction
- Examples of low and medium tension, respectively:

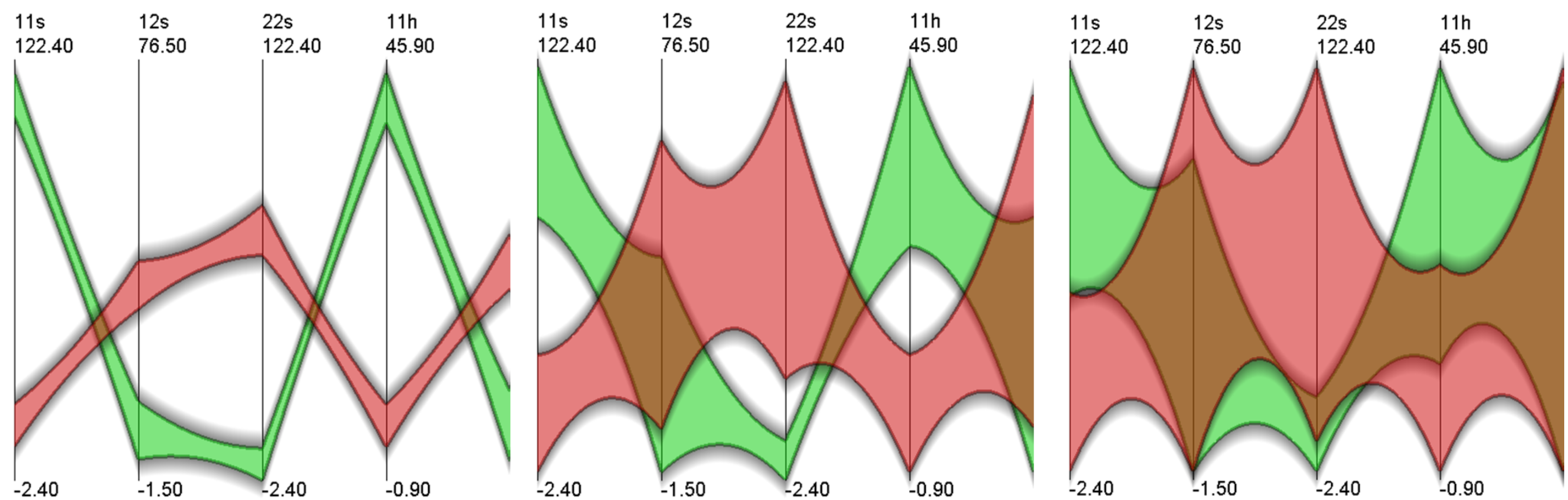




Technique 2: Cluster Rendering

In traditional PC, clusters are often rendered as heavy line segments on top of the dataset

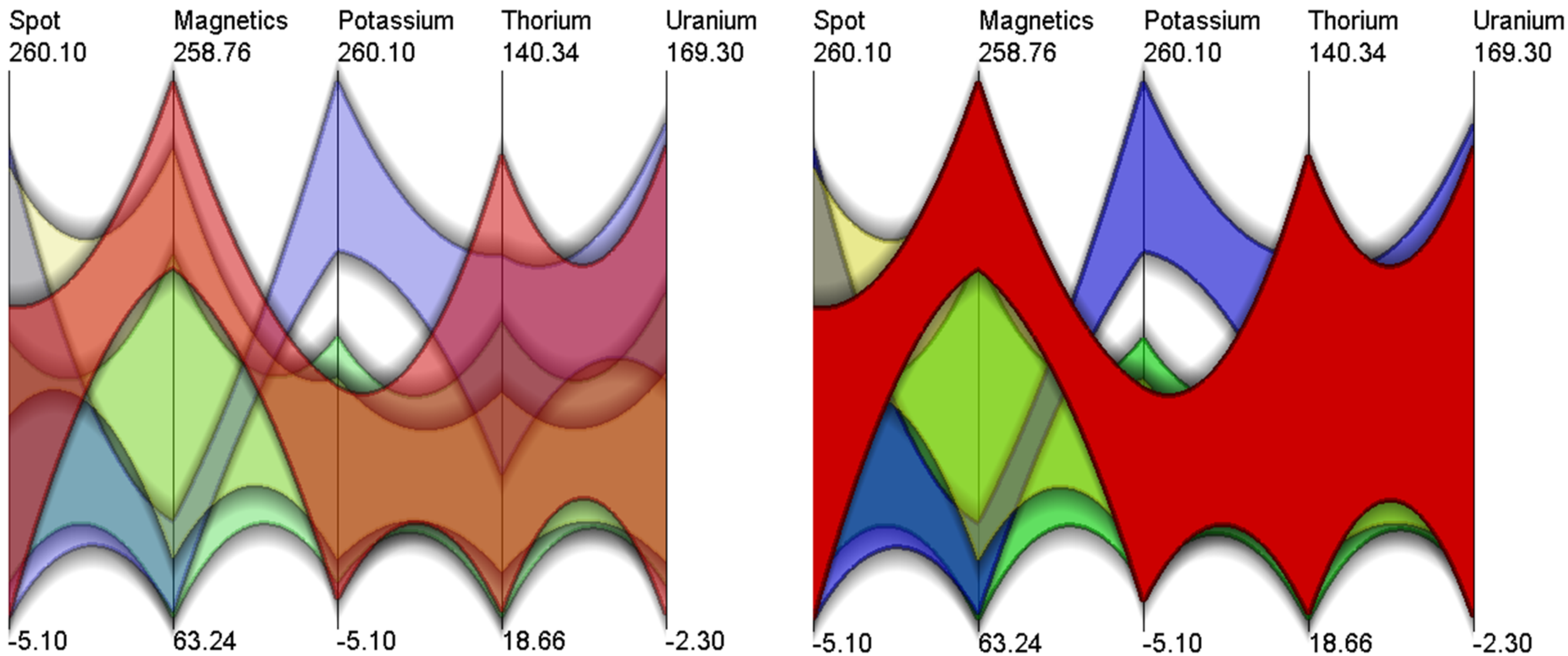
- in IPC we render the clusters as polygonal meshes
- helps to show the ranges of each cluster along axes



Technique 3: Opacity Hints

Allows context to be preserved

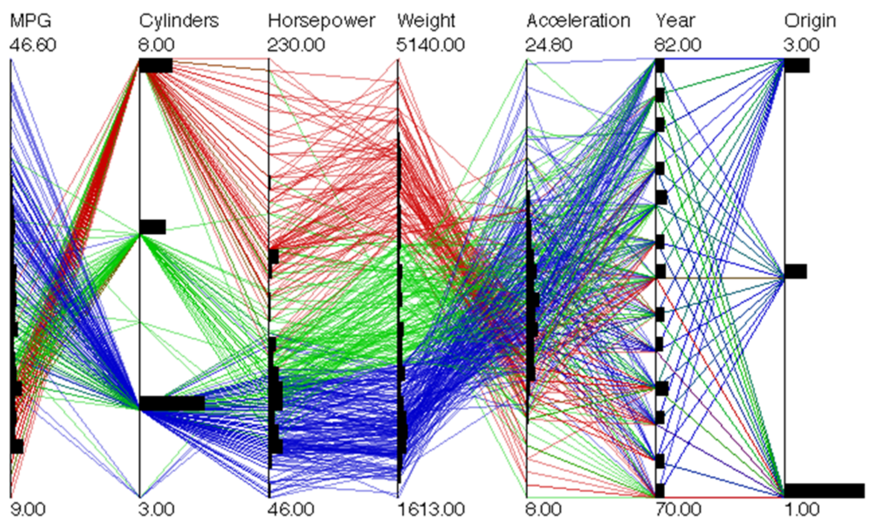
Important clusters can be made more opaque

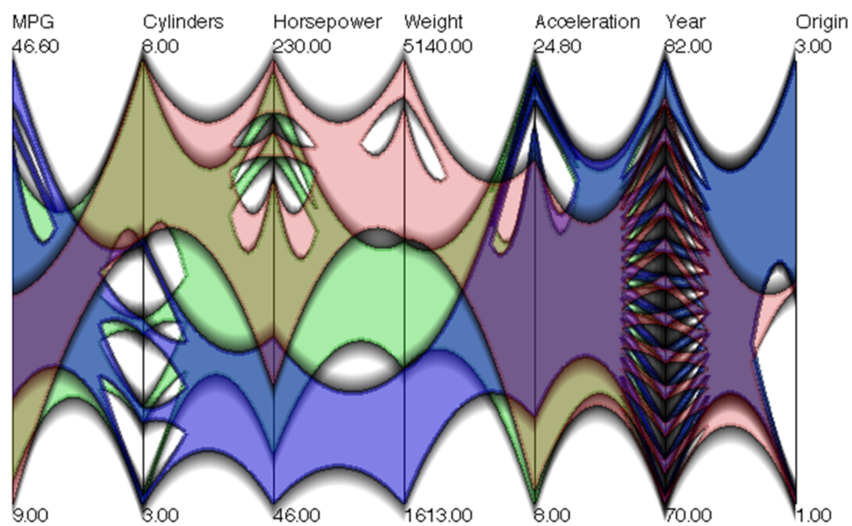


Technique 4: Branched Clusters

To illustrate the distribution of the data long each axis, it is possible to split the clusters

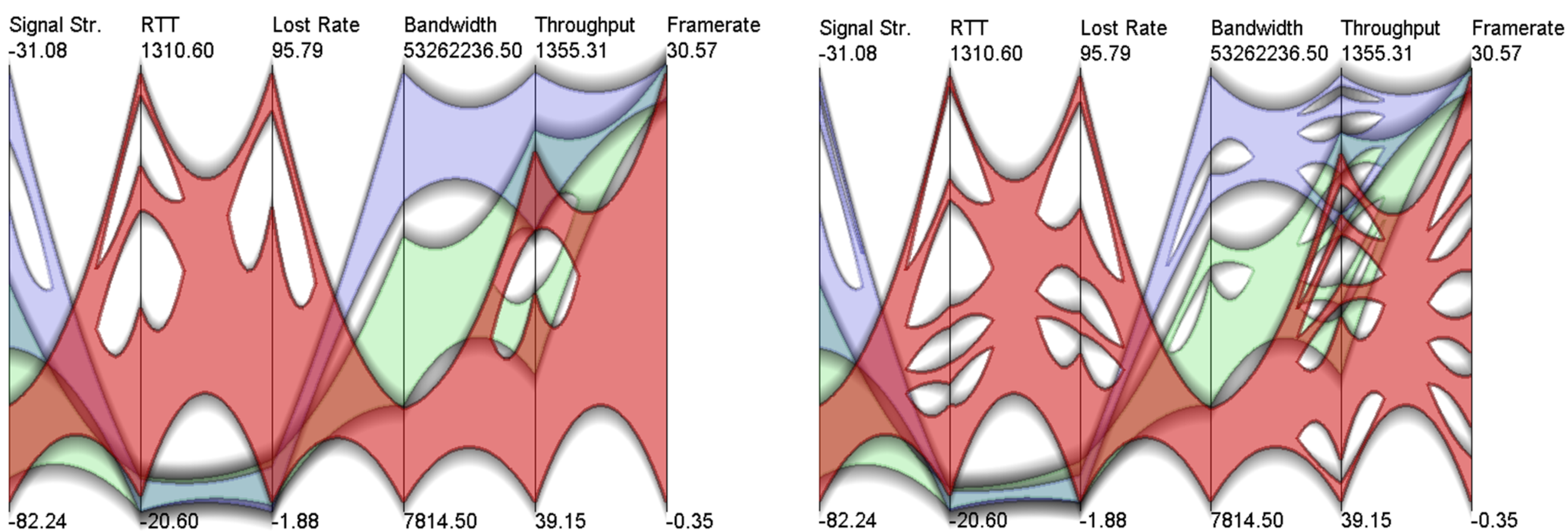
Branches provide an alternative to the display of histograms for visualizing data distributions





Branched Clusters (cont.)

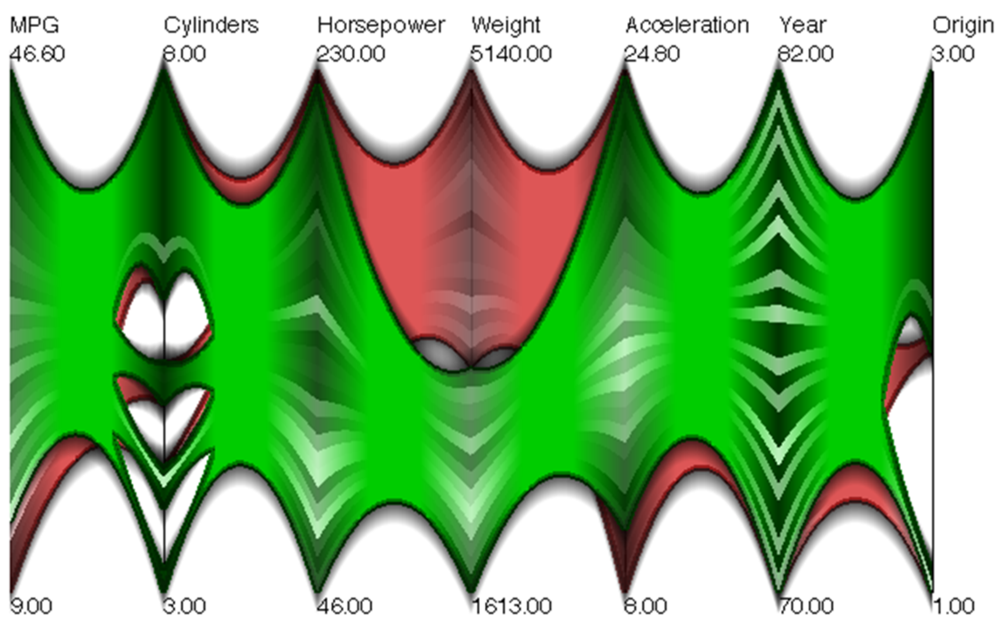
A parameter allows one to tune the visualization and change the minimum branch thickness



Technique 5: Per-Cluster Histograms

Histograms are typically used in parallel coordinate plots to show distributions along individual axes

We introduce the idea of using histograms on a per-cluster basis to reveal distribution



One More Flavor …

Lots of unstructured data on the web

We need to add structure to:

- make it machine readable
- reason with it

Humans can easily segment:

- references into author, title, etc.
- images into objects
- videos into scenes

Machine Learning Approaches

Supervised learning

- requires large amounts of *user-tagged* data
- further, data is *dynamic*
  - we might need to supplement the tagged data

Automatic learning [Raina 2007]

- Highly time intensive

Semi-Automatic Visual Learning

Keep the user in the learning loop, but:

- allow interaction with data as a whole

Use clustering methods to visually group similar objects

- helps the user mark an entire set as one category

In absence of feature vectors for a given data set

- identify important features
- allow user to adjust relative weights

→ *Visual* Active Learning

Garg, Ramakrishnan, Mueller, VAST 2010

A Good Feature Vector Is Key

Given a good feature vector:

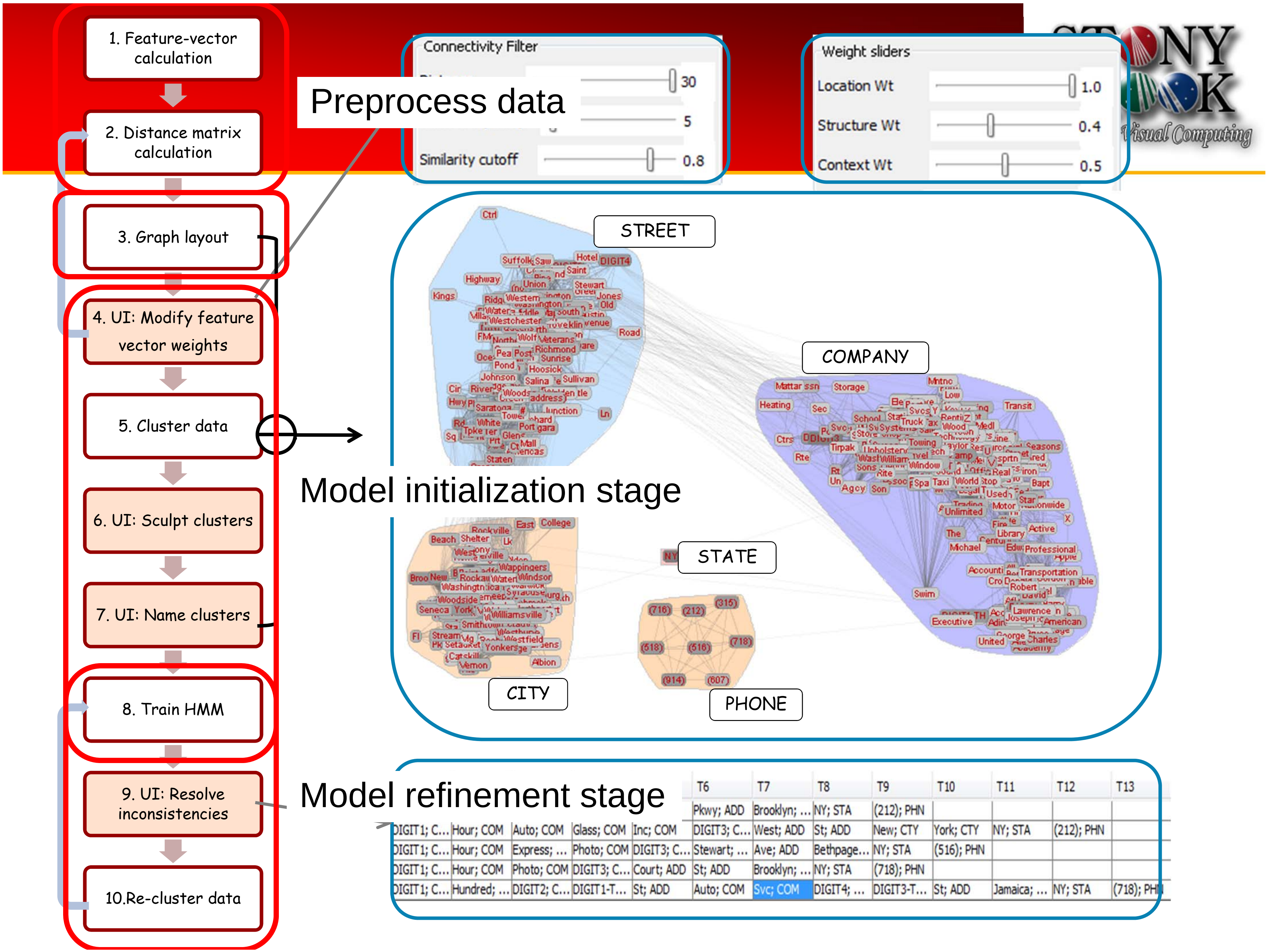
- similar points will be close-by in feature vector space

If tokens in a dataset don't have an explicit feature vector create one based on:

- structure
- context
- location
- semantics

Semantics can also simplify the problem

- e.g. in an address dataset, all numbers of the same length are interchangeable



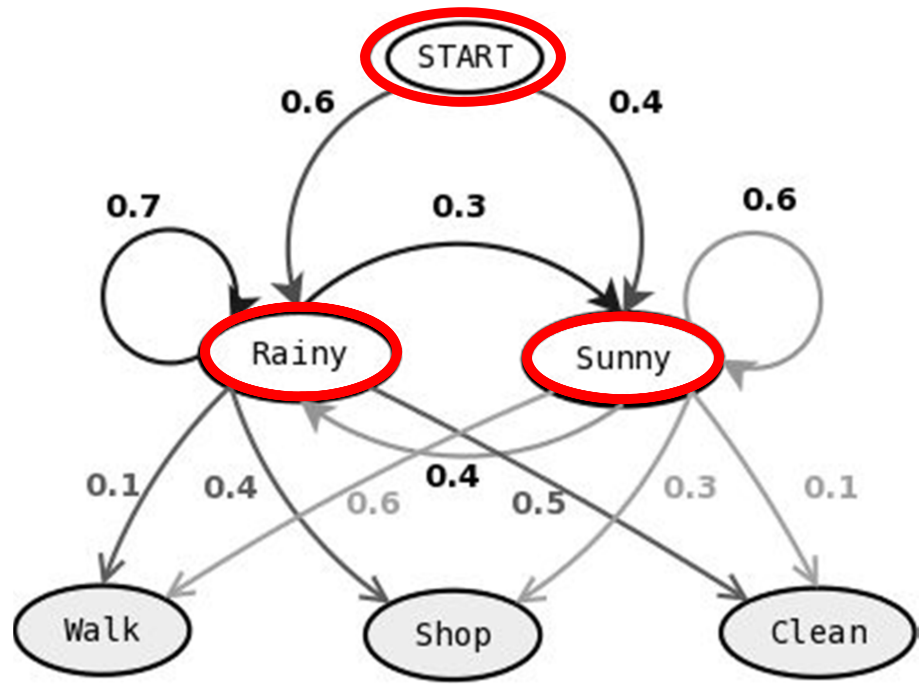
					T6	T7	T8	T9	T10	T11	T12	T13
					Pkwy; ADD	Brooklyn; ...	NY; STA	(212); PHN				
DIGIT1; C...	Hour; COM	Auto; COM	Glass; COM	Inc; COM	DIGIT3; C...	West; ADD	St; ADD	New; CTY	York; CTY	NY; STA	(212); PHN	
DIGIT1; C...	Hour; COM	Express; ...	Photo; COM	DIGIT3; C...	Stewart; ...	Ave; ADD	Bethpage...	NY; STA	(516); PHN			
DIGIT1; C...	Hour; COM	Photo; COM	DIGIT3; C...	Court; ADD	St; ADD	Brooklyn; ...	NY; STA	(718); PHN				
DIGIT1; C...	Hundred; ...	DIGIT2; C...	DIGIT1-T...	St; ADD	Auto; COM	Svc; COM	DIGIT4; ...	DIGIT3-T...	St; ADD	Jamaica; ...	NY; STA	(718); PH

Hidden Markov Model (HMM)

Statistical model used for data segmentation

Contains

- Set of (hidden) states S

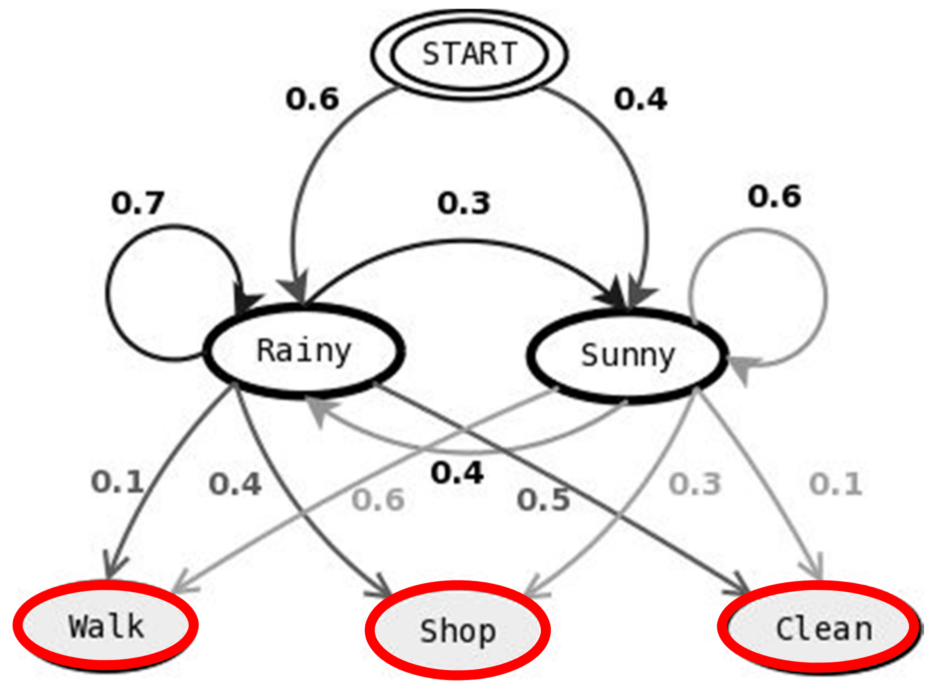


Hidden Markov Model (HMM)

Statistical model used for data segmentation

Contains

- Set of states S
- Set of observations W

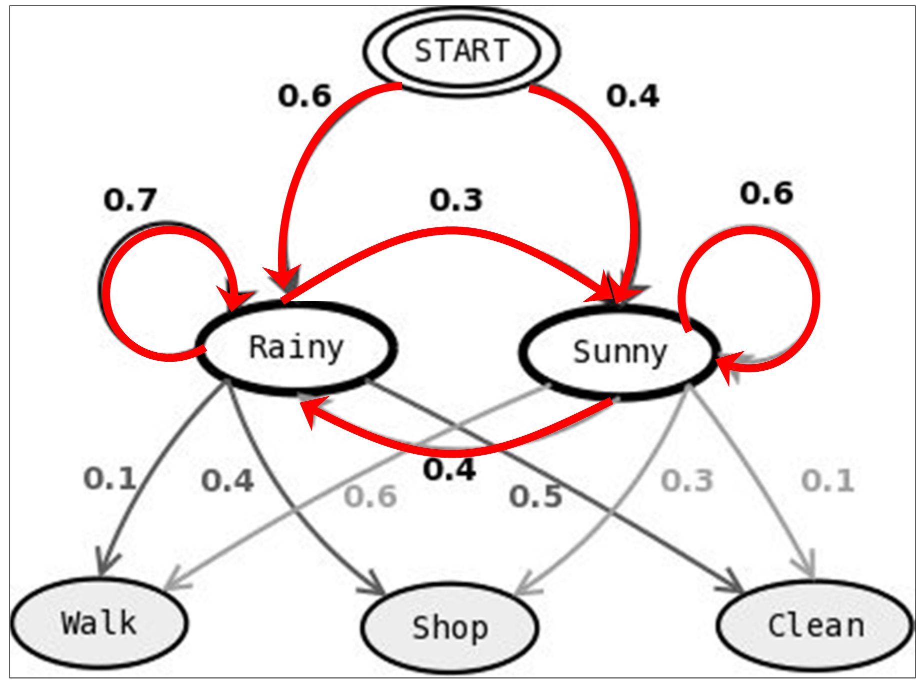


Hidden Markov Model (HMM)

Statistical model used for data segmentation

Contains

- Set of states S
- Set of observations W
- Transition model: $P(s_t \mid s_t - 1)$

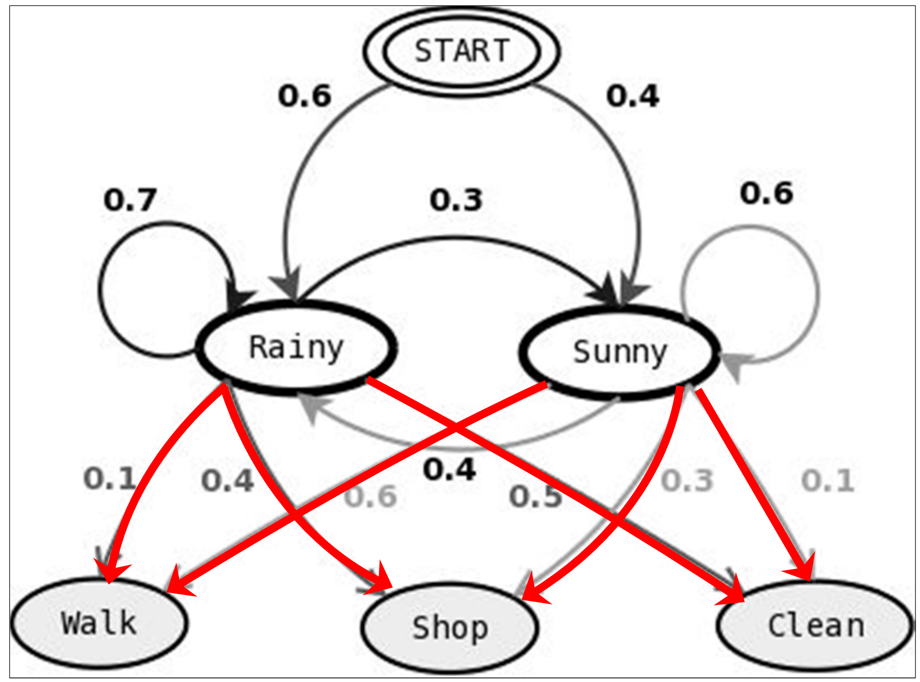


Hidden Markov Model (HMM)

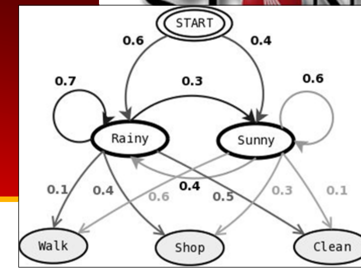
Statistical model used for data segmentation

Contains

- Set of states S
- Set of observations W
- Transition model: $P(s_t | s_t$
- Emission model: $P(w | s)$



HMM



Baum-Welch algorithm learns the model given:

- transition probabilities
- emission probabilities
- set of observations

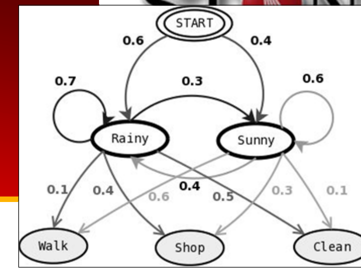
Requires hand tagged data

Gets infeasible with data size

Our solution:

- cluster the data based on feature vectors
- tag coherent data groups as a whole
- tag ambiguous data one by one

HMM: Text Segmentation



Viterbi algorithm

- returns most probable sequence of states

<COMPANY, STREET,CITY, STATE, PHONE>

Input:

- The Grand America Hotel 555 South Main Street Salt Lake City UT (800)621-4505

Output:

The Grand America Hotel,
555 South Main Street,
Salt Lake City,
UT,
(800)621-4505

Preprocessing – Windowing Approach

Window 1	Window 2	Window 3	Window 4	Window 5
1 Hour Auto	Glass Inc 403	West St	New York	NY (212)
4 Star Auto	Sound & Sec	Inc 2481 Central	Park Ave Yonkers	NY (914)
1 Hour Photo	& Copy Center	2140a White	Plains Rd Bronx	NY (718)
Westfield Agency	Inc 105	E Main	St Westfield	NY (716)
A C	P 65-09	Brook Av	Deer Park	NY (516)
A A M	C A R	303 W 96th	St New York	NY (212)

Windowing Approach

Window 1	Window 2	Window 3	Window 4	Window 5
1 Hour **Auto**	Glass Inc 403	West St	New York	NY (212)
4 Star **Auto**	Sound & Sec	Inc 2481 Central	Park Ave Yonkers	NY (914)
1 Hour Photo	& Copy Center	2140a White	Plains Rd Bronx	NY (718)
Westfield Agency	Inc 105	E Main	St Westfield	NY (716)
A C	P 65-09	Brook Av	Deer Park	NY (516)
A A M	C A R	303 W 96th	St New York	NY (212)
2	0	0	0	0

Feature Vectors in a Text Dataset

Structure

- What type of characters does the token contain

Context

er

Word	Has letter	Has digit	Has symbol	Has caps	All caps	Length
Liberty	1	0	0	1	0	7

Word	Neigh-bors	Has letter	Has digit	Has symbol	Has caps	All caps	Length 1-3	Length 4-6	Length 7+
Liberty	Av. Avenue 1344 A-1	1 1 0 1	0 0 1 1	1 0 0 1	1 1 0 1	0 0 0 0	1 0 0 1	0 1 1 0	0 0 0 0
Final F-vec		3	2	2	3	0	2	2	0

Distance matrix

Given feature vectors, calculate all pairs of distances

$$sim(x,y) = \frac{\sum_i w_i * sim_i(x,y)}{\sum_i w_i}$$

User modifiable

Token Visualization: Random Layout

Avenue

Hour

Yonkers

Agency

Highway

Alabama

Washington

1505

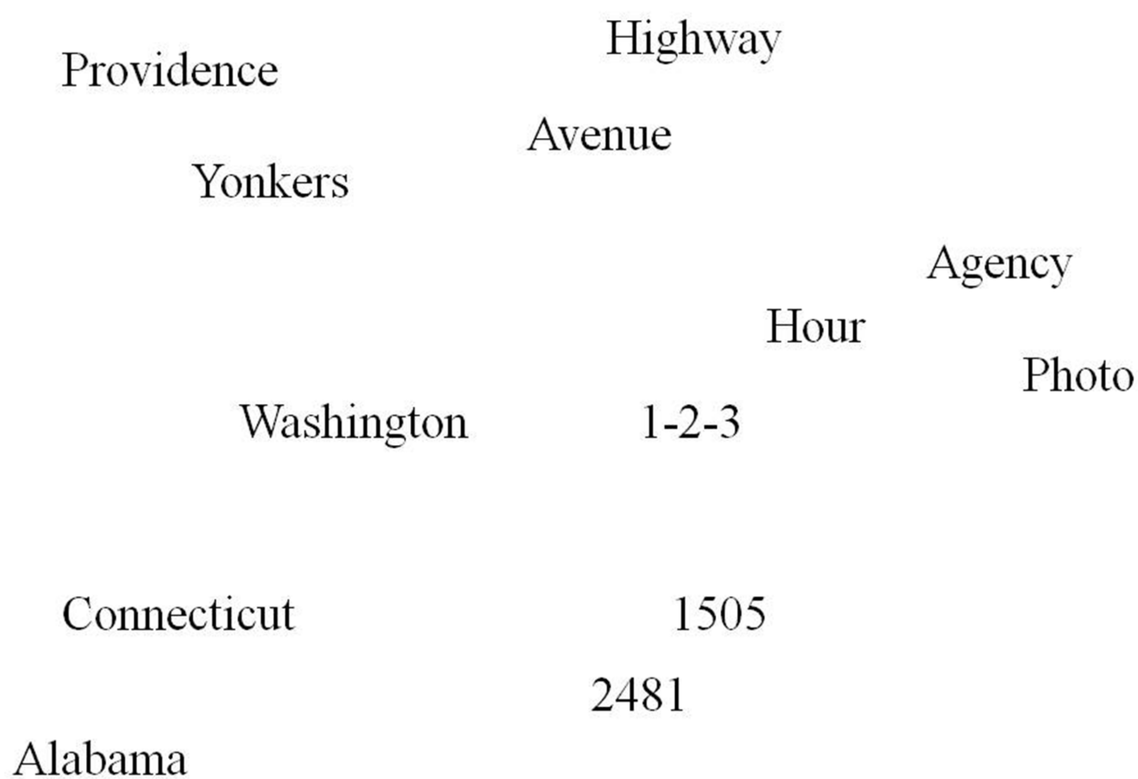
1-2-3

Connecticut

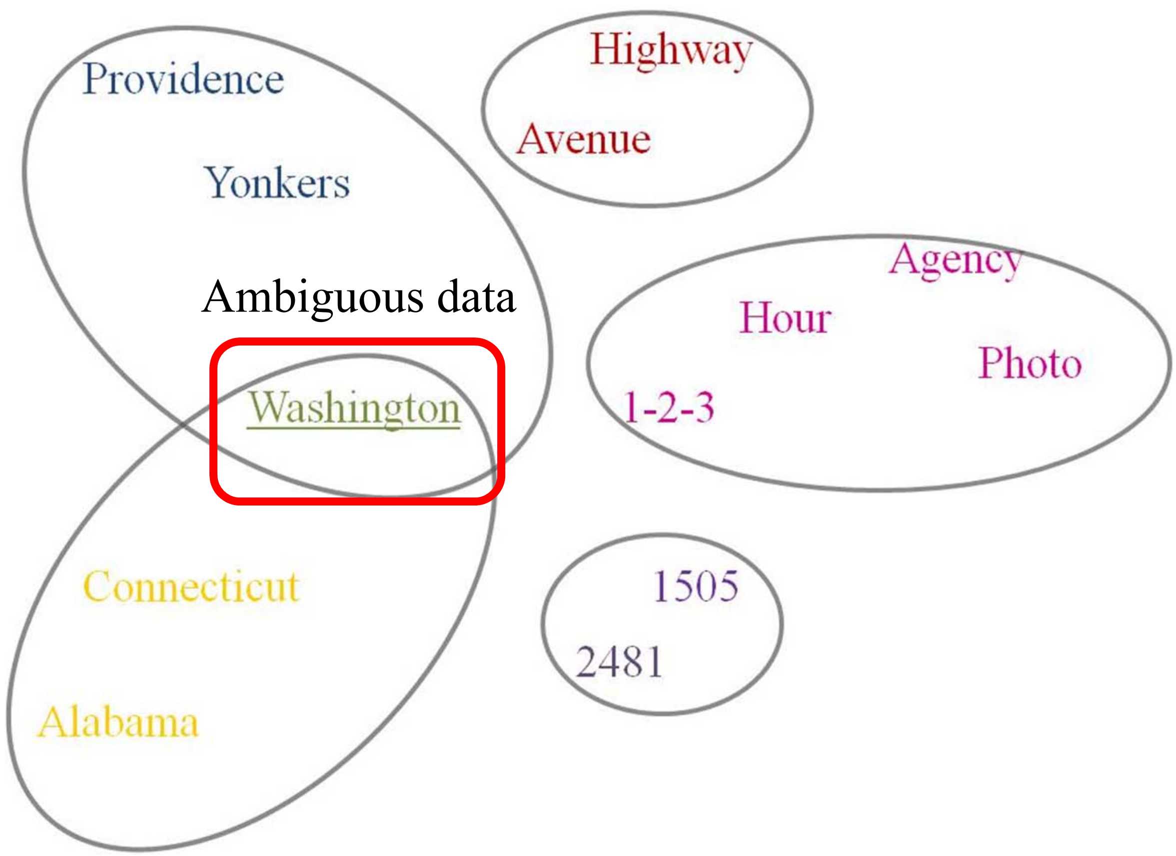
Photo

Providence

Token Visualization: Distance Based Layout



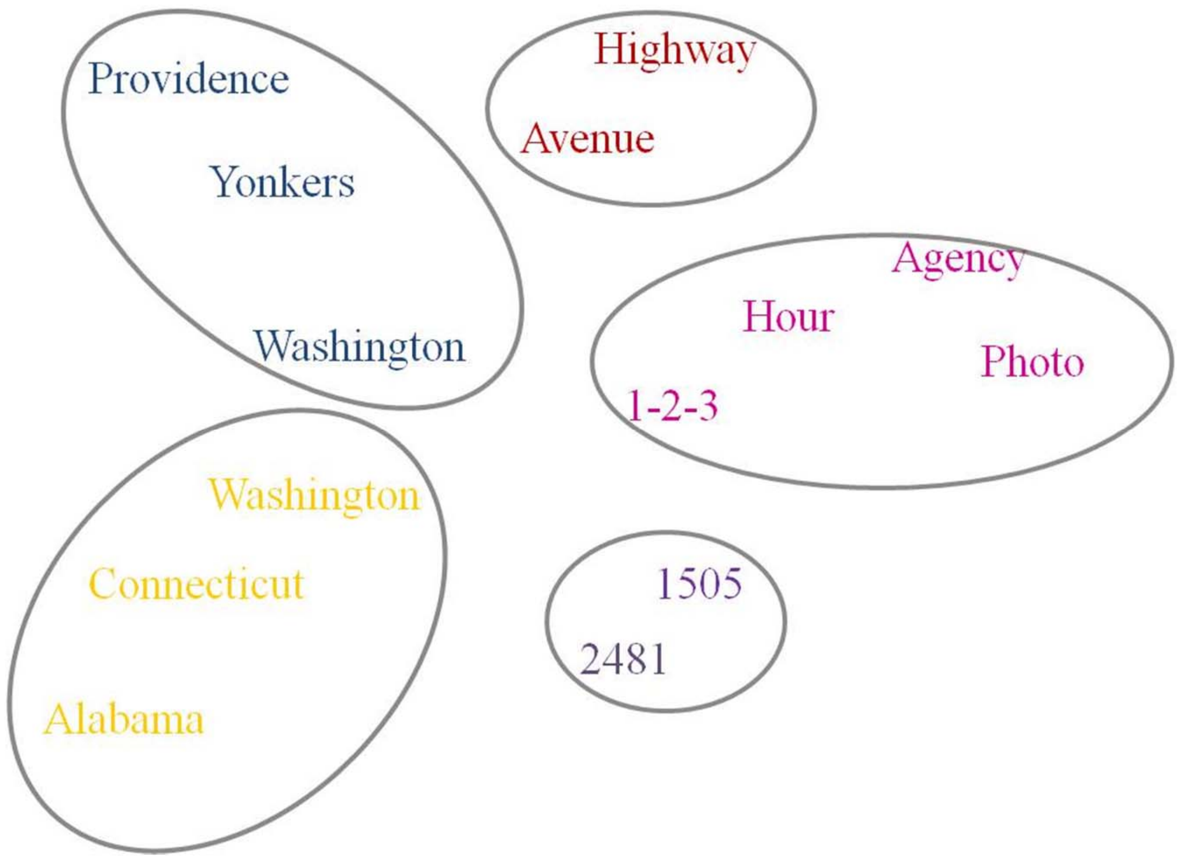
Token Visualization: User Assigned Categories



Token Visualization: Disambiguation

Window 1	Window 2	Window 3	Window 4	Window 5
Corte Salon	1019 U	St NW 2nd	Fl **Washington**	DC 20001
Glover Park	Hardware 2251	Wisconsin Ave	NW **Washington**	DC 20007
Laura Bee	Designs 6418	20th Ave	NW Seattle	**Washington** 98107
Bob's Quality	Meats 4861	Rainier Avenue	S Seattle	**Washington** 98118

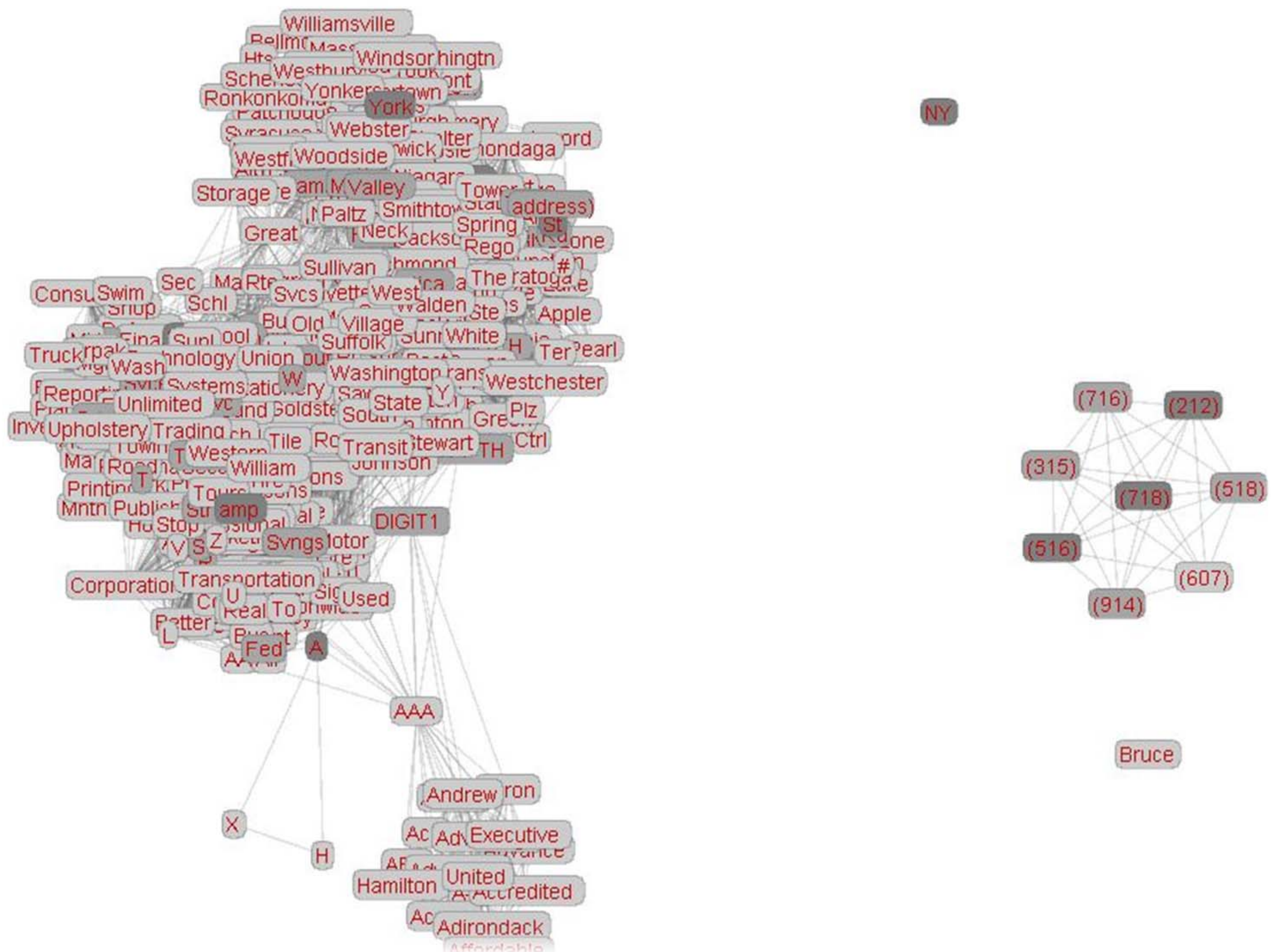
Token Visualization: Disambiguation



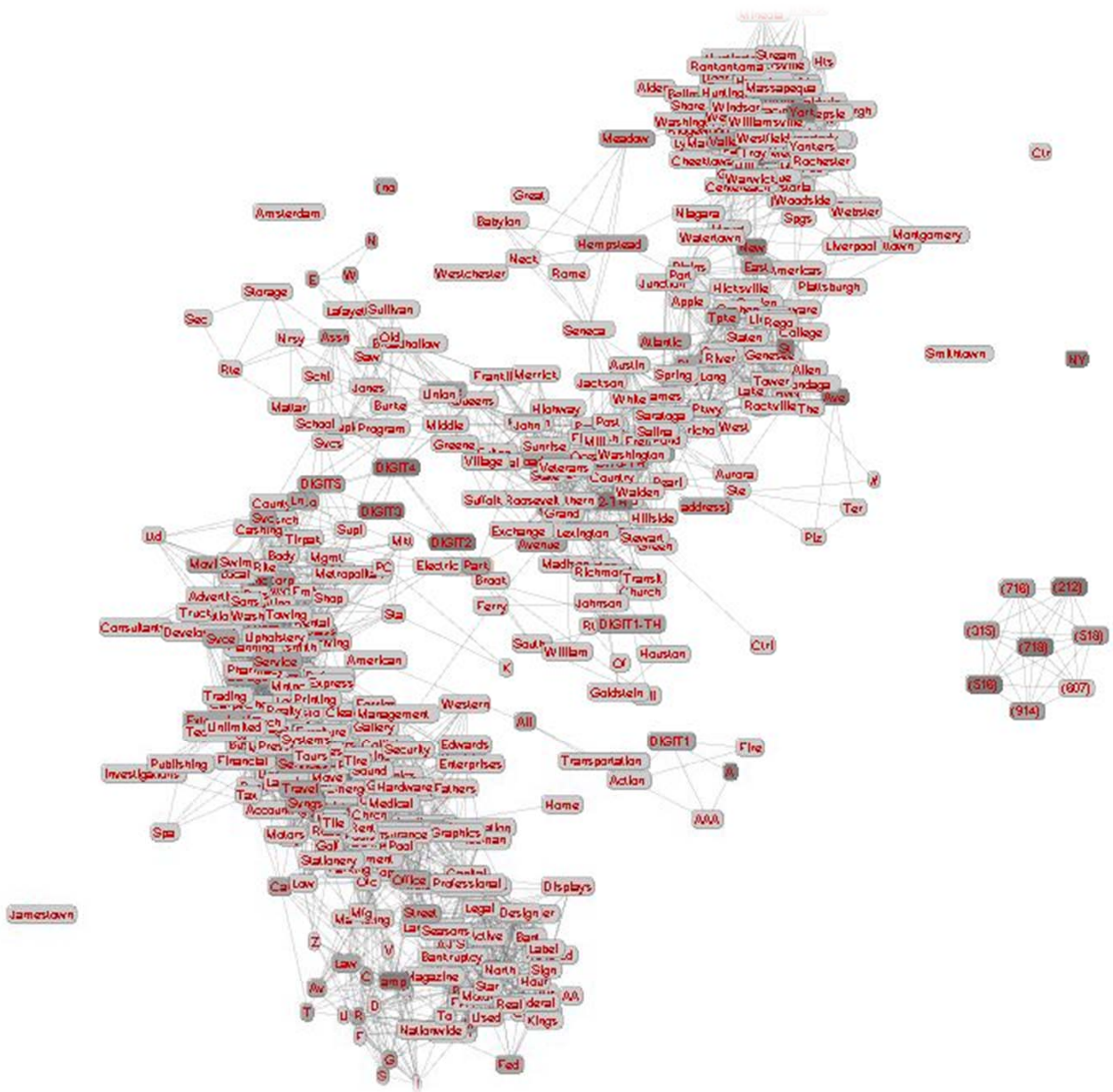
Results: Address Data Set

Segmenting an address dataset of NY businesses

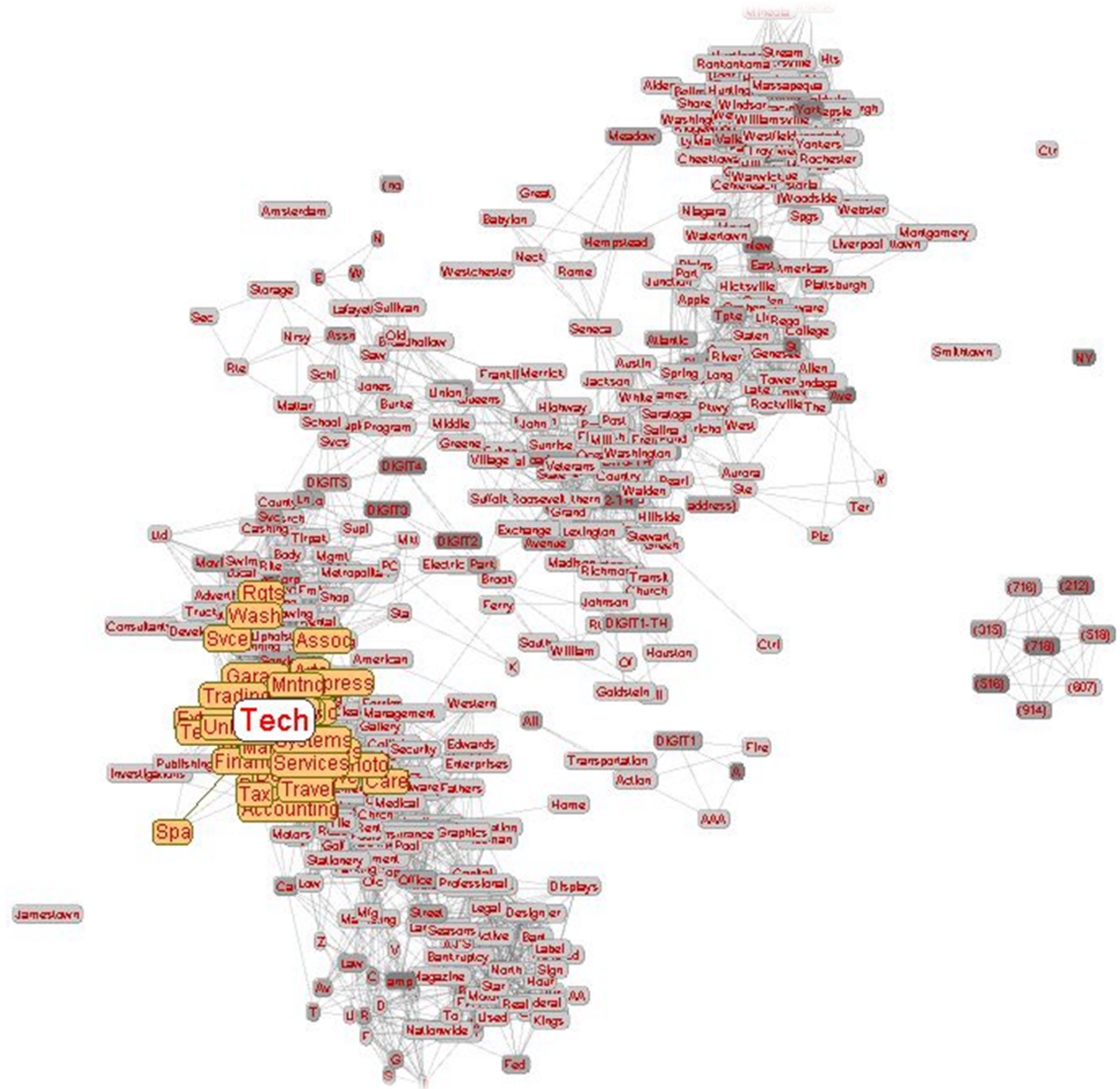
Initial Layout



Layout After Tweaking Feature Vector Weights

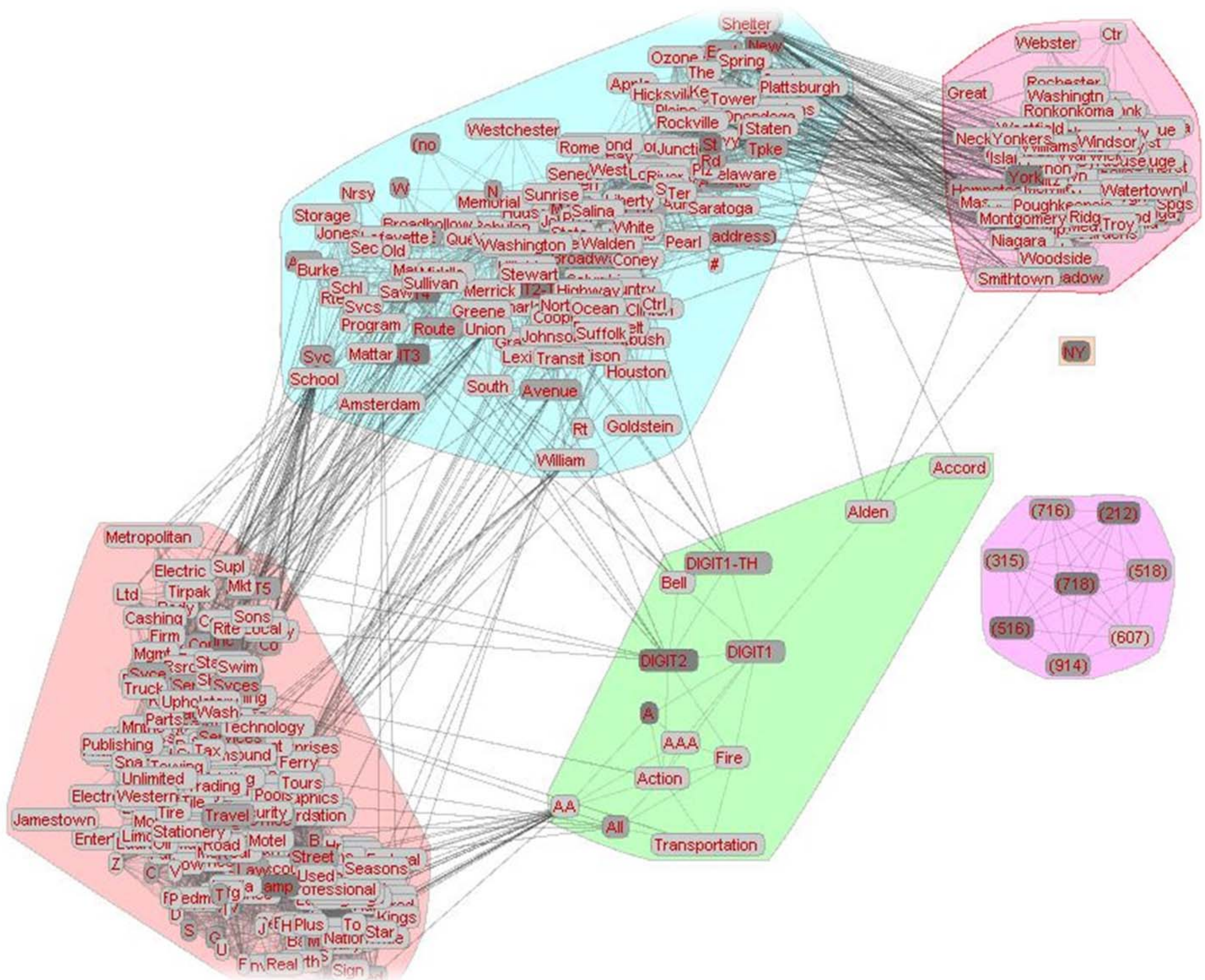


Zooming In

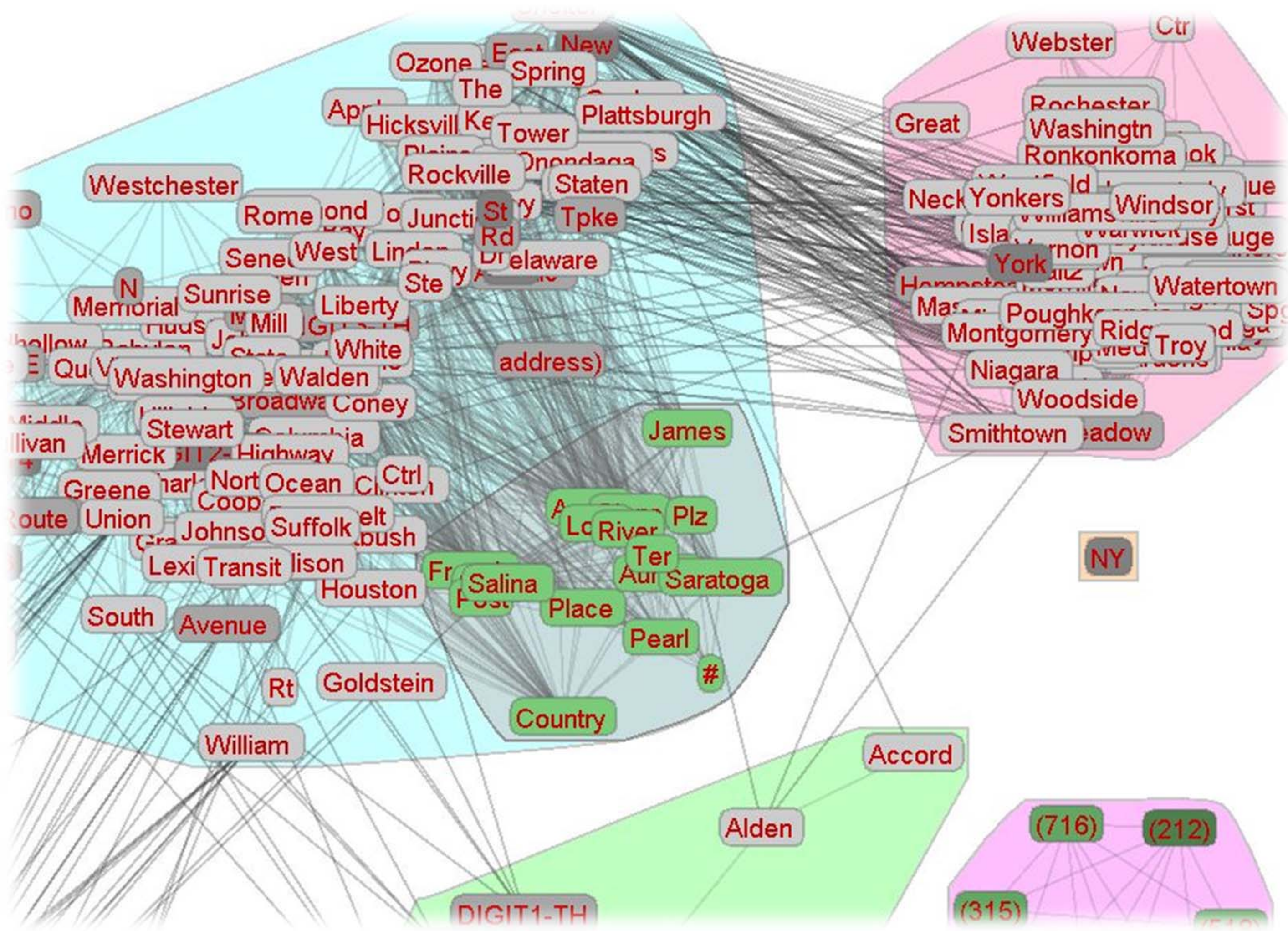


Tech
Rqts
Wash
Svce
Assoc
Spa
Tax
Travel
Accounting
Services

Layout After Clustering Using Markov Cluster Algorithm



Cluster Naming Using Inner Core



Cluster Editing

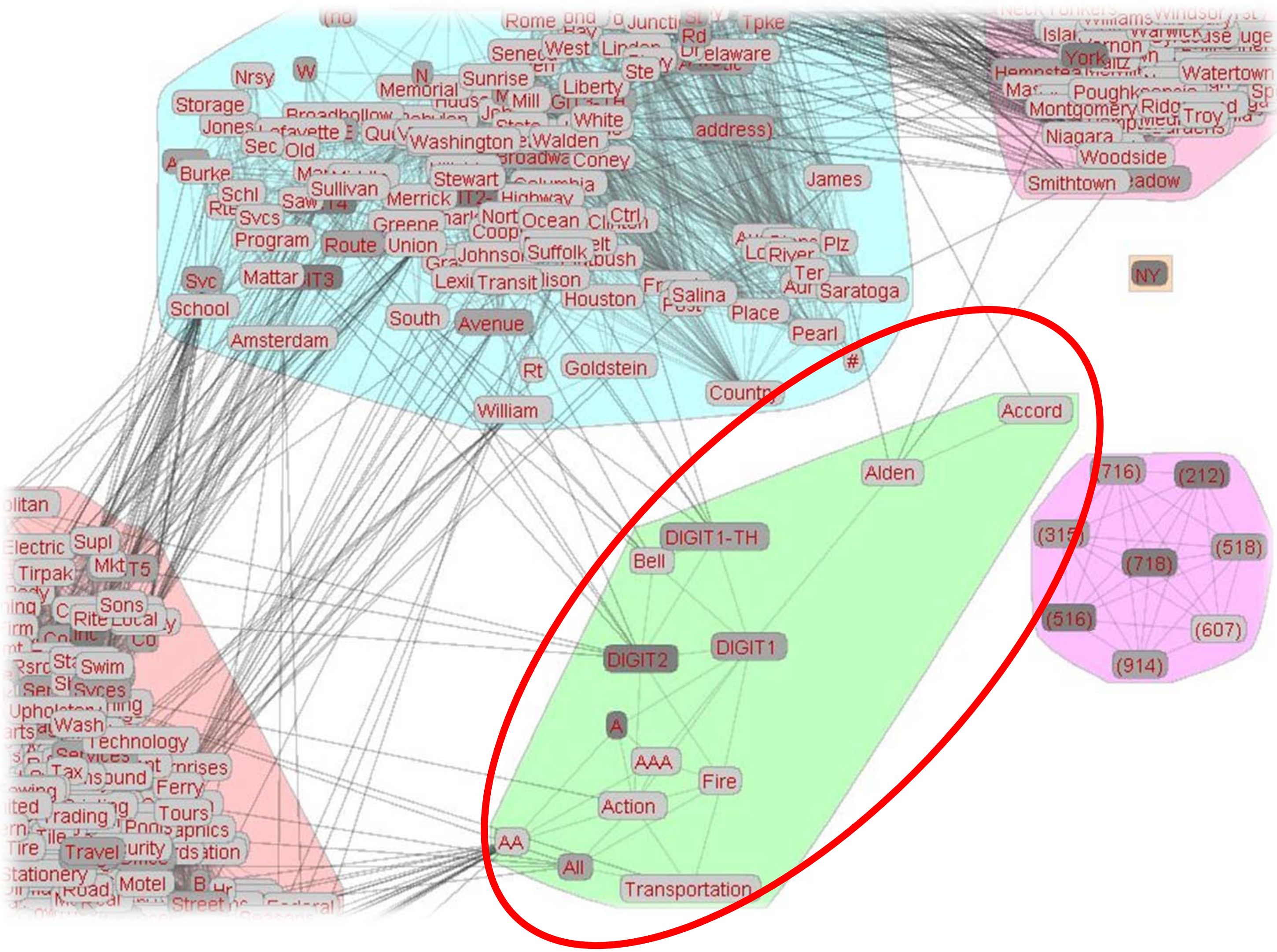
If the clusters don't lend themselves to categories

- re-cluster using a different *refinement* level

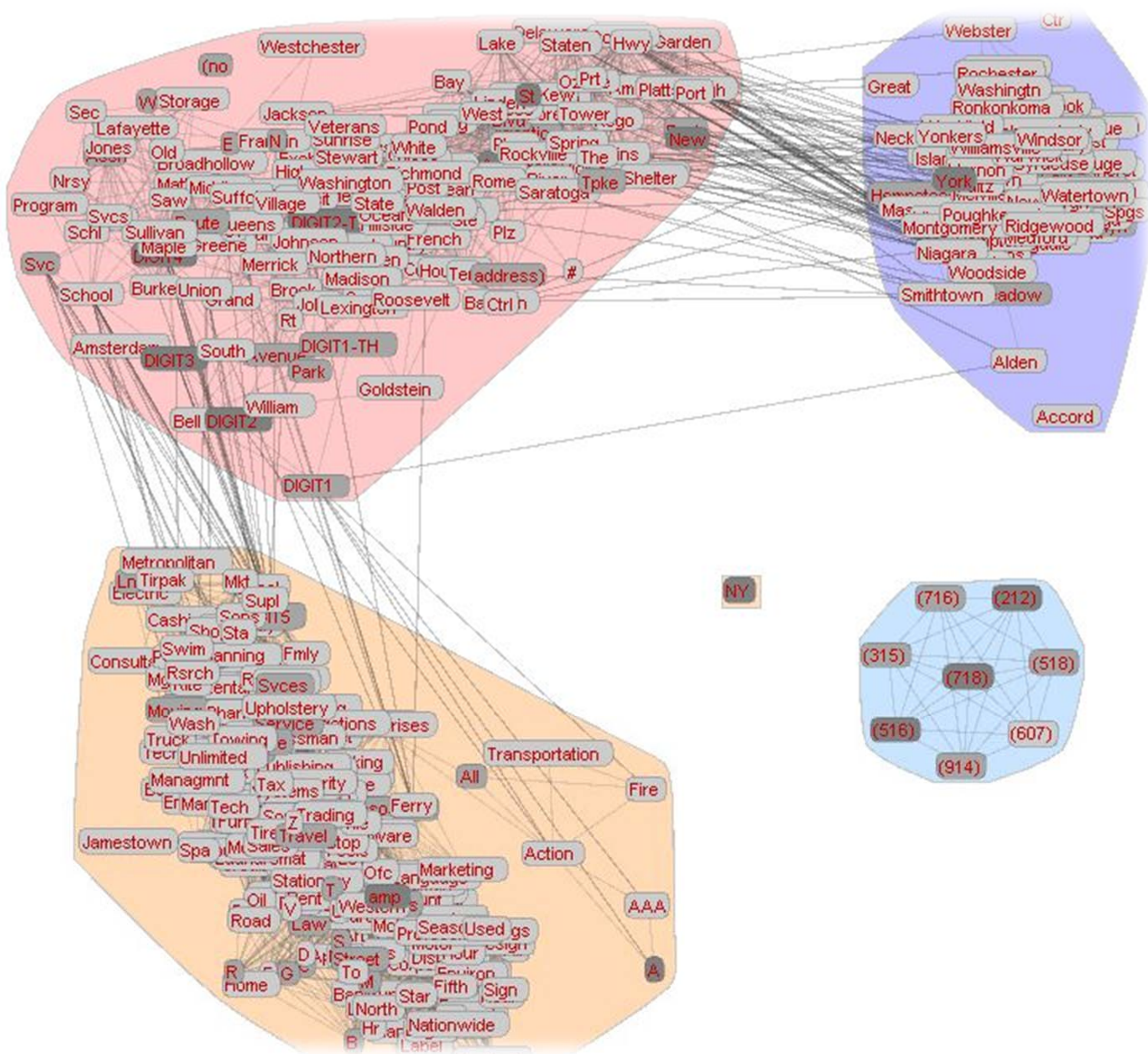
The user can modify the clusters as follows:

- merge clusters
- split clusters
- create a new cluster using nodes from multiple clusters
- name the clusters

Cluster Editing



Cluster Editing



Debugging

Show entries with ambiguously labeled tokens

This involves tokens that:

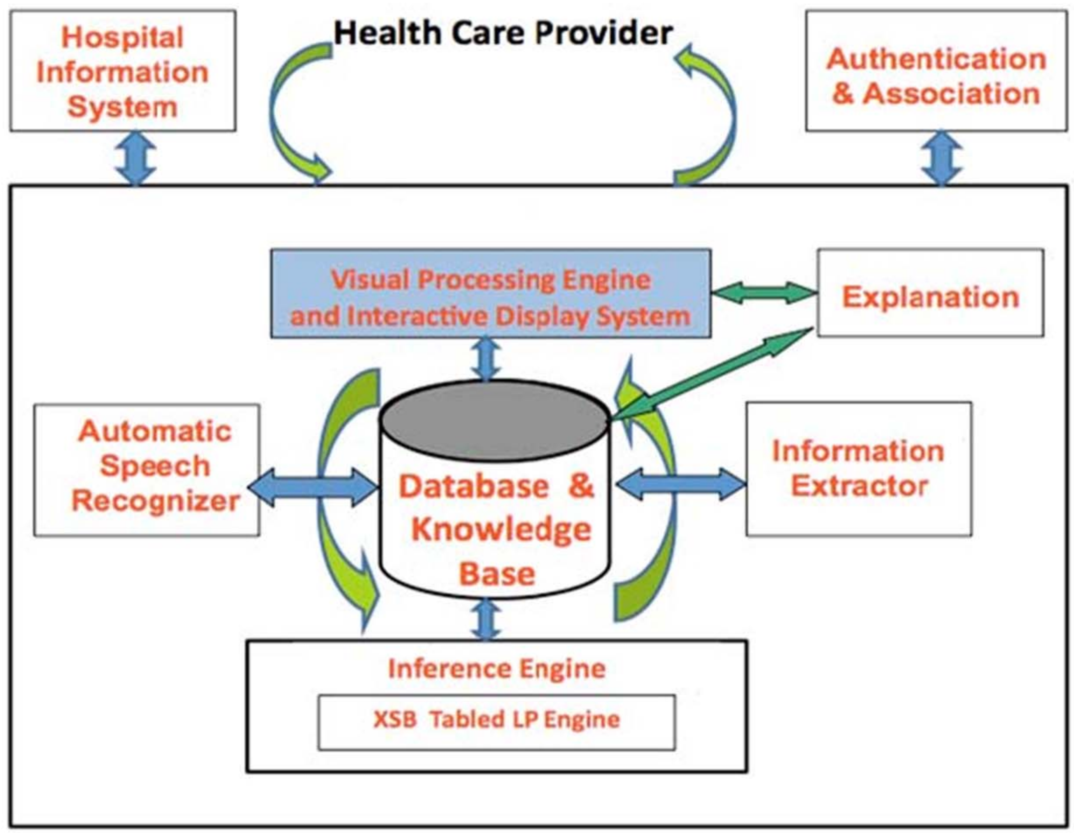
- belong to multiple categories
- occur on border of 2 categories

The visualization steps through the entry showing the class assigned to each token

Current Work

Application to Health Analytics

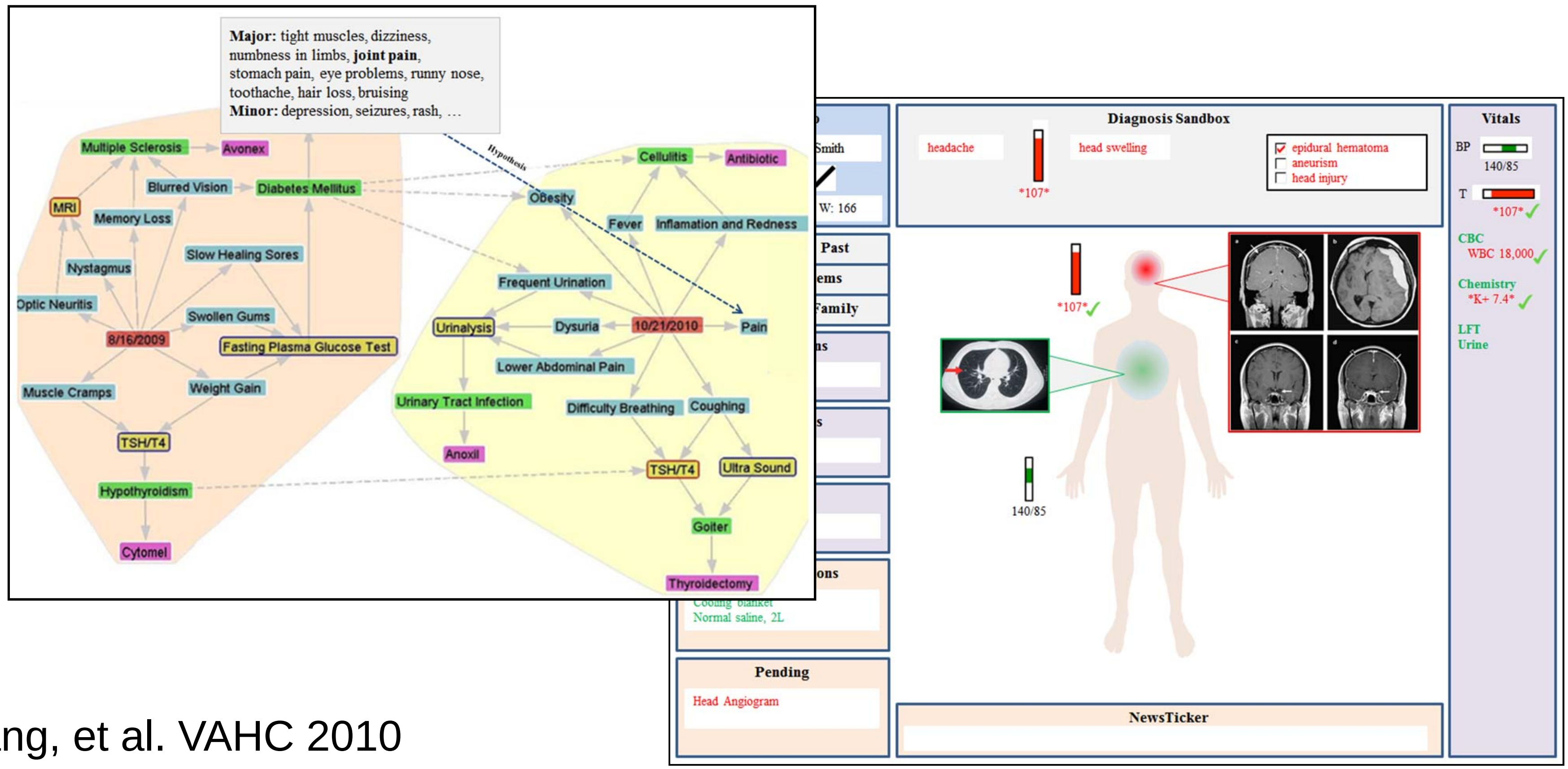
- decision support for emergency room physicians



Current Work

Application to Health Analytics

- decision support for emergency room physicians



Zhang, et al. VAHC 2010

Thanks

Support from NSF, NIH, DOE, BNL, PNL, CEWIT

Collaborators:

- Dr. Alla Zelenyuk, Dr. Dan Imre (formerly BNL, now PNL)
- Dr. IV Ramakrishan (Stony Brook University)
- Dr. Kevin McDonnell (Dowling College)

MS/PhD Students

- Peter Imrich, Yiping Han, Julia EunJu Nam, Supriya Garg, Hyunjung Lee, Zhiyuan Zhang

More information at http://www.cs.sunysb.edu/~mueller