Hybrid Approach to Data Analysis

Eli T. Brown

Shameless Plug!

Remco Chang's lab:

[v]alt
Visual Analytics Laboratory @ Tufts

- Visualization (COMP 150-VIZ)
- Visual Analytics (COMP 150-VAN)

Motivation

Please **help me** with
my **complicated data!**

Machine Learning

- Know labels: supervised learning (classifiers)
- Know nothing: unsupervised learning (clustering, regression, dimension reduction)

- Hopefully we know a little – semisupervised
  – Provide some labels
  – Provide constraints

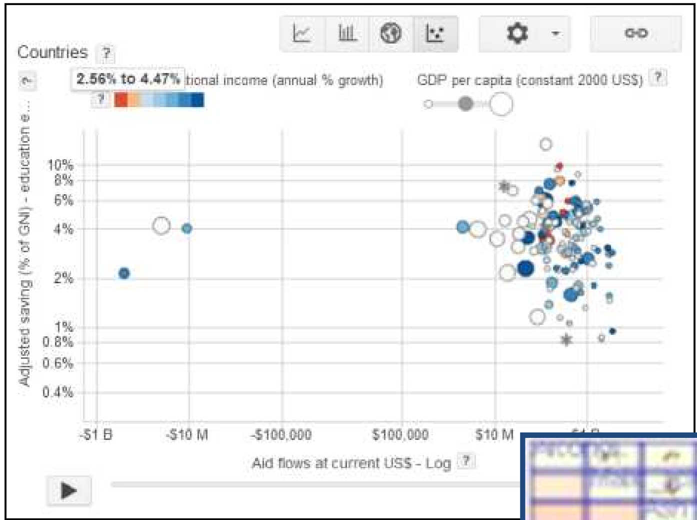
Semi-supervised Learning

- Clustering
  - COP-KMEANS (Wagstaff, et al., 2001)
  - Early separation from just building clusters (Keim, et al., 2002)

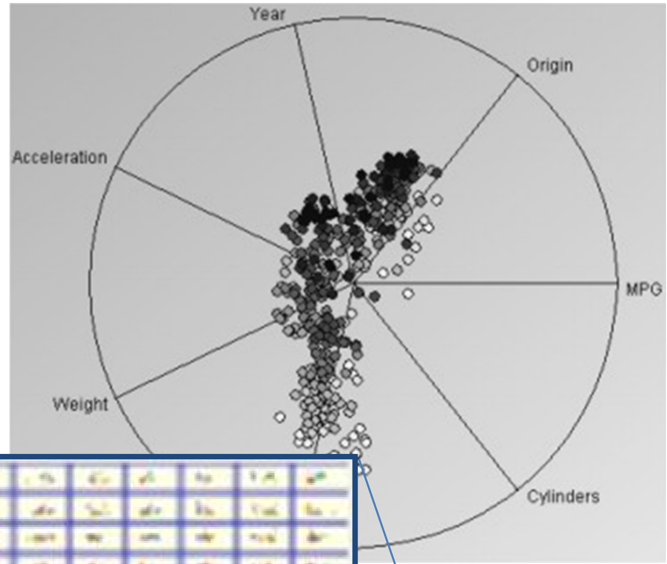
- Metric
  - Learning Mahalanobis distance metrics (Xing, et al., 2002)
    - For PSD matrix *A*: $D_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A (x_i - x_j)}$
  - Newer work has been speed improvements, accuracy analysis, applications, extensions like grouping constraints
  - Relevant Component Analysis (RCA) (Bar-Hillel, et al., 2005)
  - Distance metric learning with eigenvalue optimization (Ying and Li, 2012)

NB: Combined approach and discussion in Integrating constraints and metric learning in semi-supervised clustering (Bilenko, et al., 2004)

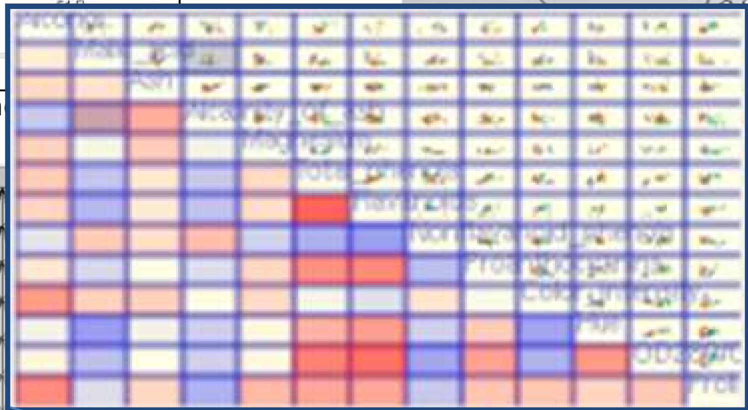
High-Dimensional Visualization

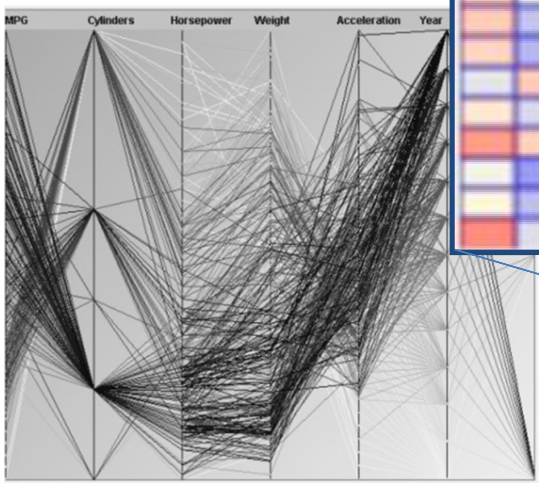


Google Data, recent screensh

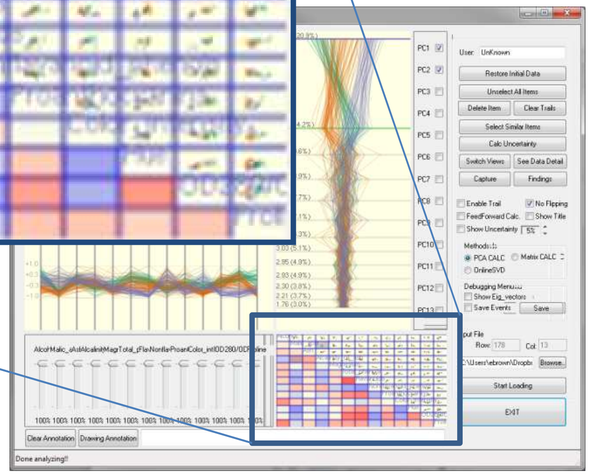


man et al. 1999





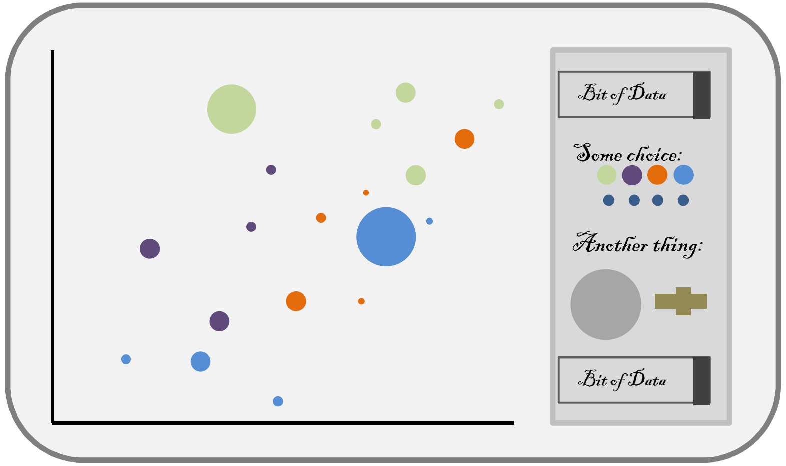
Parallel Coordinates
A. Inselberg et al. 1990



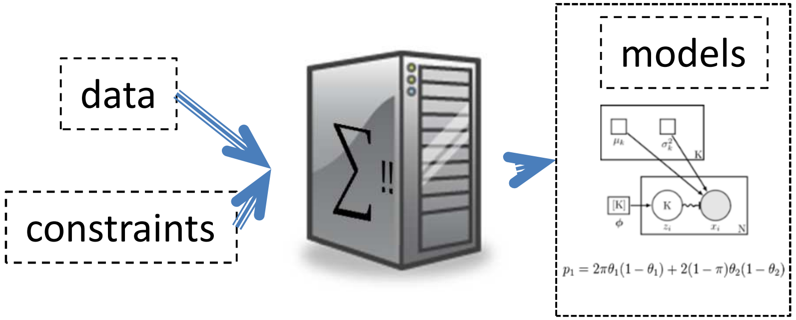
iPCA by D. Jeong, et al. 2009

Approach

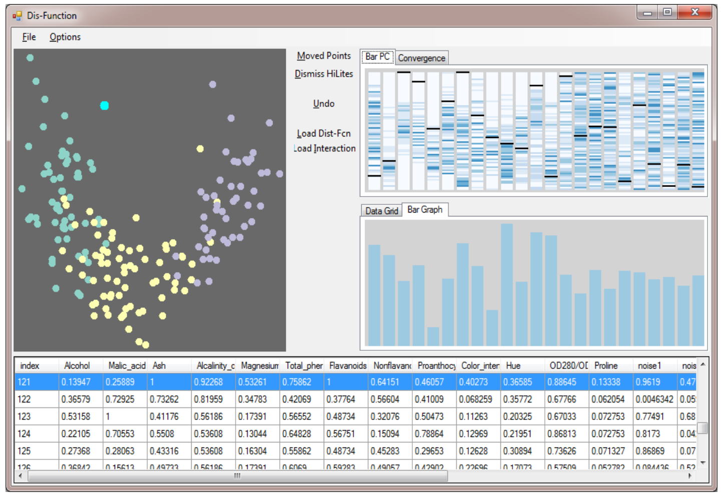
Visualization



Machine Learning



Hybrid



Dis-Function (Brown, 2012)

Visualization w/ Machine Learning

Manipulate visualization parameters
- iPCA (D. Jeong, et al., 2009)
- iVisClassifier (Choo, et al., 2010)

Interacting with Machine Learning Parameters
- Interactive MDS (Buja, et al., 2004)
- Interactive Tool… for Constrained Clustering (Okabe and Yamada, 2011)

Cluster building
- ClusterSculptor (Nam, et al., 2007)
- iCluster (Drucker, et al., 2011)

Spatial Metaphors
- ForceSPIRE (Endert, et al., 2012)
- Object Level Interaction (Endert, et al., 2011)
- Interactive Visual Clustering (desJardins, et al., 2007)
- Object-Centered MDS (Broekens, et al., 2006)

My Work:

Helping users interactively explore, analyze and understand complex data

- Visualization front end
- Machine learning back-end
- => Model hidden from the user

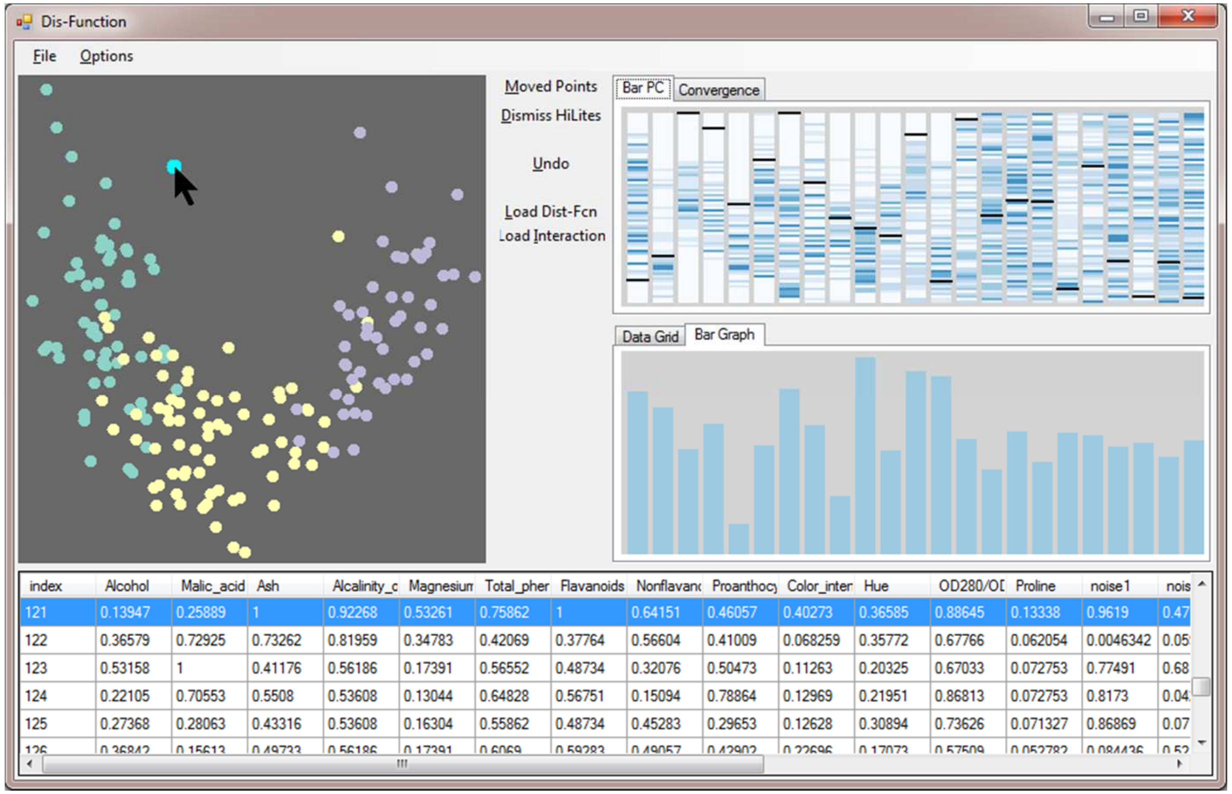
- Learn the “parameters” of the machine learning algorithms from the user’s interactions

As it turns out…

- Learning from interactions not limited to data models (parameters)
- Interactions with software encode users' analysis profile and individual differences
- Using the same scaffolding of interaction capturing, we can simultaneously do:
  - Learning of a data model
  - Learning of a model of the user

Dis-Function:

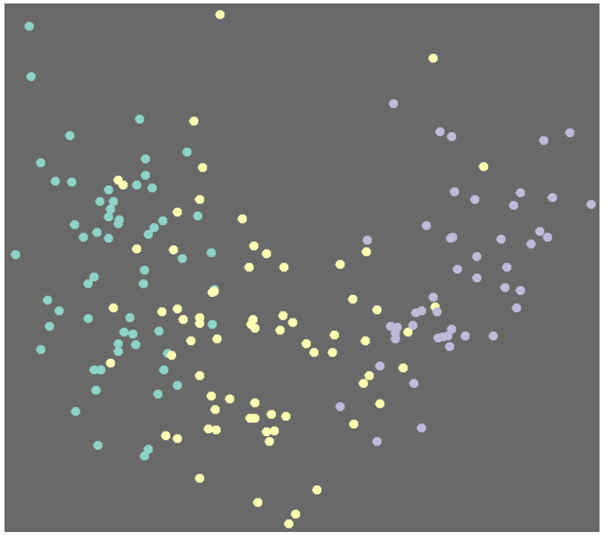
Learning Distance Functions Interactively*



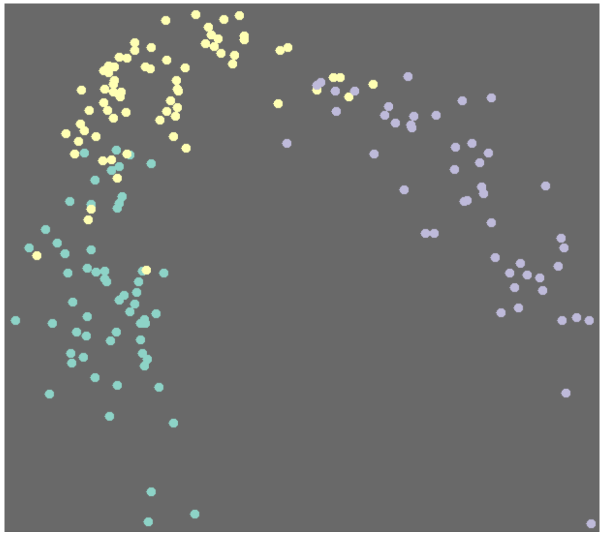
*Brown, Eli T., Liu, J., Brodley, C. E. and Chang, R. Dis-Function: Learning Distance Functions Interactively. In Proceedings of the IEEE VAST (2012)

Weighted Euclidean Distance Functions

$$\sum_{k=1}^{M} (x_{ik} - x_{jk})^2$$

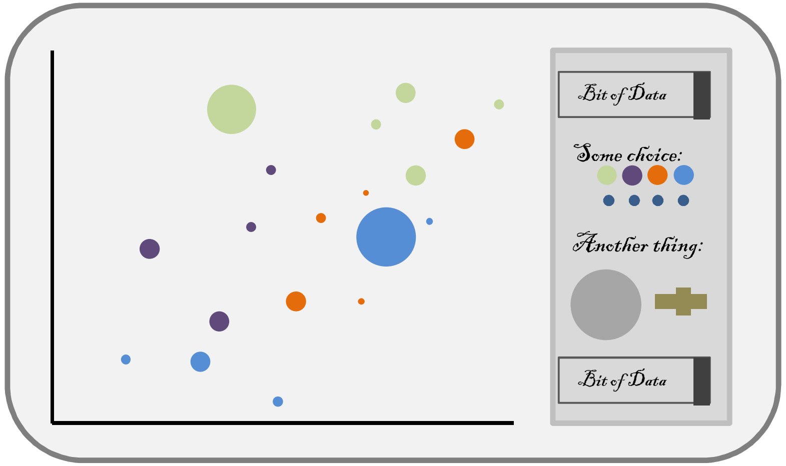


$$\sum_{k=1}^{M} \theta_k (x_{ik} - x_{jk})^2$$

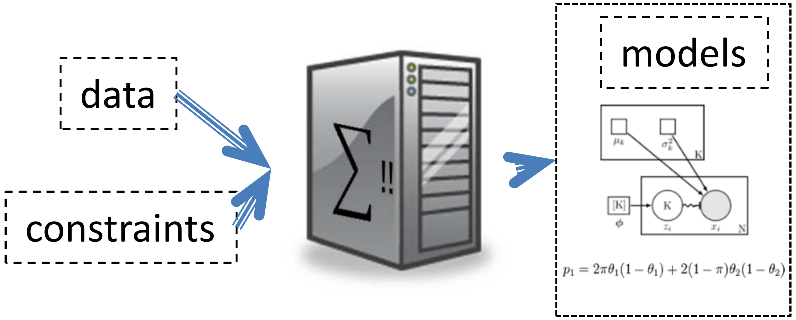


Approach

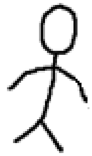
Visualization



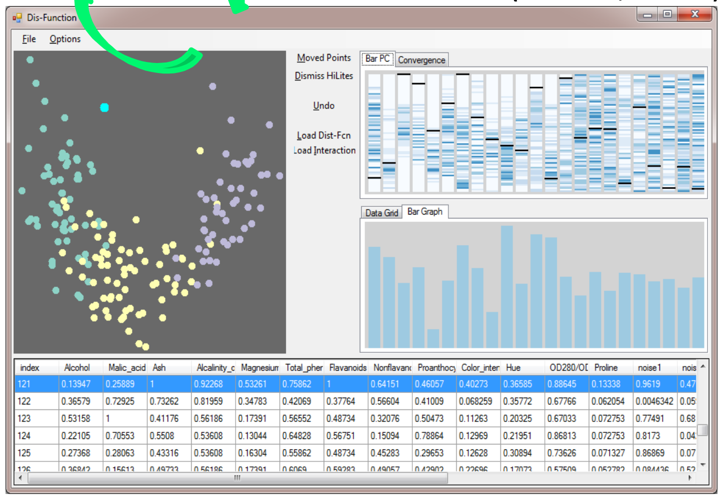
Machine Learning



Dis-Function



Dis-Function (Brown, 2012)



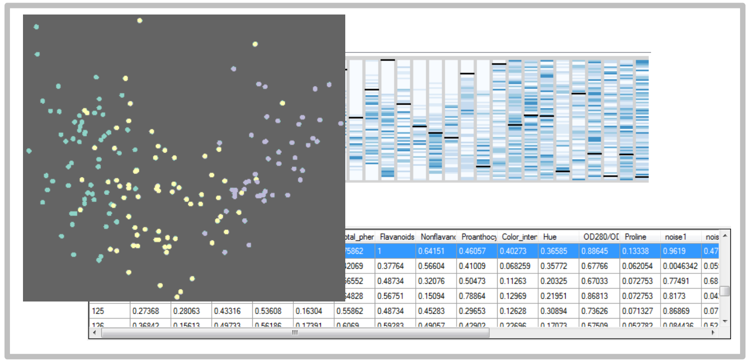
distance function

$D(x, y, \theta)$

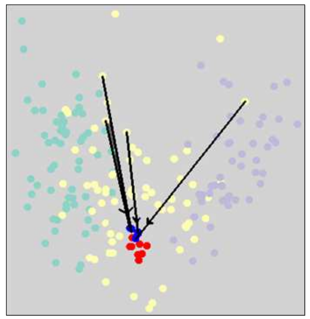
Iterative Process:

1. Start: standard distance function (all $\Theta_k = 1/M$)

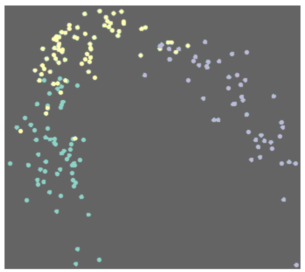
2. Discover feedback opportunities



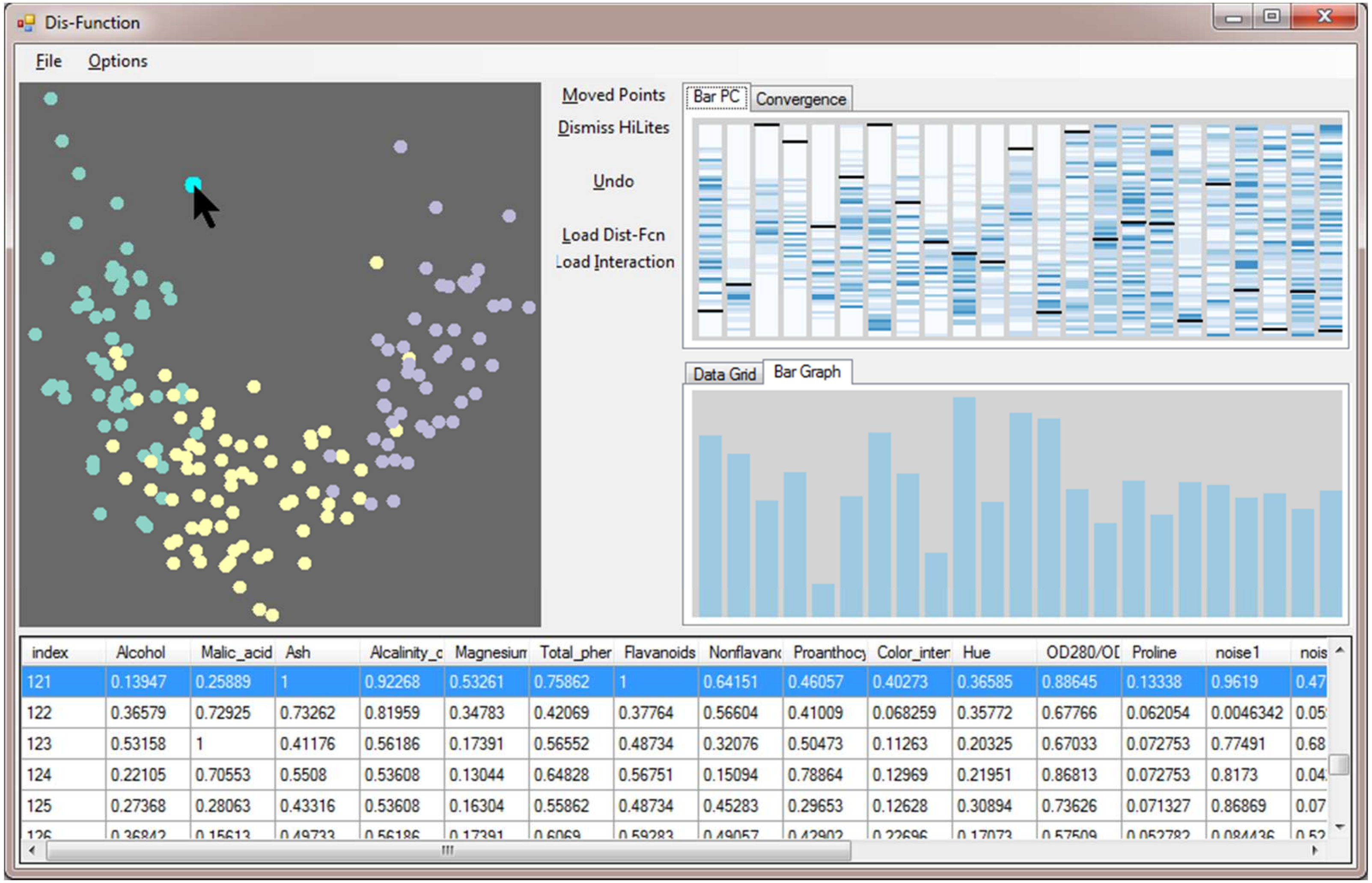
3. Provide feedback by directly moving points



4. Dis-Function makes new Θ , new projection



DEMO


Dis-Function
File
Options
Moved Points
Dismiss HiLites
Undo
Load Dist-Fcn
Load Interaction
Bar PC
Convergence
Data Grid
Bar Graph
index Alcohol Malic_acid Ash Alcalinity_c Magnesiur Total_pher Flavanoids Nonflavanc Proanthocy Color_inter Hue OD280/OI Proline noise1 nois
121 0.13947 0.25889 1 0.92268 0.53261 0.75862 1 0.64151 0.46057 0.40273 0.36585 0.88645 0.13338 0.9619 0.47
122 0.36579 0.72925 0.73262 0.81959 0.34783 0.42069 0.37764 0.56604 0.41009 0.068259 0.35772 0.67766 0.062054 0.0046342 0.05
123 0.53158 1 0.41176 0.56186 0.17391 0.56552 0.48734 0.32076 0.50473 0.11263 0.20325 0.67033 0.072753 0.77491 0.68
124 0.22105 0.70553 0.5508 0.53608 0.13044 0.64828 0.56751 0.15094 0.78864 0.12969 0.21951 0.86813 0.072753 0.8173 0.04
125 0.27368 0.28063 0.43316 0.53608 0.16304 0.55862 0.48734 0.45283 0.29653 0.12628 0.30894 0.73626 0.071327 0.86869 0.07

Interaction:

$Y_1 = \{x_{14}, x_{17}, x_{121}, x_{145}\}$ (blue points)

$Y_2 = \{x_{25}, x_{28}, x_{44}, x_{45}, x_{48}, x_{119}, x_{158}, x_{159}\}$ (red points)

$intended\_distance = \left(x_i^t - x_j^t\right)^2$

$original\_distance = \left(x_i^{t-1} - x_j^{t-1}\right)$

x_i^t indicates point x_i (in projected space) at time t (after user moved it for this round)

$$U_{ij} = \begin{cases} \dfrac{intended\_distance}{original\_projected\_distance} & \text{if } (x_i, x_j) \in Y_1 \times Y_2, \\ 1 & \text{otherwise.} \end{cases}$$

Tells how much user has moved each pair of points (x_i and x_j)

Updating Distance Function:

Optimization to pick new Θ:

$$\boxed{\Theta^t = \arg\min_{\Theta^t} \sum_{i<j\leq N} L_{ij}^t \left(D(x_i, x_j|\Theta^t) - U_{ij}^t \cdot D(x_i, x_j|\Theta^{t-1}) \right)^2}$$

N, M	Number of Dimensions, Points
x_i	Point i of the data
x_{ij}	Value of feature k of data point x_i
Θ^{t-1}, Θ^t	Vectors containing the weight of each dimension for a distance function after or before (resp.) an interaction step
θ_k	Weight of feature k in Θ
$D(x_i, x_j\|\Theta)$	Distance between x_i and x_j with parameter Θ (dimension weight vector)

$$L_{ij}^t = \begin{cases} \frac{N(N-1)}{|Y_1^t||Y_2^t|} - 1 & \text{if } (x_i, x_j) \in Y_1^t \times Y_2^t, \\ 1 & \text{otherwise.} \end{cases}$$

Scalar weight emphasizing terms of the sum related to user input

$$D(x_i, x_j|\Theta) = \sum_{k=1}^{M} \theta_k (x_{ik} - x_{jk})^2$$

Updating Distance Function:

Optimization to pick new Θ:

$$\boxed{\Theta^t = \arg\min_{\Theta^t} \sum_{i<j\leq N} L_{ij}^t \left(D(x_i, x_j|\Theta^t) - U_{ij}^t \cdot D(x_i, x_j|\Theta^{t-1}) \right)^2}$$

Pick a new Θ to minimize,

Updating Distance Function:

Optimization to pick new Θ:

$$\Theta^t = \arg\min_{\Theta^t} \sum_{i<j\le N} L_{ij}^t \left(D(x_i, x_j | \Theta^t) - U_{ij}^t \cdot D(x_i, x_j | \Theta^{t-1}) \right)^2$$

Pick a new Θ to minimize,
Over the sum of all points

Updating Distance Function:

Optimization to pick new Θ:

$$\Theta^t = \arg\min_{\Theta^t} \sum_{i<j\leq N} L^t_{ij} \left(D(x_i, x_j | \Theta^{\boxed{t}}) - U^t_{ij} \cdot D(x_i, x_j | \Theta^{\boxed{t-1}}) \right)^2$$

Pick a new Θ to minimize,
Over the sum of all points,
The change in total distance
between those points
before and after update

Updating Distance Function:

Optimization to pick new Θ:

$$\boxed{\Theta^t = \arg\min_{\Theta^t} \sum_{i<j\leq N} L^t_{ij}\left(D(x_i, x_j|\Theta^t) - U^t_{ij}\cdot D(x_i, x_j|\Theta^{t-1})\right)^2}$$

Pick a new Θ to minimize,
Over the sum of all points,
The change in total distance
between those points
before and after update,
Only when signaled by user input

Updating Distance Function:

Optimization to pick new Θ:

$$\boxed{\Theta^t = \arg\min_{\Theta^t} \sum_{i<j\leq N} L^t_{ij} \left(D(x_i, x_j | \Theta^t) - U^t_{ij} \cdot D(x_i, x_j | \Theta^{t-1}) \right)^2}$$

Pick a new Θ to minimize,
Over the sum of all points,
The change in total distance
  between those points
  before and after update,
Only when signaled by user input

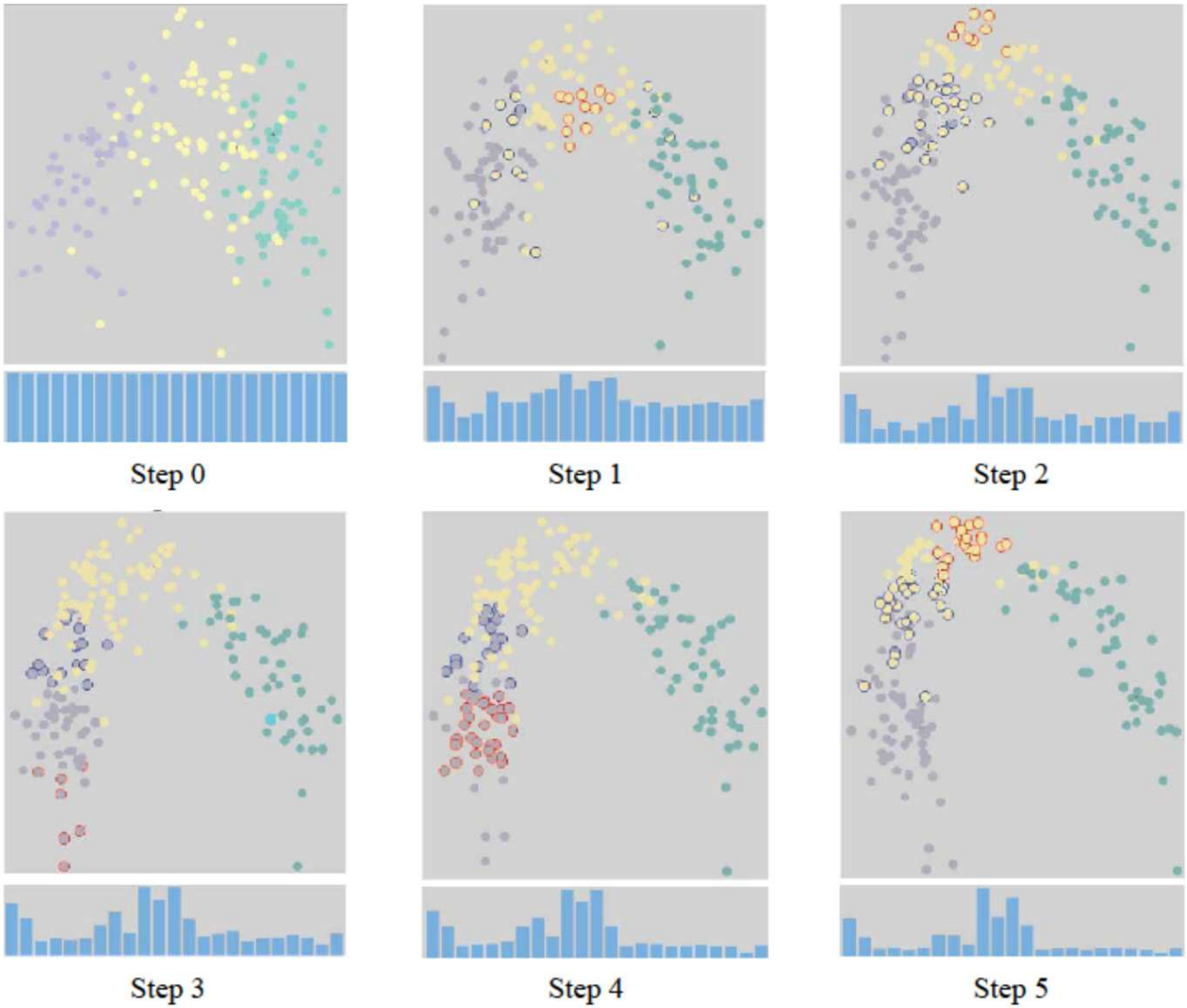
Experiments:

1. 10 participants used software
   - Simulated expertise by coloring points by true class

2. Augmented dataset with 10 dimensions filled with uniformly distributed noise in the same range as the data

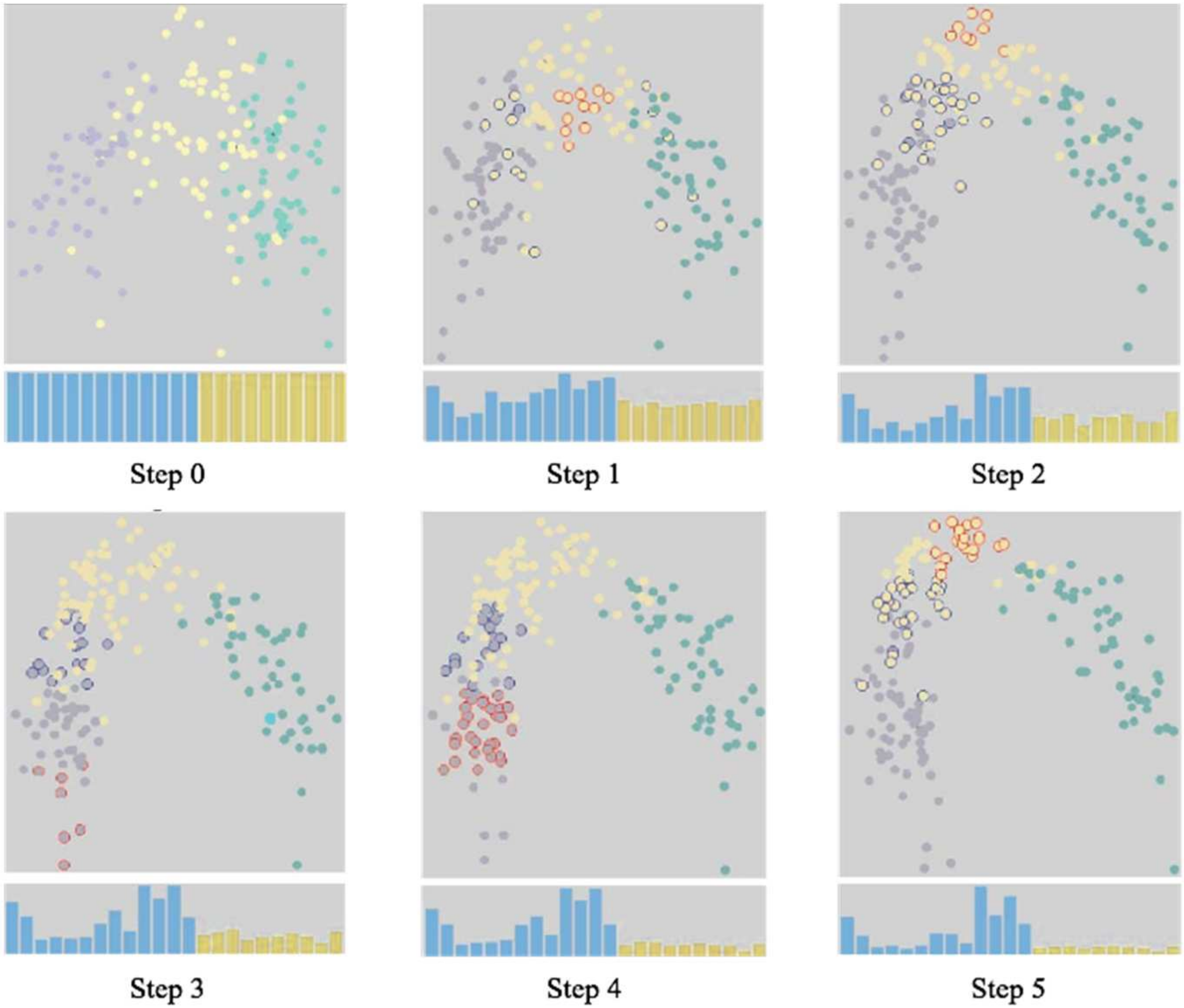
3. Used *k*-NN to evaluate resulting distance functions
   - Evaluated for classification under leave-one-out cross-validation for *k* = 1, 3, 5, 7

Example Set of Interactions:



Projections from successive iterations of user feedback with Wine dataset.
Corresponding bar graph shows weight of each dimension's coefficient.

Example Set of Interactions:



Projections from successive iterations of user feedback with Wine dataset.
Corresponding bar graph shows weight of each dimension's coefficient.

Experiments:

1. 10 participants used software
   - Simulated expertise by coloring points by true class

2. Augmented dataset with 10 dimensions filled with uniformly distributed noise in the same range as the data

3. Used *k*-NN to evaluate resulting distance functions
   - Evaluated for classification under leave-one-out cross-validation for $k = 1, 3, 5, 7$

Results:

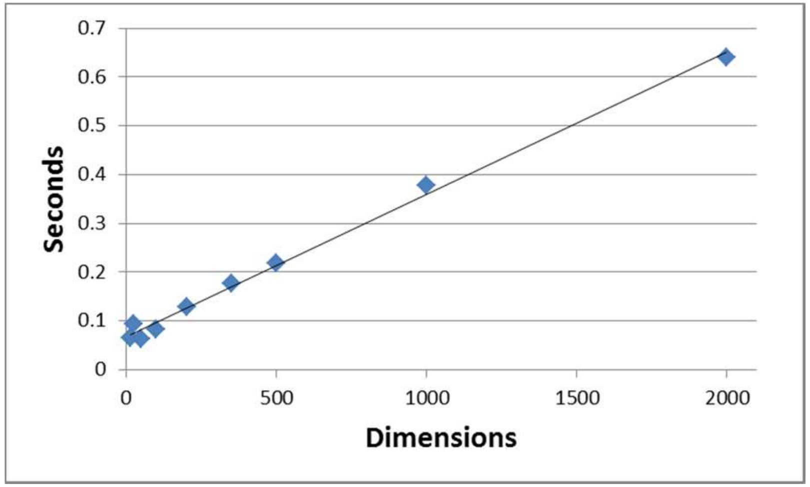
Distance Function	Accuracy*
Even Weights	90 %
User 1	97 %
User 2	94 %
User 3	94 %
User 4	96 %
User 5	97 %
User 6	96 %
User 7	96 %
User 8	96 %
User 9	96 %
User 10	97 %

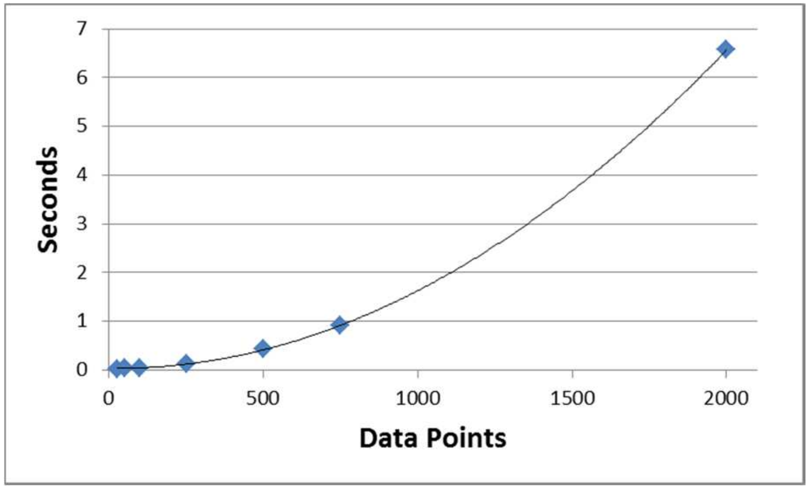
* K Nearest Neighbor accuracy, average of k= 1, 3, 5, 7; Leave-one-out cross validation

Empirical Runtime Performance:

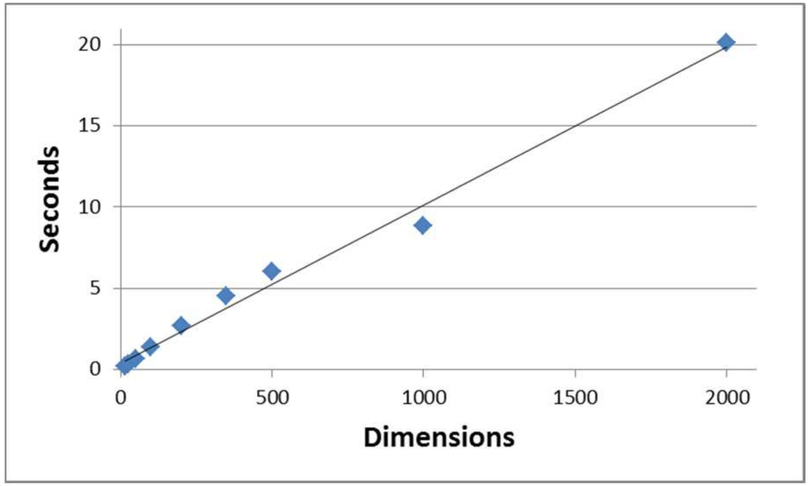
Complexity (#dimensions) | **Size (#instances)**

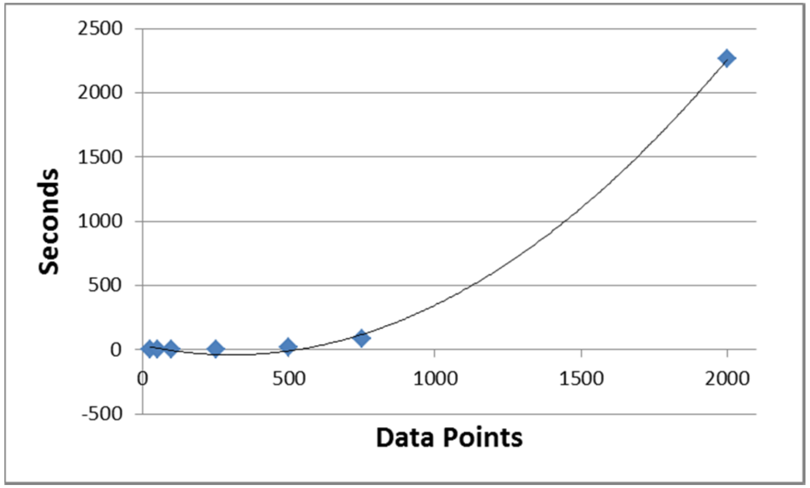
Optimization





Projection





Dis-Function Summary:

- Can learn a simple human- and computer-readable representation of an expert's knowledge about data
- Distance functions learned here are usable in many machine learning applications (including k-NN)
- Can read importance of dimensions off distance function, which helps user understanding
- Performance is promising for small data

Future Work:

- High Dimensional Understanding and Interaction
  - Different representations, new interaction semantics
  - Non-real-valued data
  - Traditional tools: filtering, faceted search, RadVis, parallel coordinates
  - Topological, hierarchical views
- Active Learning
  - Intelligently help user discover points of interest (maybe useful ones for learning distance function)
- Applications:
  - MS patient data
  - Phylogenetic trees

Future Work:

- High Dimensional Understanding and Interaction
    - Different representations, new interaction semantics
    - Non-real-valued data
    - Traditional tools: filtering, faceted search, RadVis, parallel coordinates
    - Topological, hierarchical views

- Active Learning
    - Intelligently help user discover points of interest (maybe useful ones for learning distance function)

- **Applications:**
    - **MS patient data**
    - **Phylogenetic trees**

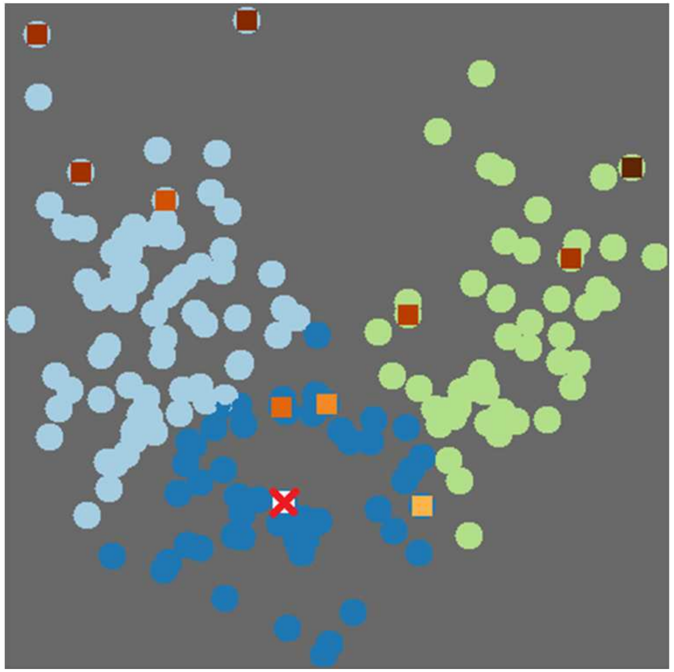
Future Work:

- High Dimensional Understanding and Interaction
  - Different representations, new interaction semantics
  - Non-real-valued data
  - Traditional tools: filtering, faceted search, RadVis, parallel coordinates
  - Topological, hierarchical views

- **Active Learning**
  - **Intelligently help user discover points of interest (maybe useful ones for learning distance function)**

- Applications:
  - MS patient data
  - Phylogenetic trees

EigenSense



Brown, Eli T and Chang, Remco. *EigenSense: Saving User Effort with Active Metric Learning*. KDD IDEA Workshop 2014.

EigenSense: Motivation

- How does the user know what points to choose?
- Might like points that have the most impact (most effect for effort)

- From machine learning: *active learning*
- Determines which instances are most helpful to the model

EigenSense: Inspiration

- Need an active learning method for our inter-active distance function learning
- Representation of the model: pairwise distance matrix

- Eigenvectors represent asymptotic behavior of matrix transformations
  - PageRank
  - Biology: population models

EigenSense at Work

Algorithm 1: EigenScore

Input: Data points x_i, x_j, distance matrix D

Output: $ES_{ij} \in [0,1]$

Calculate $v_1 = eigs(D,1)$ [dominant eigenvector]

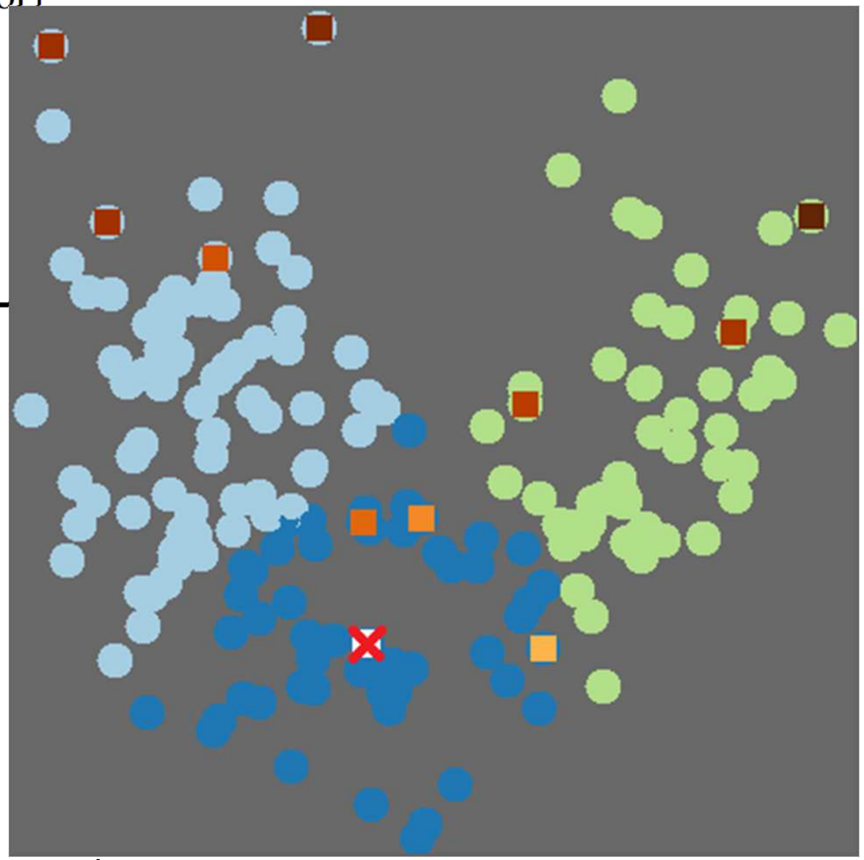
Let $D' = D$

Set $D'_{ij} = D'_{ji} = 0$

Calculate $v'_1 = eigs(D',1)$

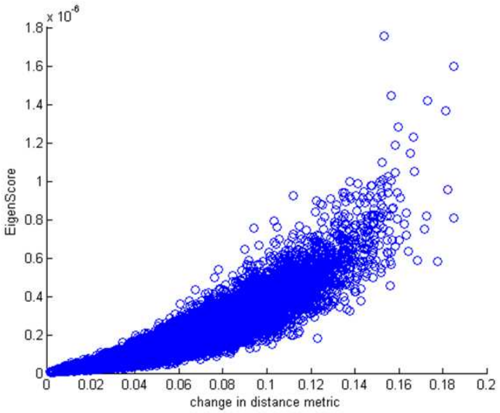
Set $ES_{ij} = 1 - CosineSimilarity(v_1, v'_1)$

return ES_{ij}

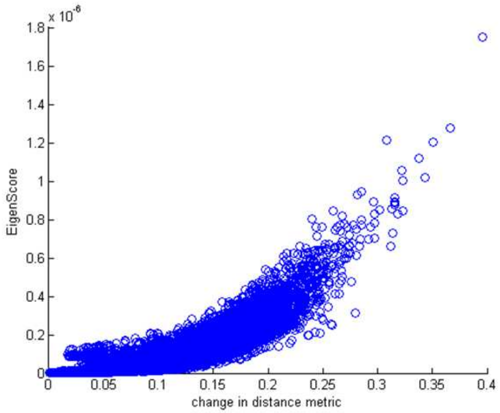


Working in Dis-Function

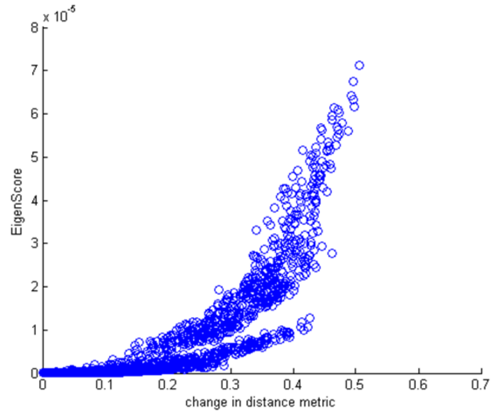
EigenSense Evidence



wine



ion



parkinsons

Future Work

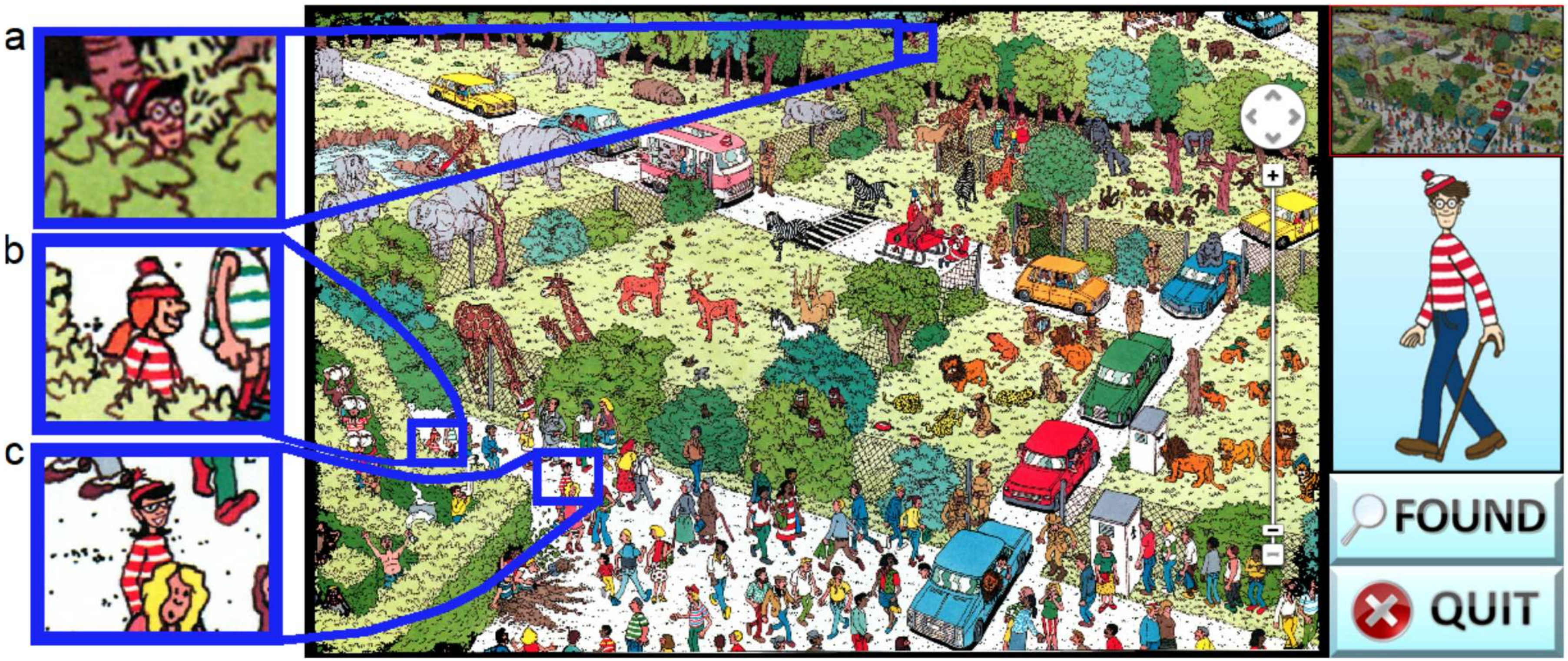
- Evaluation!

- Does this type of guidance actually help people?
- Is our method for choosing guidance a good one?

Learning a User Model

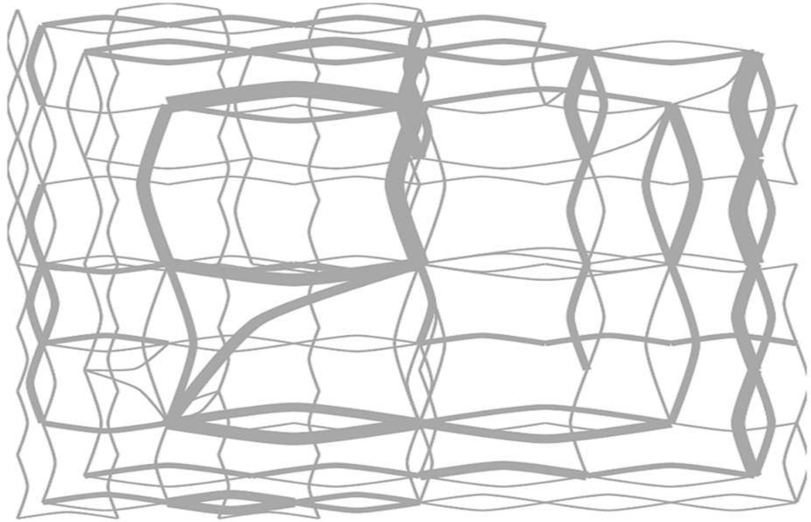
- Finding Waldo (TVCG 2014)
  - User performs a visual search task (Where's Waldo)
  - The system records and learns a user's characteristics
    - Completion time
    - Cognitive Factors (Locus of Control, Big Five Personality)

Waldo Study Interface

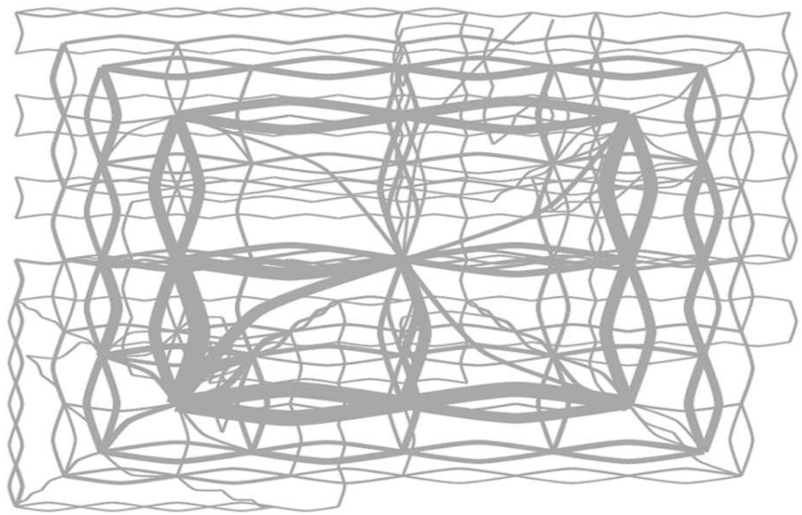


Waldo Results: Completion Time

Fast

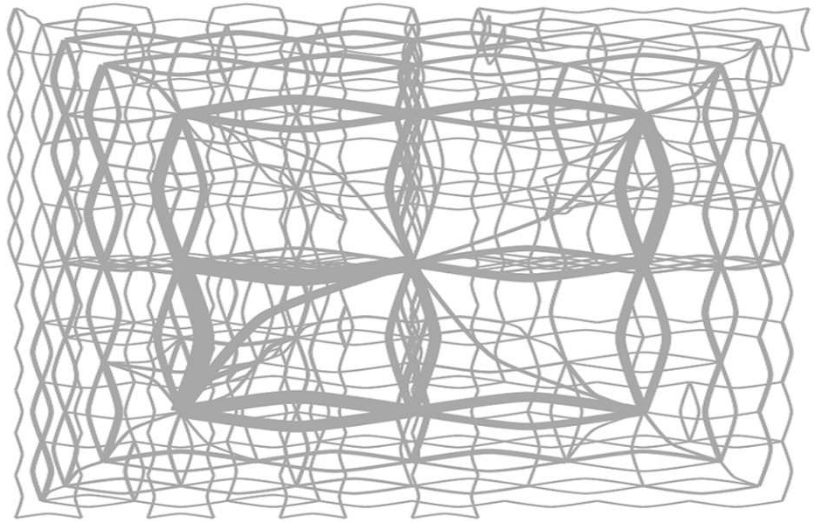


Slow

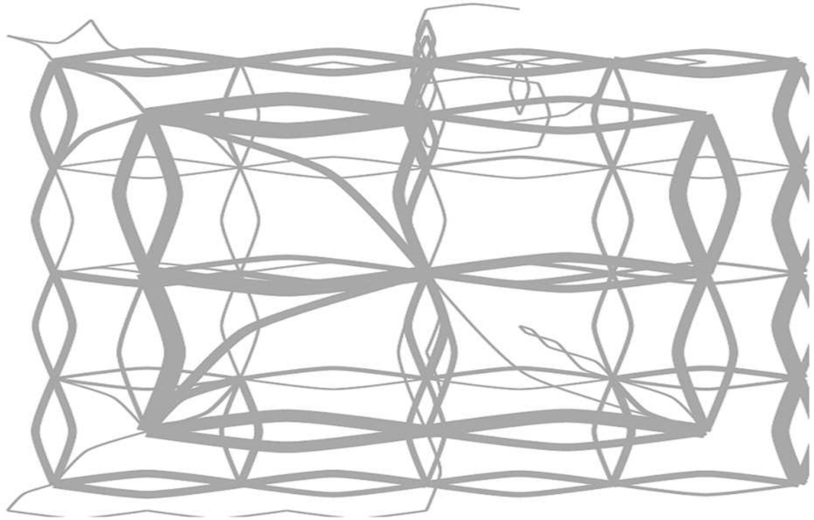


Waldo Results: Personality

External LOC

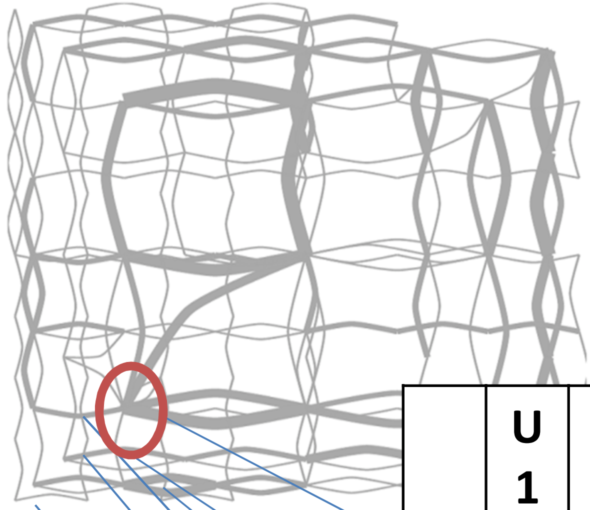


Internal LOC

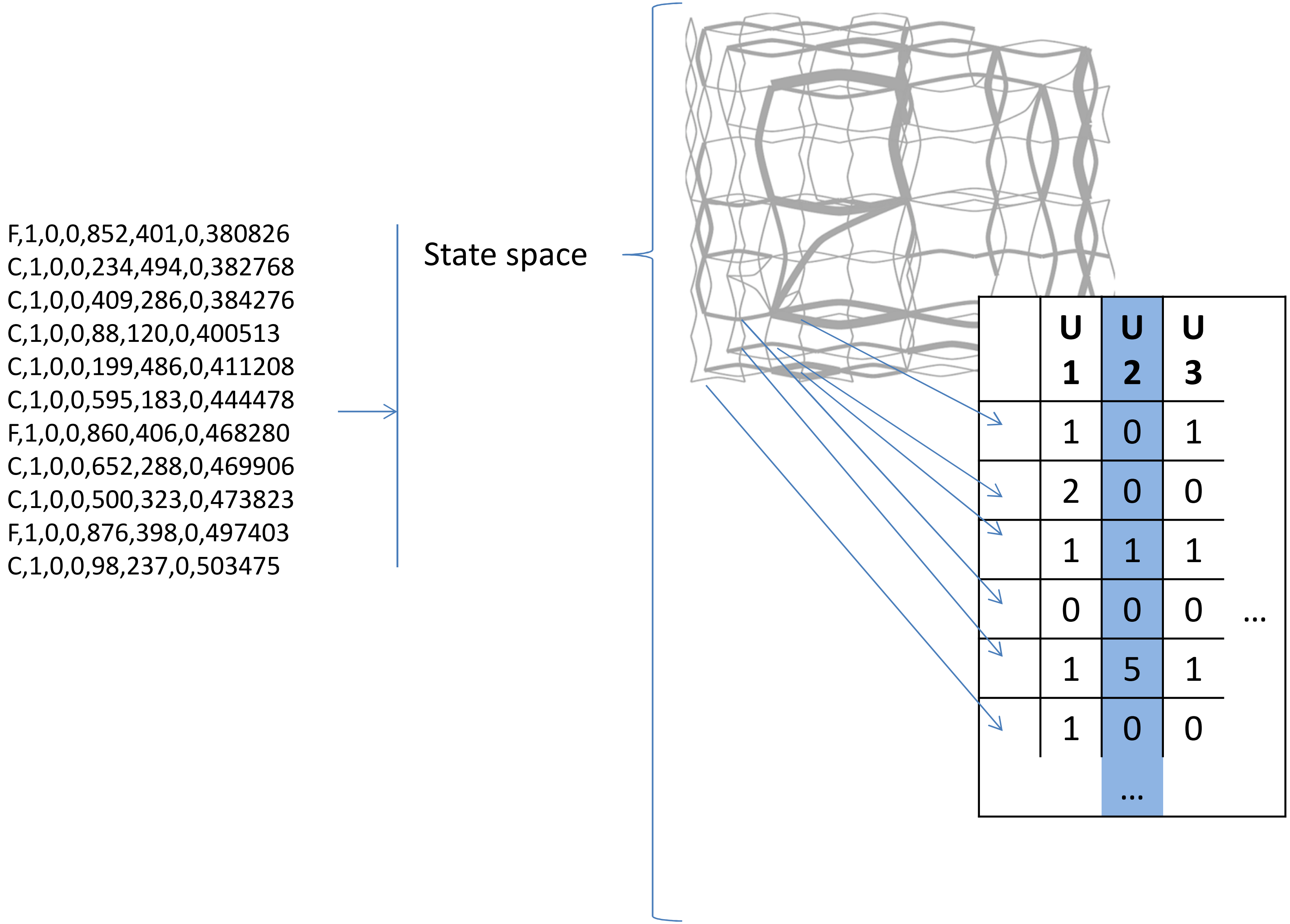


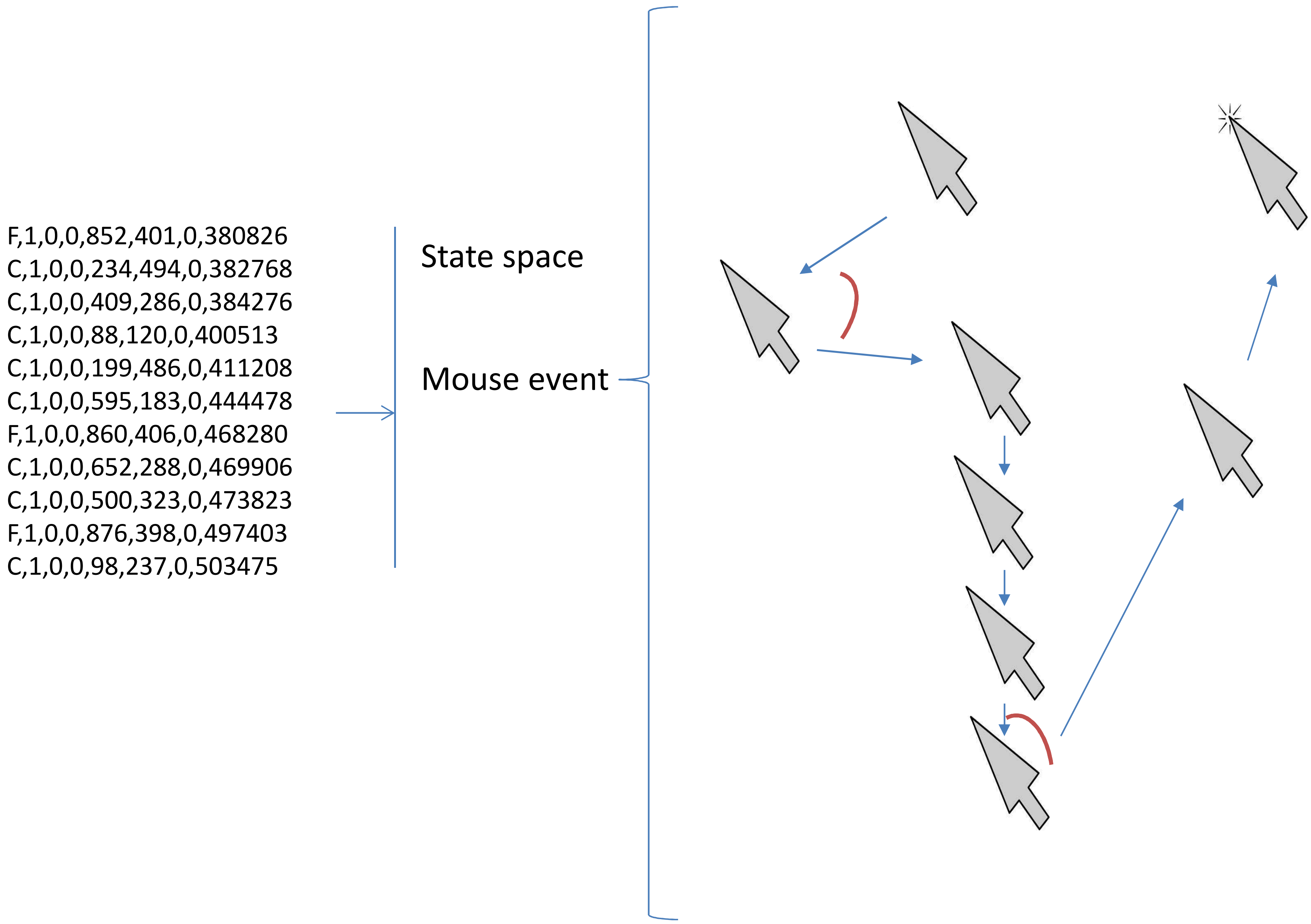
F,1,0,0,852,401,0,380826
C,1,0,0,234,494,0,382768
C,1,0,0,409,286,0,384276
C,1,0,0,88,120,0,400513
C,1,0,0,199,486,0,411208
C,1,0,0,595,183,0,444478
F,1,0,0,860,406,0,468280
C,1,0,0,652,288,0,469906
C,1,0,0,500,323,0,473823
F,1,0,0,876,398,0,497403
C,1,0,0,98,237,0,503475

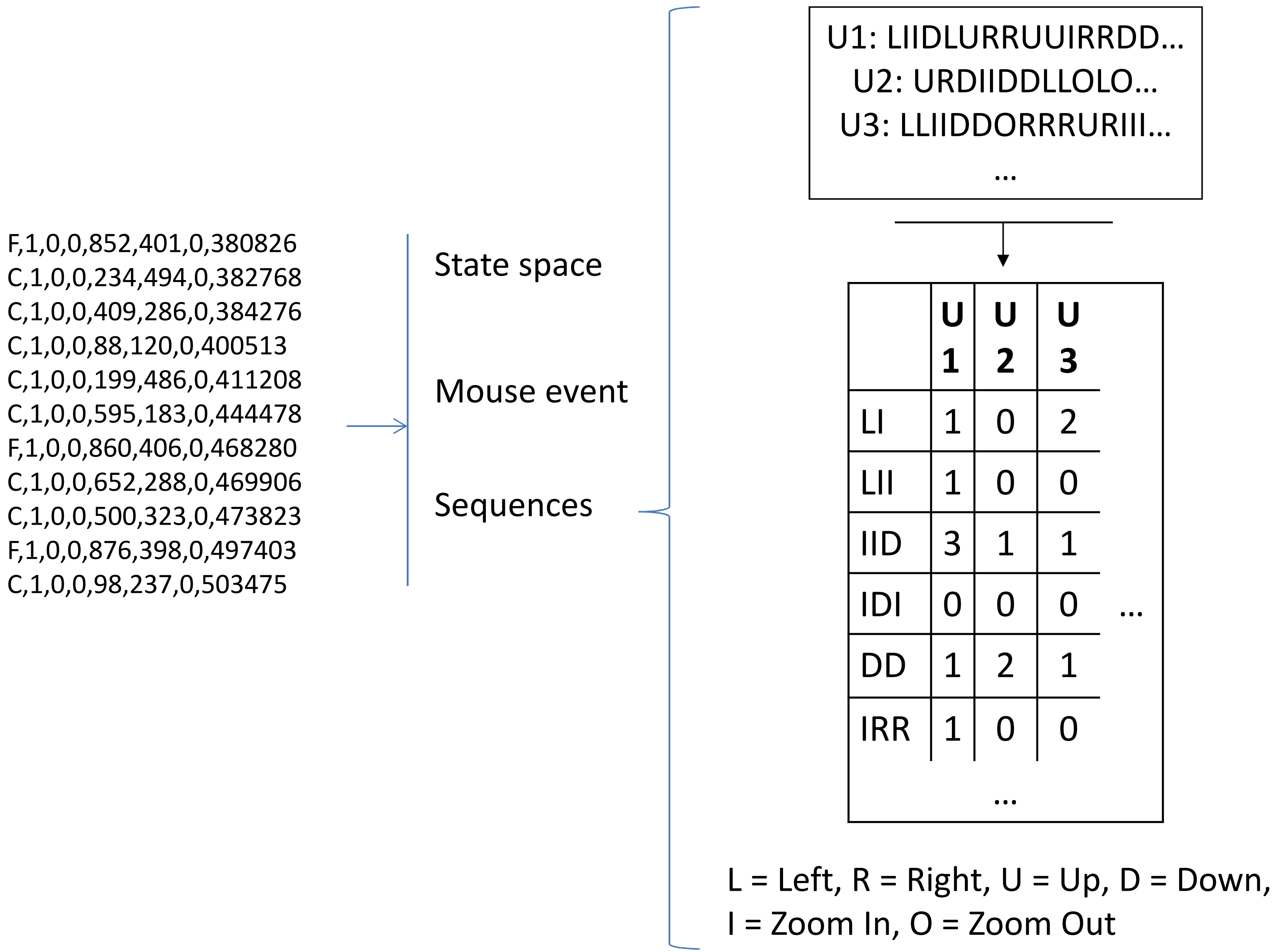
State space



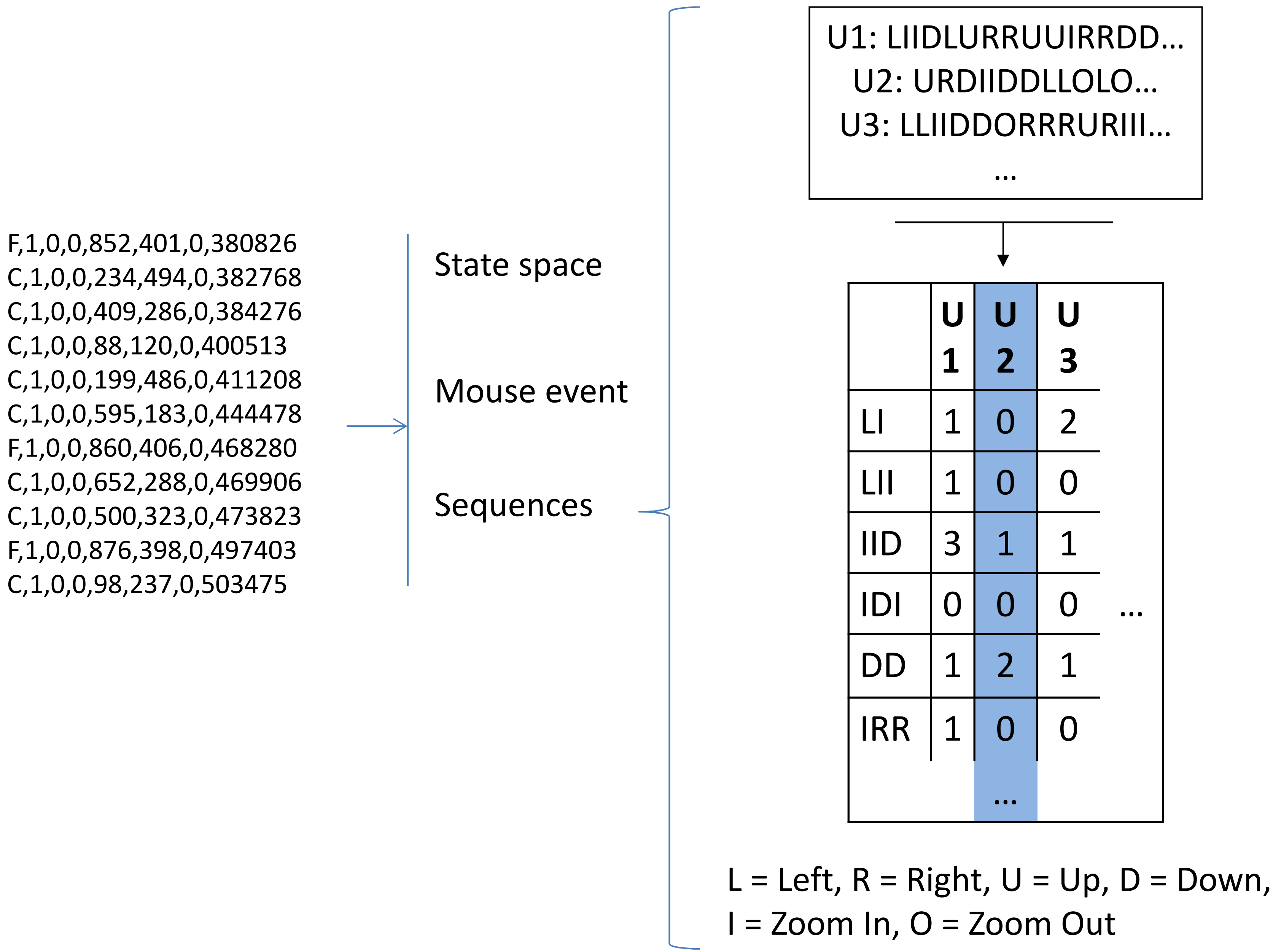
	U 1	**U 2**	**U 3**	
	1	0	1	
	2	0	0	
	1	1	1	
	0	0	0	…
	1	5	1	
	1	0	0	
	…			


F,1,0,0,852,401,0,380826
C,1,0,0,234,494,0,382768
C,1,0,0,409,286,0,384276
C,1,0,0,88,120,0,400513
C,1,0,0,199,486,0,411208
C,1,0,0,595,183,0,444478
F,1,0,0,860,406,0,468280
C,1,0,0,652,288,0,469906
C,1,0,0,500,323,0,473823
F,1,0,0,876,398,0,497403
C,1,0,0,98,237,0,503475
State space
U1 U2 U3
1 0 1
2 0 0
1 1 1
0 0 0 …
1 5 1
1 0 0
…


F,1,0,0,852,401,0,380826
C,1,0,0,234,494,0,382768
C,1,0,0,409,286,0,384276
C,1,0,0,88,120,0,400513
C,1,0,0,199,486,0,411208
C,1,0,0,595,183,0,444478
F,1,0,0,860,406,0,468280
C,1,0,0,652,288,0,469906
C,1,0,0,500,323,0,473823
F,1,0,0,876,398,0,497403
C,1,0,0,98,237,0,503475
State space
Mouse event

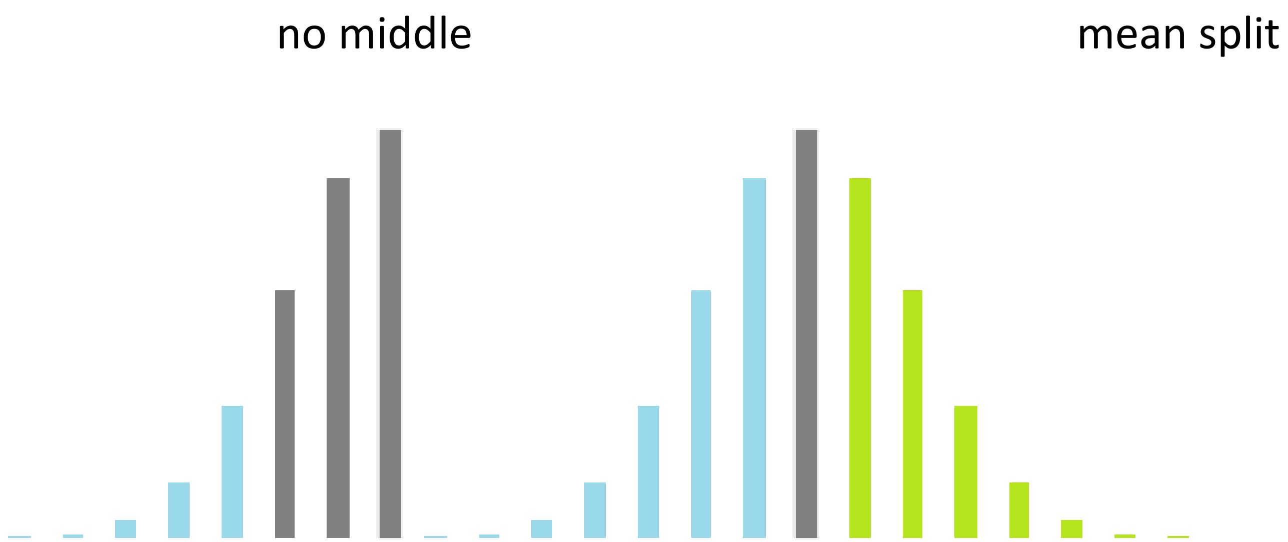


	U1	**U2**	**U3**	
LI	1	0	2	
LII	1	0	0	
IID	3	1	1	
IDI	0	0	0	…
DD	1	2	1	
IRR	1	0	0	
…				


F,1,0,0,852,401,0,380826
C,1,0,0,234,494,0,382768
C,1,0,0,409,286,0,384276
C,1,0,0,88,120,0,400513
C,1,0,0,199,486,0,411208
C,1,0,0,595,183,0,444478
F,1,0,0,860,406,0,468280
C,1,0,0,652,288,0,469906
C,1,0,0,500,323,0,473823
F,1,0,0,876,398,0,497403
C,1,0,0,98,237,0,503475
State space
Mouse event
Sequences
U1: LIIDLURRUUIRRDD…
U2: URDIIDDLLOLO…
U3: LLIIDDORRRURIII…
…
U1
U2
U3
LI 1 0 2
LII 1 0 0
IID 3 1 1
IDI 0 0 0
DD 1 2 1
IRR 1 0 0
…
…
L = Left, R = Right, U = Up, D = Down,
I = Zoom In, O = Zoom Out

Classes

- Supervised learning – need label for each participant
- Predict performance (fast vs. slow)

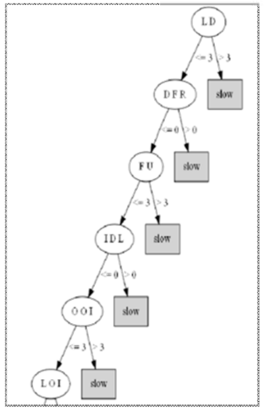


F,1,0,0,852,401,0,380826
C,1,0,0,234,494,0,382768
C,1,0,0,409,286,0,384276
C,1,0,0,88,120,0,400513
C,1,0,0,199,486,0,411208
C,1,0,0,595,183,0,444478
F,1,0,0,860,406,0,468280
C,1,0,0,652,288,0,469906
C,1,0,0,500,323,0,473823
F,1,0,0,876,398,0,497403
C,1,0,0,98,237,0,503475

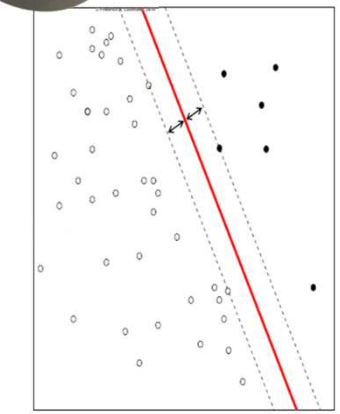
State space

Mouse event

Sequences

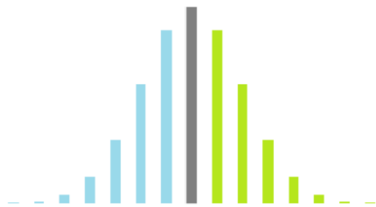


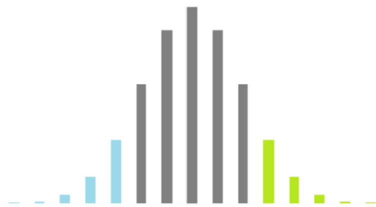
decision tree



SVM

Performance Prediction Accuracy





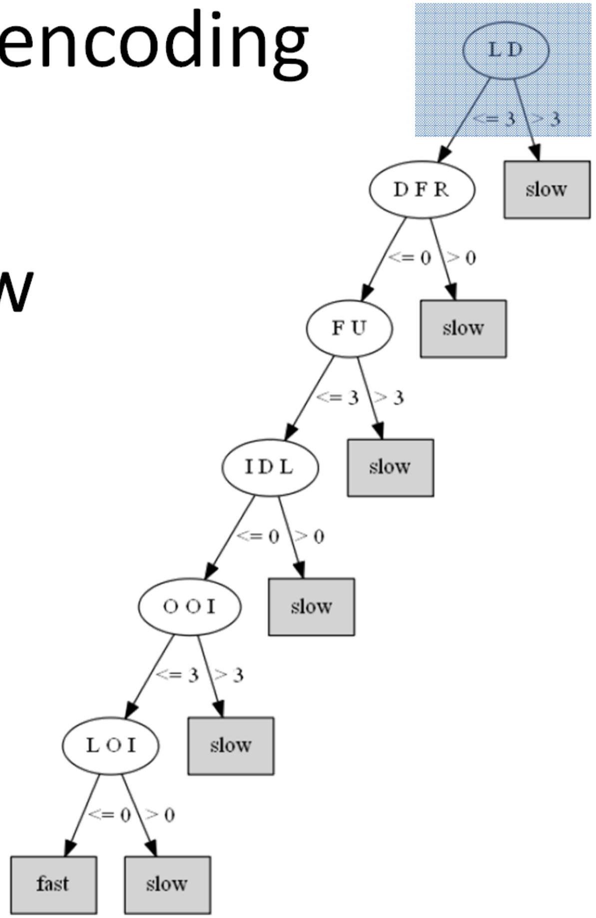
state space (SVM)	79 %	83 %
mouse events (SVM)	62 %	79 %
sequences (decision tree)	77 %	79 %

Interpreting Decision Trees

Decision tree on sequences encoding

Left, Down too much => slow

Waldo is in upper right



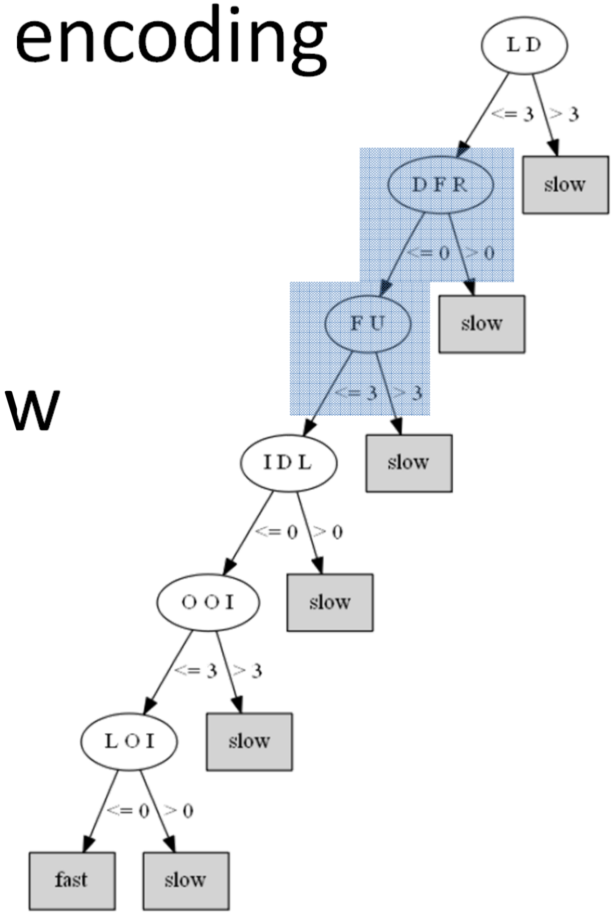
Interpreting Decision Trees

Decision tree on sequences encoding

Found, Up and

Down, *Found*, Right => slow

Clicking Found incorrectly



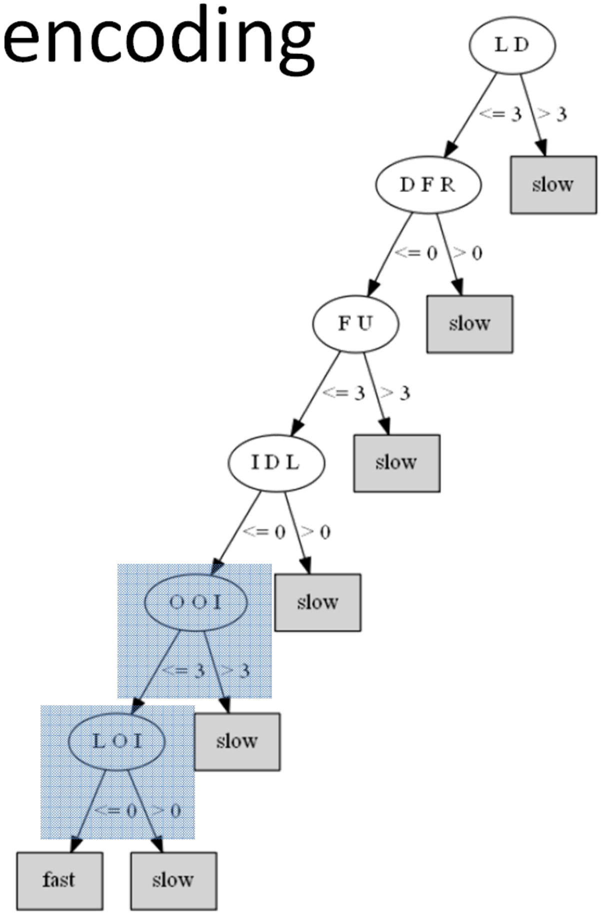
Interpreting Decision Trees

Decision tree on sequences encoding

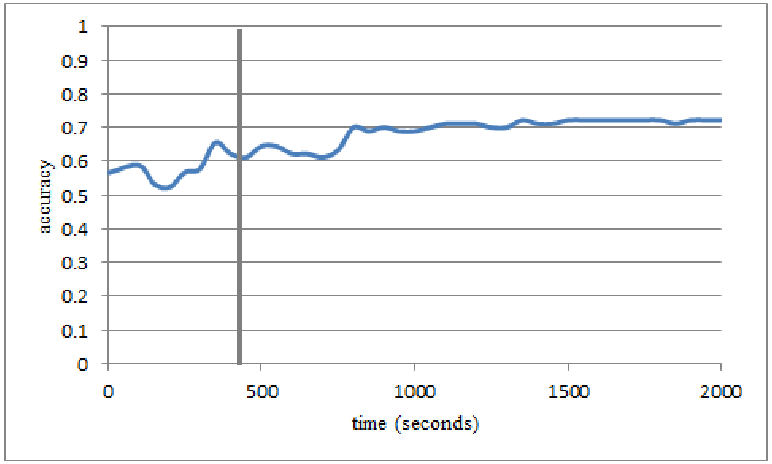
Out, Out, In and

Left, Out, In => slow

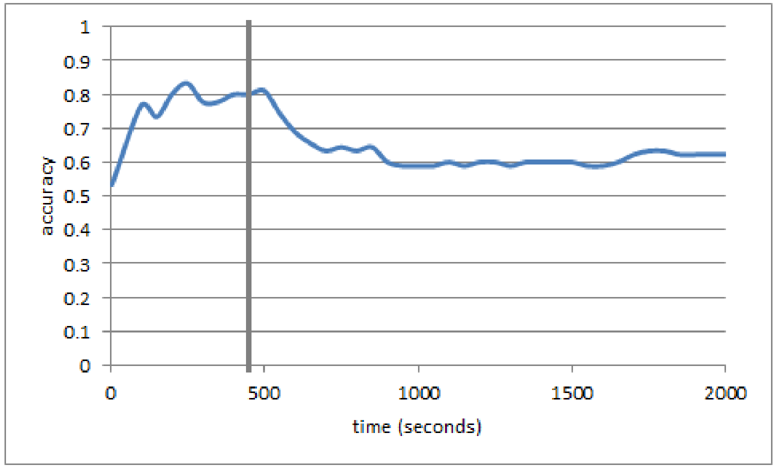
Out then in, null move?



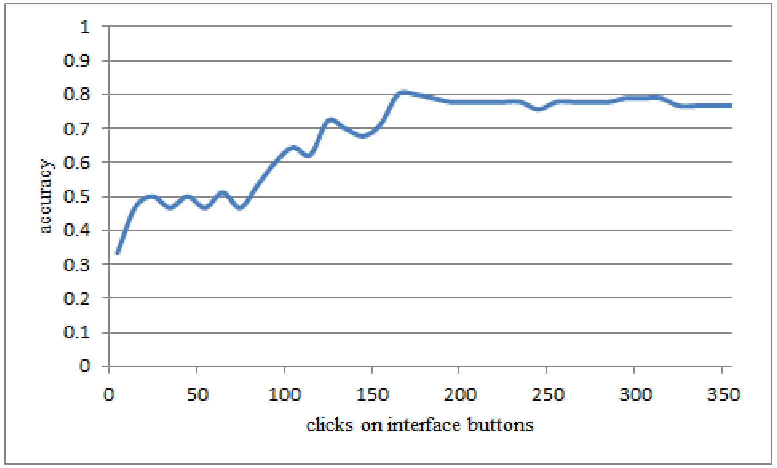
Limited Observation Time



State Space



Mouse Events



Sequences

Beyond Performance

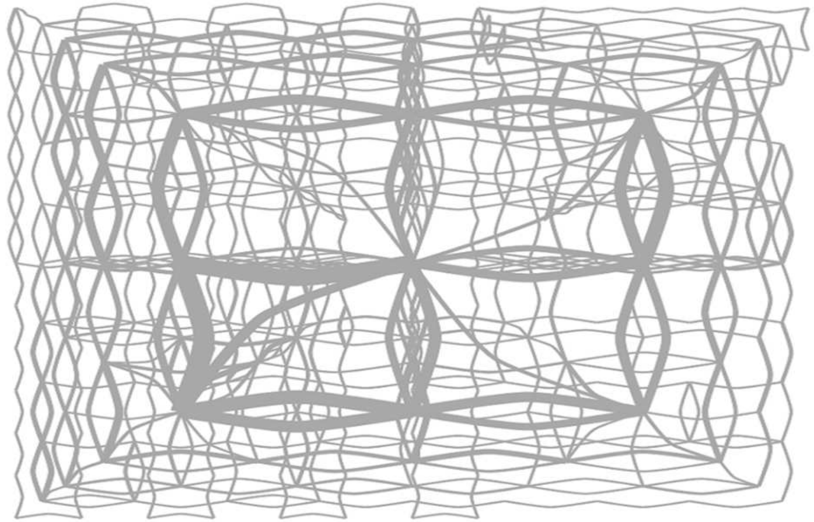
- Learning about the user

- Research has shown some testable personality traits correlate with performance on visualization tasks

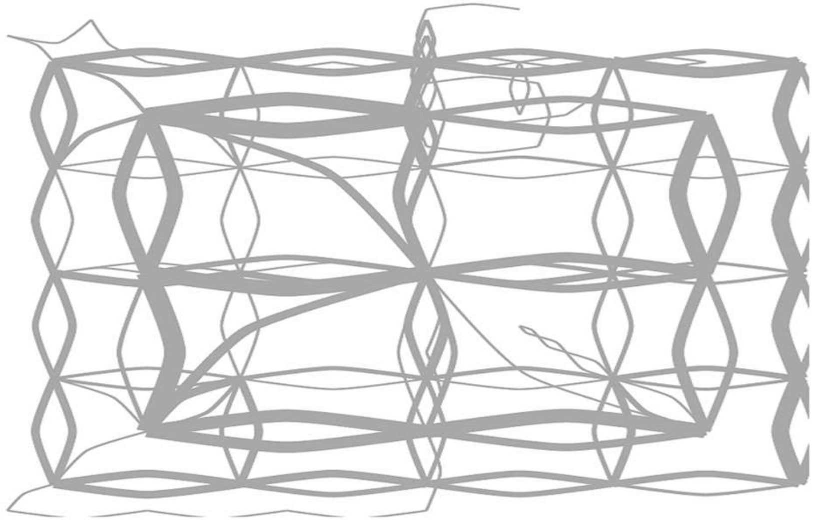
- Surveyed users:
  - Big Five personality traits
  - Locus of Control

Beyond Performance: Visualization

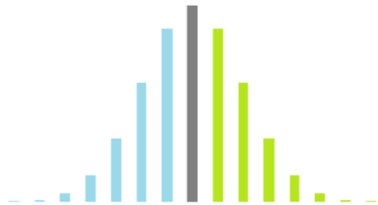
Locus of Control:
External



Locus of Control:
Internal



Beyond Performance: Accuracy



Locus of Control	67 %	sequences (decision tree)

Waldo Conclusions

- It is possible to automatically extract information about users from their interactions
  - Performance
  - Personality, though noisier
- Little observation needed
- Get an idea of strategy

Conclusion

- Machine Learning + Interactive Visualization can be the future of Visual Analytics
  - It provides an intuitive interface,
  - and hides the ugly math

- Can leverages user's interactions for:
  - Learning a *data* model
  - Learning a *user* model

Ongoing Research Agenda

- Data stakeholders

- Continue my Microsoft Research work (Interactive machine learning)

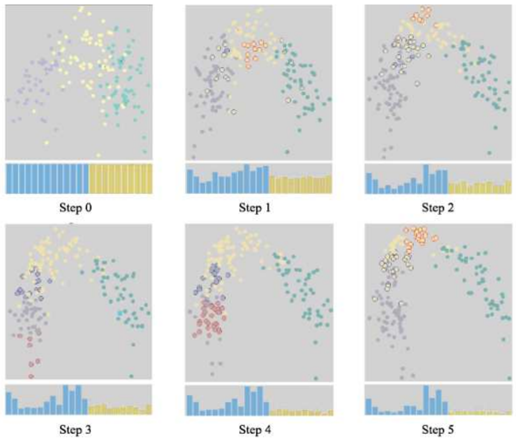
- Graduate

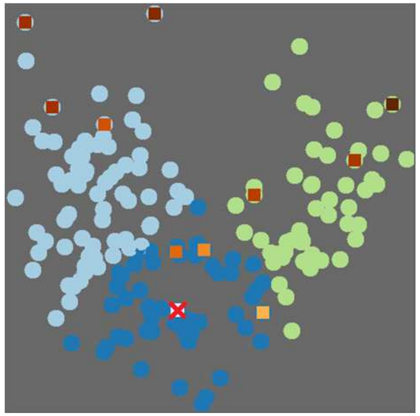
  …

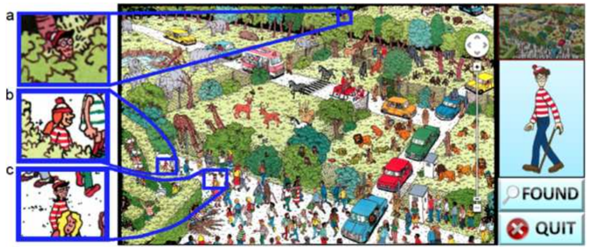
- Profit

[v]alt
Visual Analytics Laboratory @ Tufts

Tufts
UNIVERSITY







Thank You

Eli T. Brown