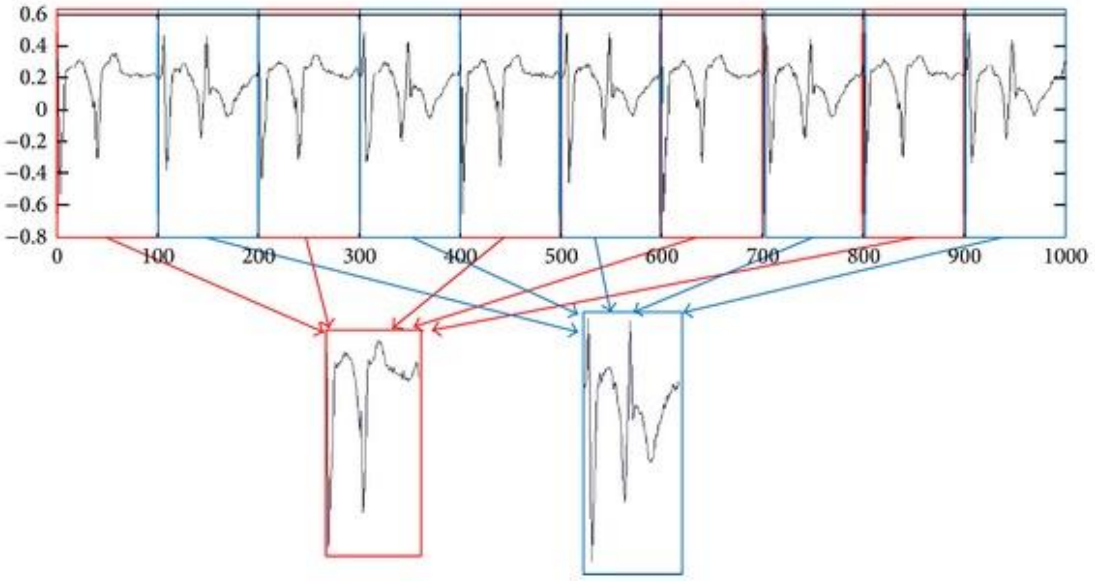
A Visualization Framework for Streaming Multivariate Data

Klaus Mueller

Visual Analytics and Imaging Lab
Computer Science Department
Stony Brook University

What's a Behavior Pattern?

A salient subsequence in a time series



- can be clustered and mined
- can be treated as a motif and associated with a scalar ID
- the scalar ID then becomes a scalar data point

What's a Multivariate Behavior Pattern?

Really just a simultaneous set of such patterns

What's a Multivariate Behavior Pattern?

Really just a simultaneous set of such patterns

- can be clustered and mined
- can be treated as motif and associated with a scalar ID
- at some discretized level
- the scalar ID then becomes a multivariate data point

Similarity Functions

Important metric

- Manhattan (L1)
- Euclidian (L2)
- cosine
- correlation
- structural

$$\mathrm{SSIM}(x,y)=\left[\frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1}\right]^{\alpha}\cdot\left[\frac{2\sigma_x\sigma_y+c_2}{\sigma_x^2+\sigma_y^2+c_2}\right]^{\beta}\cdot\left[\frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3}\right]^{\gamma}$$

luminance contrast structure

- domain-specific features

How About Sub Seq Window Size?

Can be found via

- optimization from prior samples
- possibly involving the users
- detect periodicity via wavelets and Fourier analysis

Use DTW (Dynamic Time Warping) to align two sub sequences of possibly unequal length

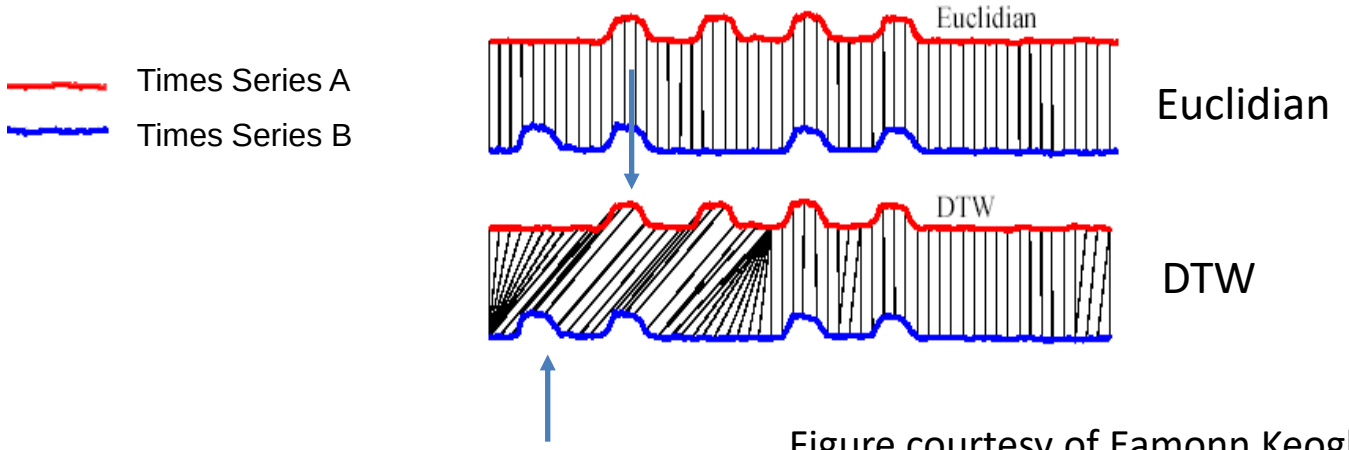
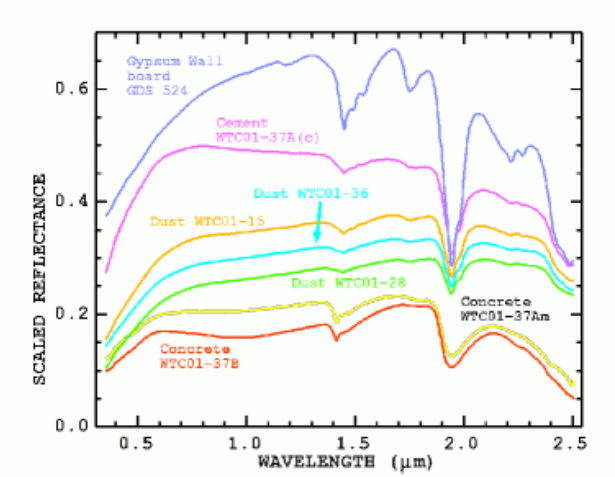
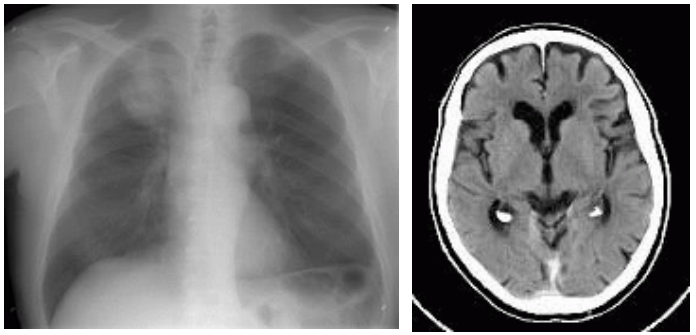


Figure courtesy of Eamonn Keogh, UC Riverside

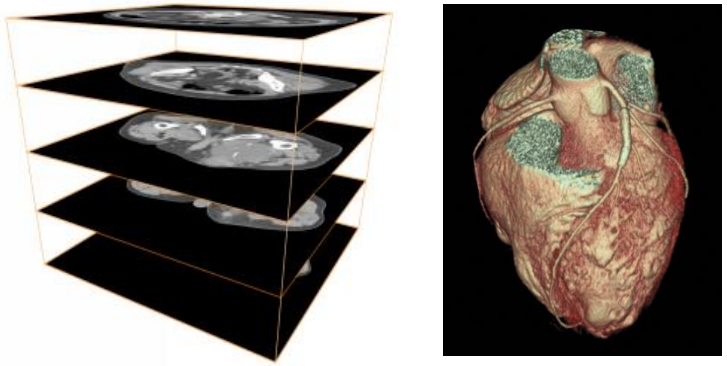
Visualization of 1-3 Dimensional Data



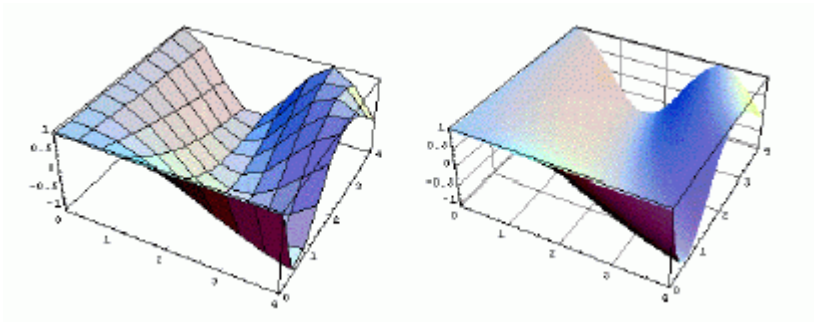
1D signal $f(x)$



2D signal $f(x, y)$



3D signal $f(x, y, z)$



2D signal, shown as height field

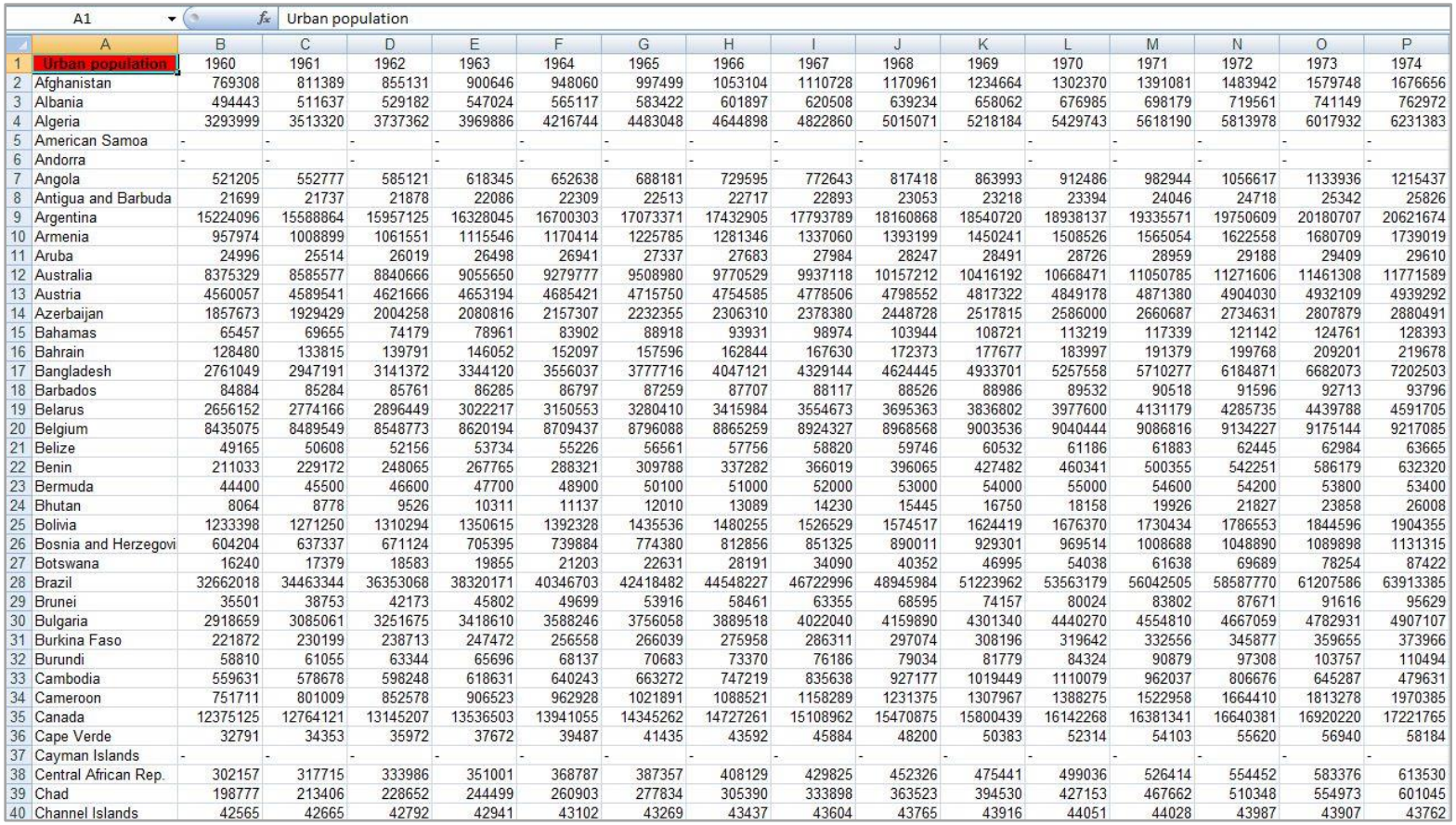
4D signal $f(x, y, z, t$=time)
example: 3D heart in motion

High-Dimensional Data

Consider the salient features of a car (not very high-D):

- miles per gallon (MPG)
- top speed
- acceleration
- number of cylinders
- horsepower
- weight
- year
- country origin
- brand
- number of seats
- number of doors
- reliability (average number of breakdowns)
- and so on...

Can You See Patterns in a Spreadsheet?



A1 | Urban population

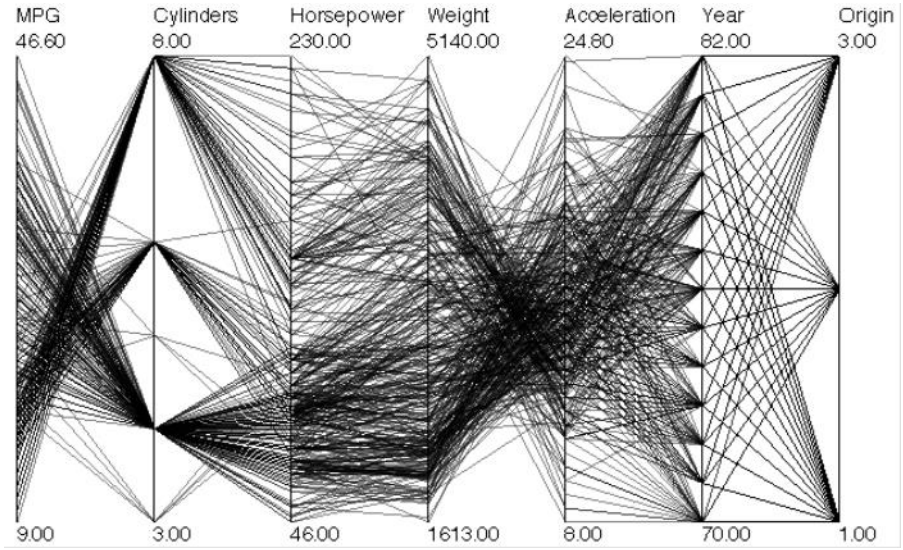
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Urban population	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
2	Afghanistan	769308	811389	855131	900646	948060	997499	1053104	1110728	1170961	1234664	1302370	1391081	1483942	1579748	1676656
3	Albania	494443	511637	529182	547024	565117	583422	601897	620508	639234	658062	676985	698179	719561	741149	762972
4	Algeria	3293999	3513320	3737362	3969886	4216744	4483048	4644898	4822860	5015071	5218184	5429743	5618190	5813978	6017932	6231383
5	American Samoa	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	Andorra	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	Angola	521205	552777	585121	618345	652638	688181	729595	772643	817418	863993	912486	982944	1056617	1133936	1215437
8	Antigua and Barbuda	21699	21737	21878	22086	22309	22513	22717	22893	23053	23218	23394	24046	24718	25342	25826
9	Argentina	15224096	15588864	15957125	16328045	16700303	17073371	17432905	17793789	18160868	18540720	18938137	19335571	19750609	20180707	20621674
10	Armenia	957974	1008899	1061551	1115546	1170414	1225785	1281346	1337060	1393199	1450241	1508526	1565054	1622558	1680709	1739019
11	Aruba	24996	25514	26019	26498	26941	27337	27683	27984	28247	28491	28726	28959	29188	29409	29610
12	Australia	8375329	8585577	8840666	9055650	9279777	9508980	9770529	9937118	10157212	10416192	10668471	11050785	11271606	11461308	11771589
13	Austria	4560057	4589541	4621666	4653194	4685421	4715750	4754585	4778506	4798552	4817322	4849178	4871380	4904030	4932109	4939292
14	Azerbaijan	1857673	1929429	2004258	2080816	2157307	2232355	2306310	2378380	2448728	2517815	2586000	2660687	2734631	2807879	2880491
15	Bahamas	65457	69655	74179	78961	83902	88918	93931	98974	103944	108721	113219	117339	121142	124761	128393
16	Bahrain	128480	133815	139791	146052	152097	157596	162844	167630	172373	177677	183997	191379	199768	209201	219678
17	Bangladesh	2761049	2947191	3141372	3344120	3556037	3777716	4047121	4329144	4624445	4933701	5257558	5710277	6184871	6682073	7202503
18	Barbados	84884	85284	85761	86285	86797	87259	87707	88117	88526	88986	89532	90518	91596	92713	93796
19	Belarus	2656152	2774166	2896449	3022217	3150553	3280410	3415984	3554673	3695363	3836802	3977600	4131179	4285735	4439788	4591705
20	Belgium	8435075	8489549	8548773	8620194	8709437	8796088	8865259	8924327	8968568	9003536	9040444	9086816	9134227	9175144	9217085
21	Belize	49165	50608	52156	53734	55226	56561	57756	58820	59746	60532	61186	61883	62445	62984	63665
22	Benin	211033	229172	248065	267765	288321	309788	337282	366019	396065	427482	460341	500355	542251	586179	632320
23	Bermuda	44400	45500	46600	47700	48900	50100	51000	52000	53000	54000	55000	54600	54200	53800	53400
24	Bhutan	8064	8778	9526	10311	11137	12010	13089	14230	15445	16750	18158	19926	21827	23858	26008
25	Bolivia	1233398	1271250	1310294	1350615	1392328	1435536	1480255	1526529	1574517	1624419	1676370	1730434	1786553	1844596	1904355
26	Bosnia and Herzegovi	604204	637337	671124	705395	739884	774380	812856	851325	890011	929301	969514	1008688	1048890	1089898	1131315
27	Botswana	16240	17379	18583	19855	21203	22631	28191	34090	40352	46995	54038	61638	69689	78254	87422
28	Brazil	32662018	34463344	36353068	38320171	40346703	42418482	44548227	46722996	48945984	51223962	53563179	56042505	58587770	61207586	63913385
29	Brunei	35501	38753	42173	45802	49699	53916	58461	63355	68595	74157	80024	83802	87671	91616	95629
30	Bulgaria	2918659	3085061	3251675	3418610	3588246	3756058	3889518	4022040	4159890	4301340	4440270	4554810	4667059	4782931	4907107
31	Burkina Faso	221872	230199	238713	247472	256558	266039	275958	286311	297074	308196	319642	332556	345877	359655	373966
32	Burundi	58810	61055	63344	65696	68137	70683	73370	76186	79034	81779	84324	90879	97308	103757	110494
33	Cambodia	559631	578678	598248	618631	640243	663272	747219	835638	927177	1019449	1110079	962037	806676	645287	479631
34	Cameroon	751711	801009	852578	906523	962928	1021891	1088521	1158289	1231375	1307967	1388275	1522958	1664410	1813278	1970385
35	Canada	12375125	12764121	13145207	13536503	13941055	14345262	14727261	15108962	15470875	15800439	16142268	16381341	16640381	16920220	17221765
36	Cape Verde	32791	34353	35972	37672	39487	41435	43592	45884	48200	50383	52314	54103	55620	56940	58184
37	Cayman Islands	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
38	Central African Rep.	302157	317715	333986	351001	368787	387357	408129	429825	452326	475441	499036	526414	554452	583376	613530
39	Chad	198777	213406	228652	244499	260903	277834	305390	333898	363523	394530	427153	467662	510348	554973	601045
40	Channel Islands	42565	42665	42792	42941	43102	43269	43437	43604	43765	43916	44051	44028	43987	43907	43762

Very hard….

Parallel Coordinates

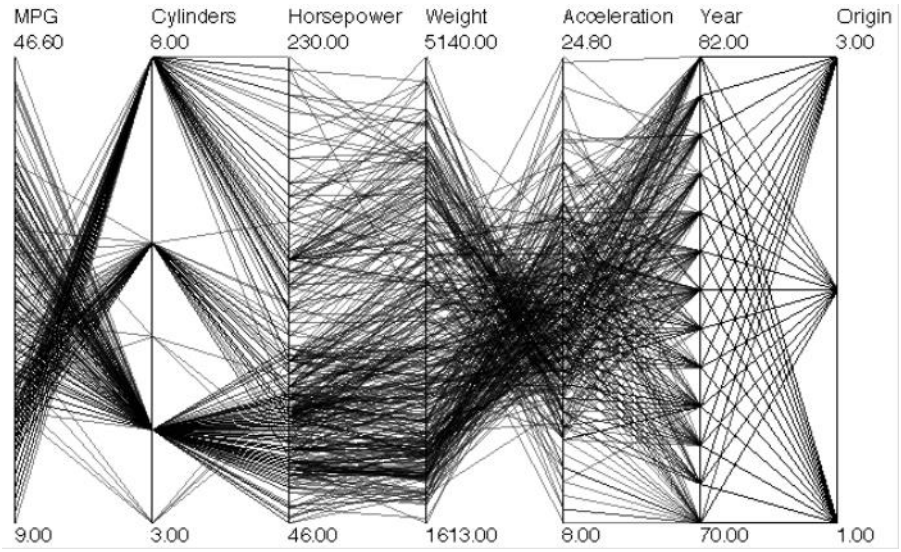
a car as a 7-dimensional data point

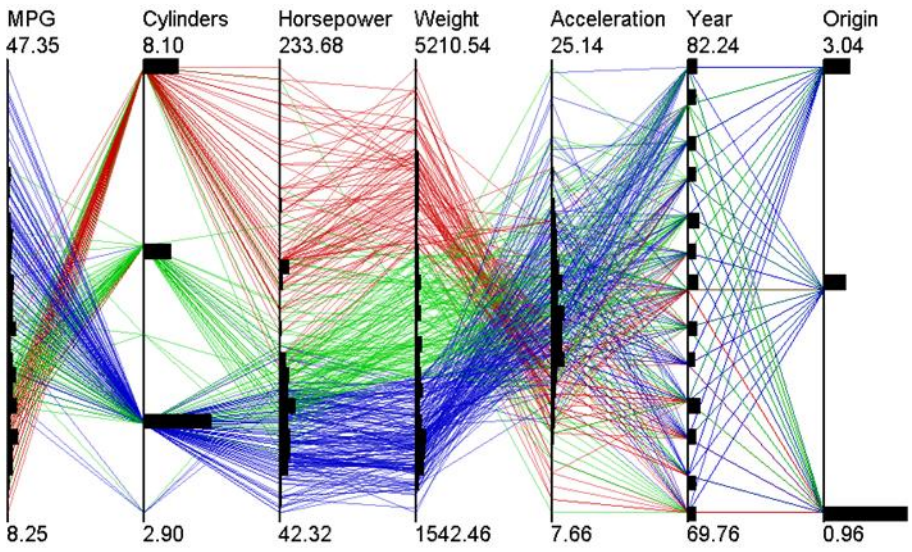
Parallel Coordinates



a database of cars

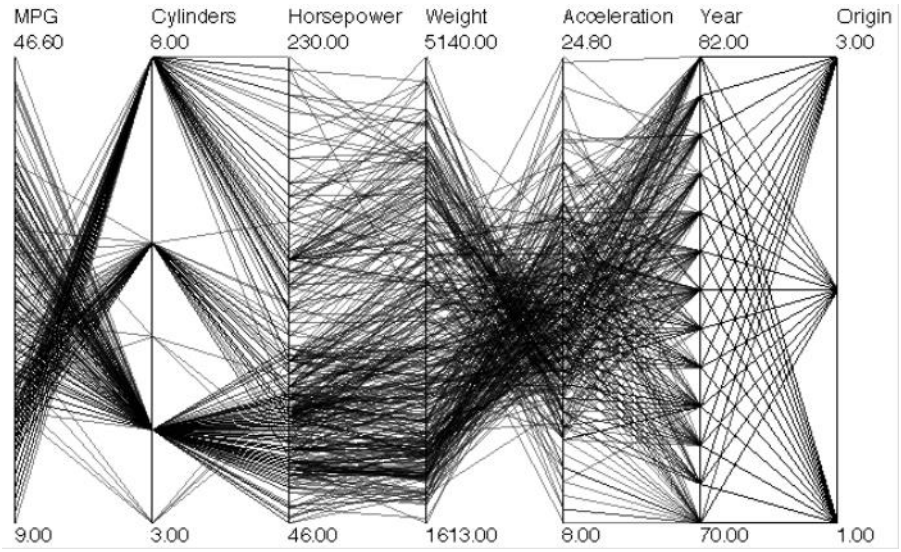
Parallel Coordinates

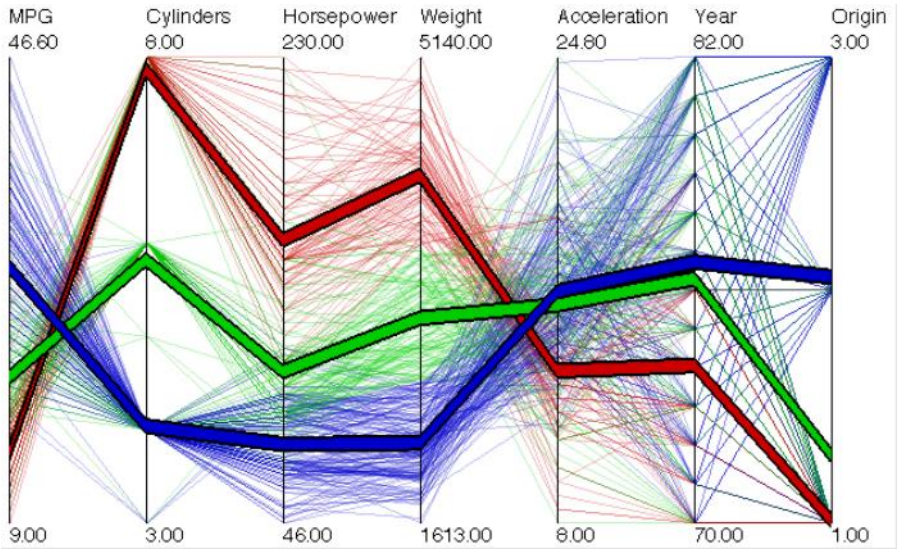




after some clustering

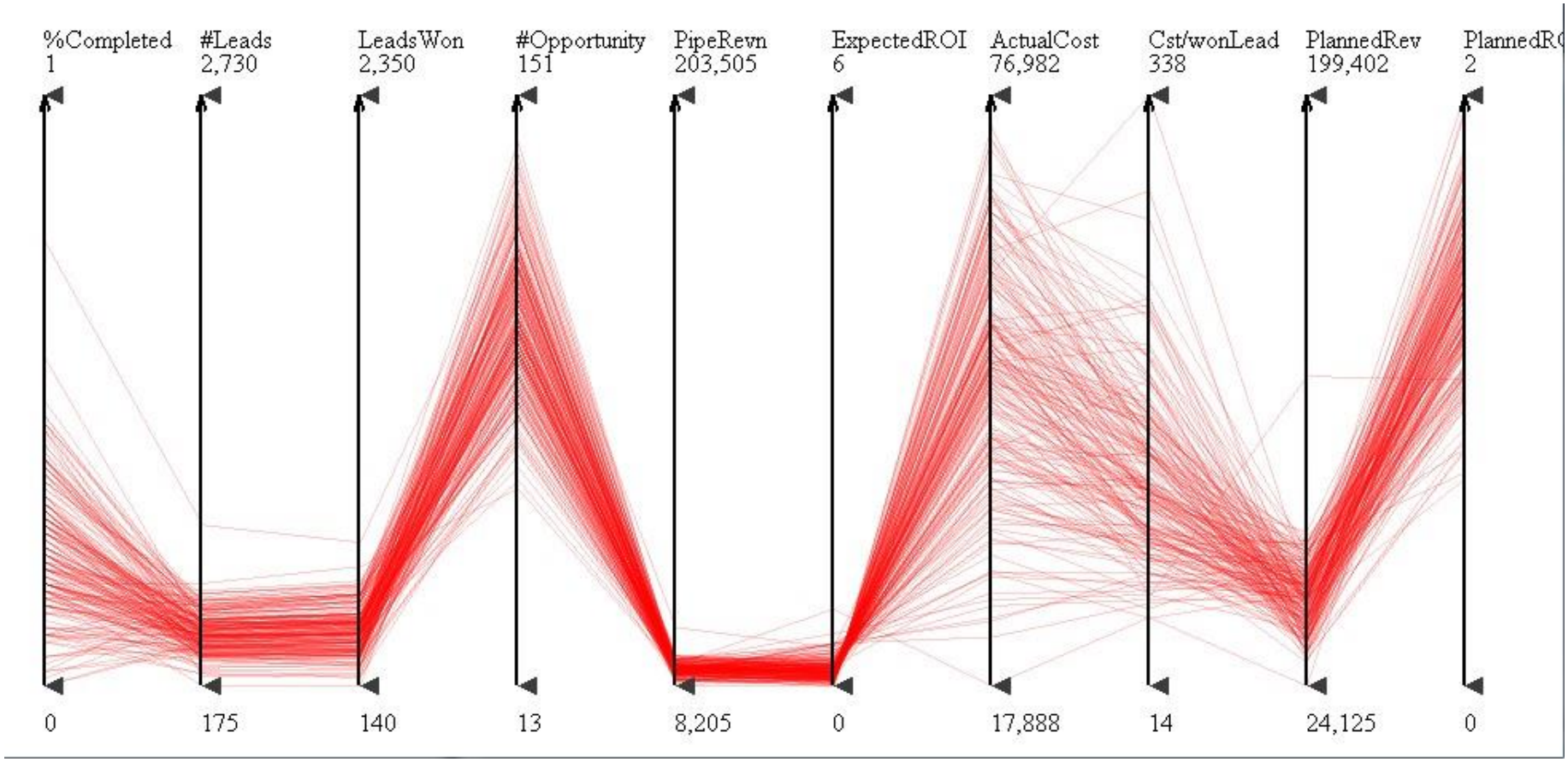
Parallel Coordinates





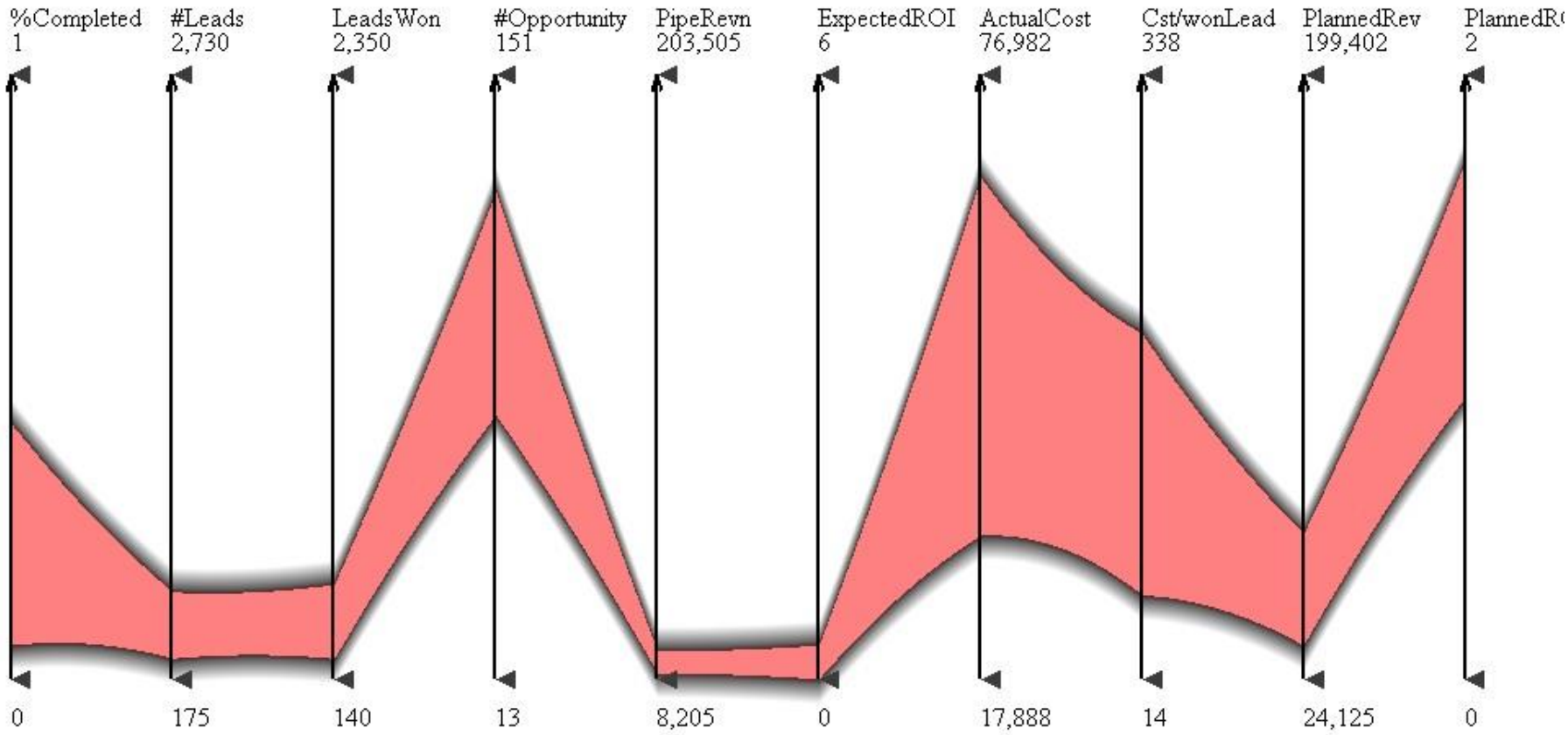
with mean trend

PC With Illustrative Abstraction



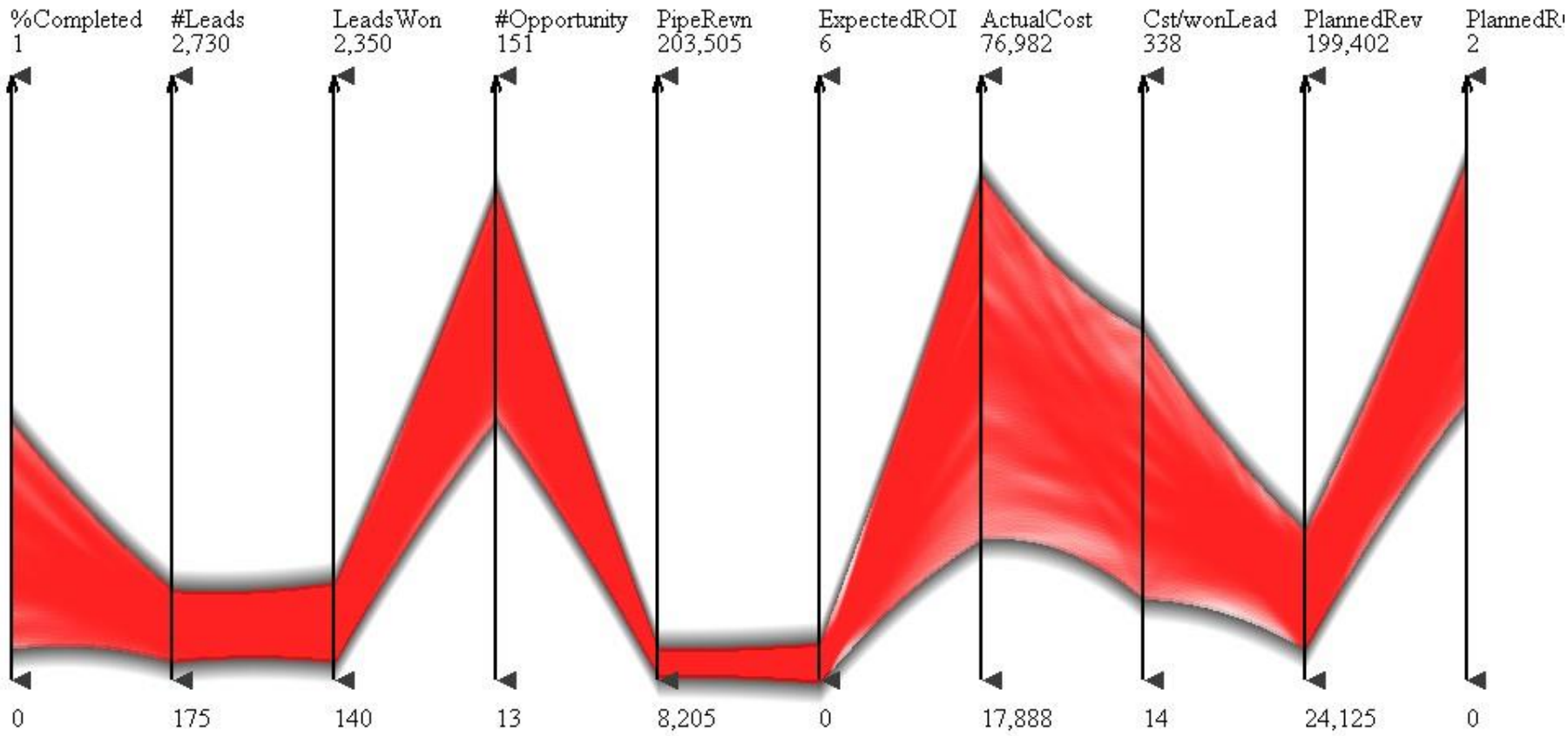
individual polylines

PC With Illustrative Abstraction



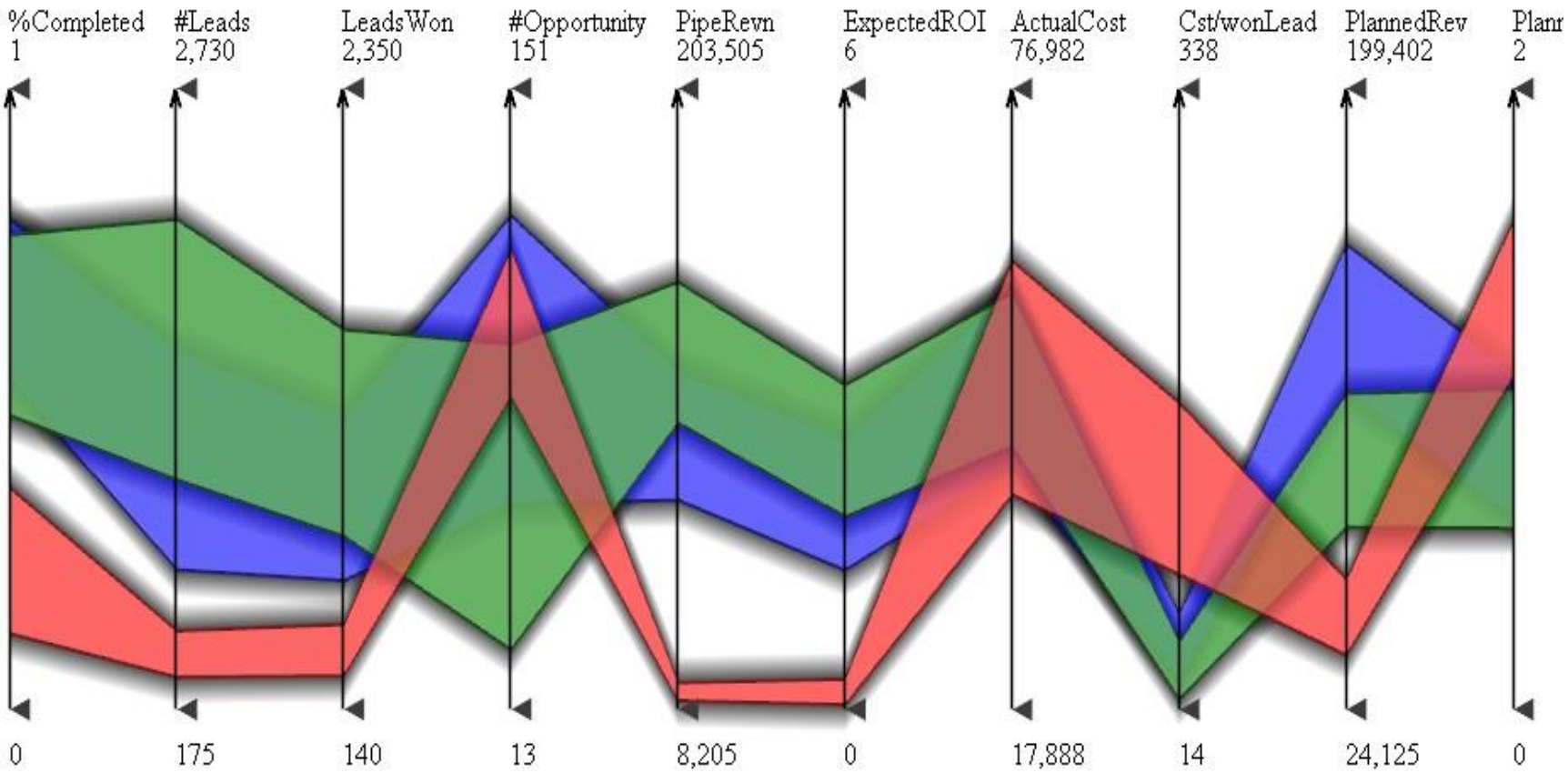
completely abstracted away

PC With Illustrative Abstraction



blended partially

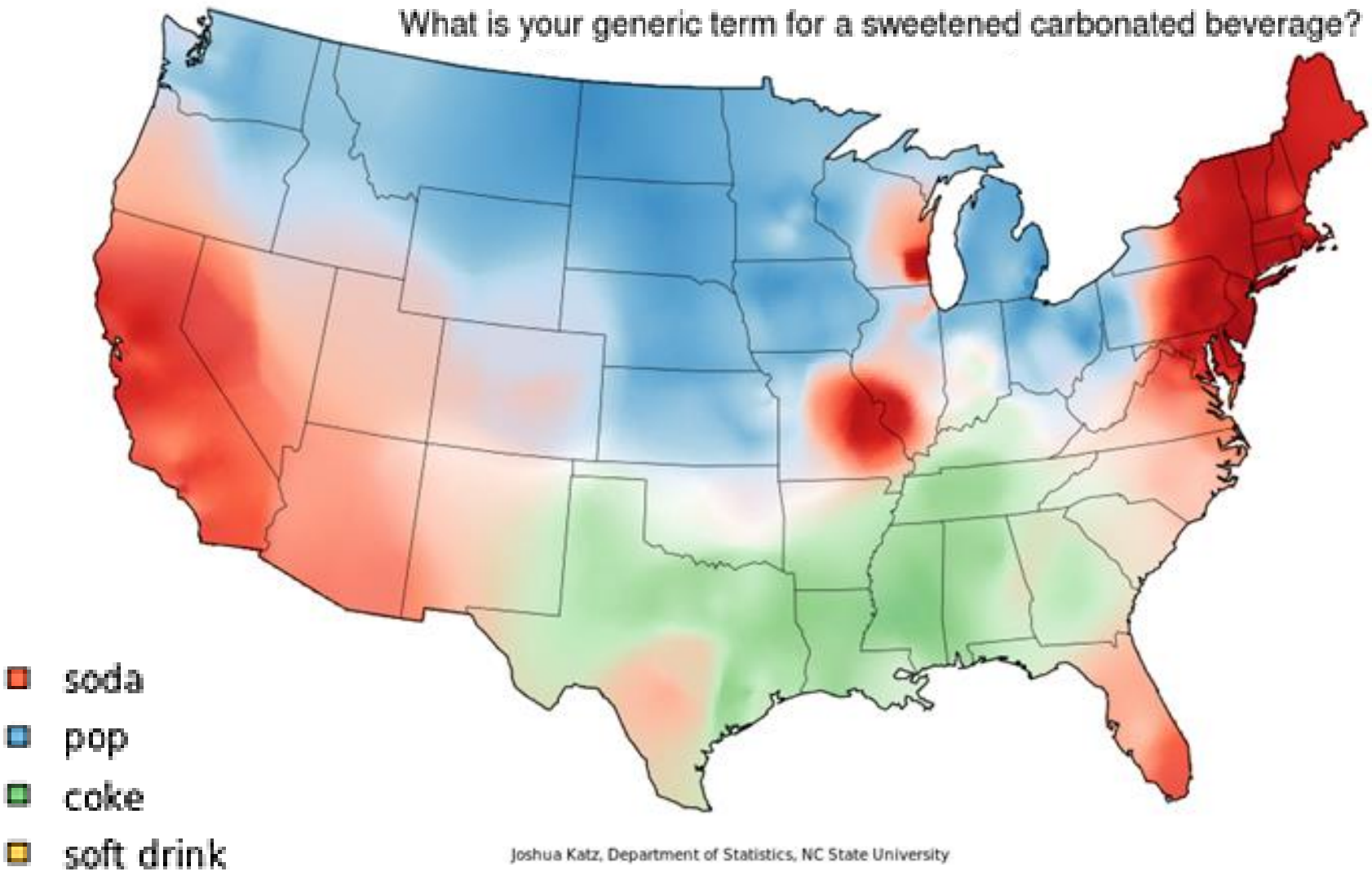
PC With Illustrative Abstraction



all put together – three clusters

Interaction in Parallel Coordinate

Visualization Via Maps



Data Map Via Space Embedding

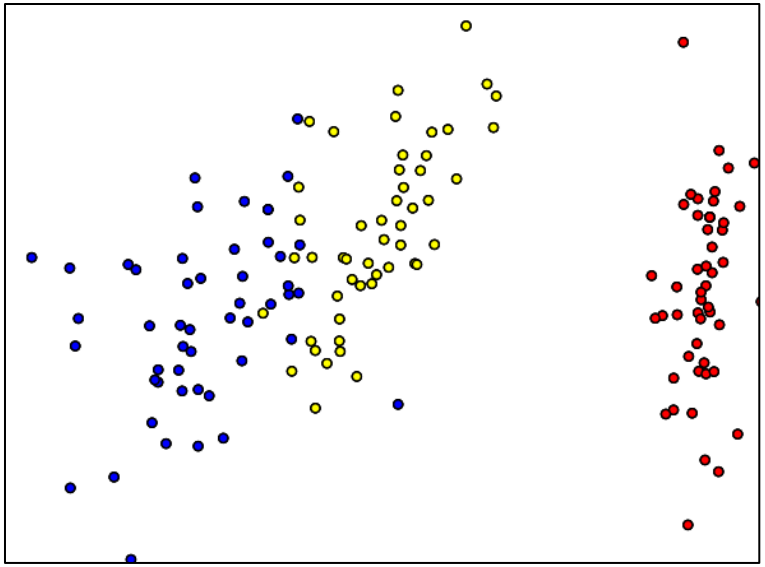
General idea:

- preserve N-D space distances δ_{ij} in 2-D space d_{ij}
- comes to down to an optimization problem
- minimize

$$stress = \sqrt{\frac{\sum_{ij}(d_{ij} - \delta_{ij})^2}{\sum_{ij}\delta_{ij}^{\ 2}}}$$

- Multi-Dimensional Scaling (MDS)
- similar data map to similar places

  → Similarity Map

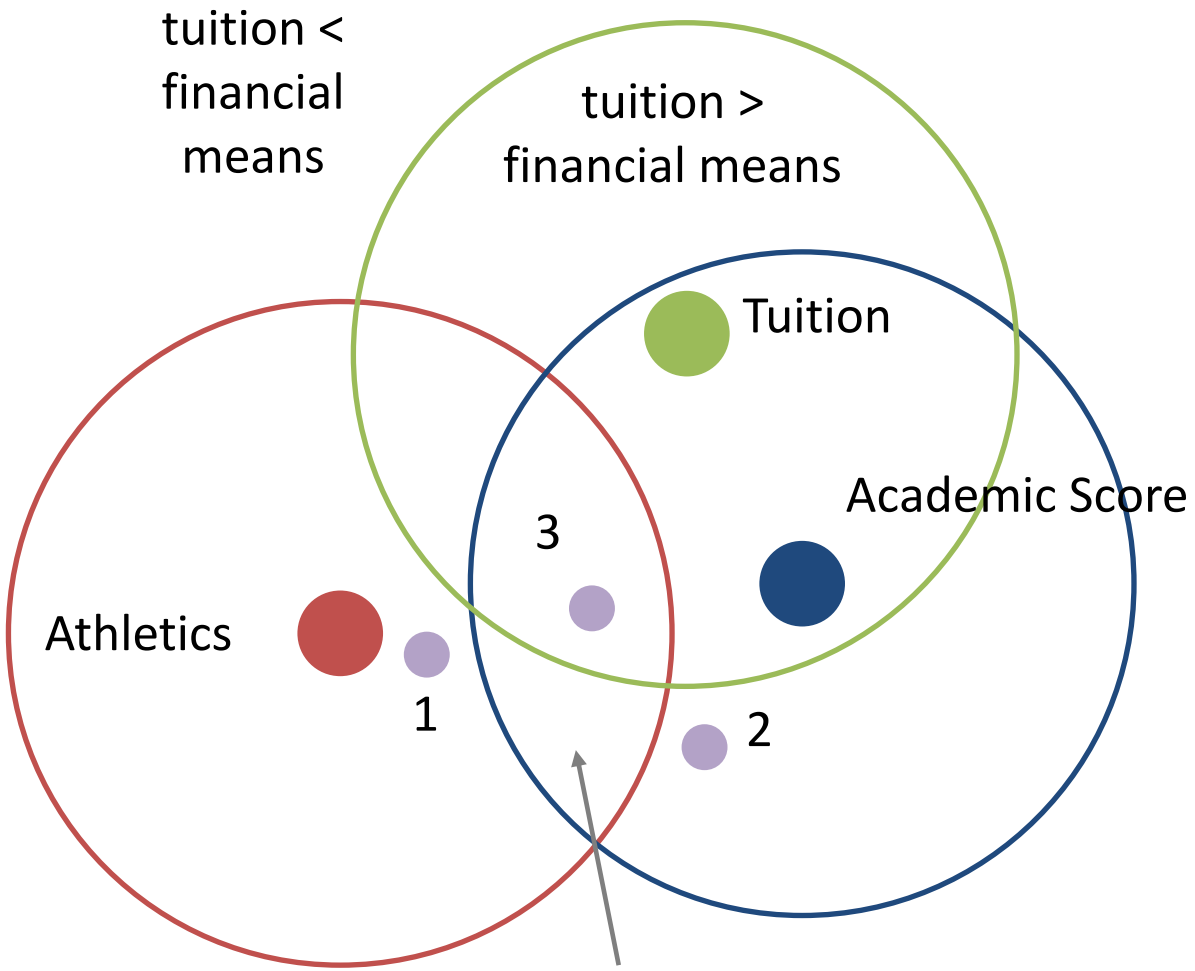


But….

…are these clusters so different?

We Need to Map The Attributes, Too

Example College Selection

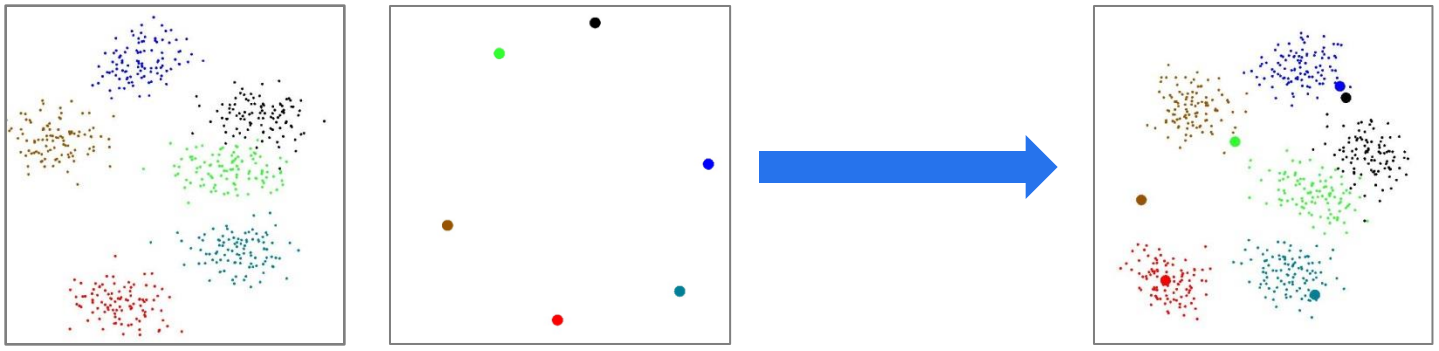


no dream school here: good athletics, low tuition, high academic score

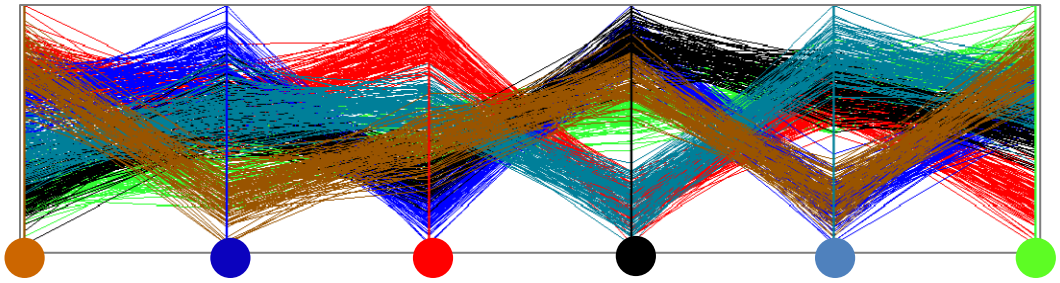
The Data Context Map

Best of both worlds

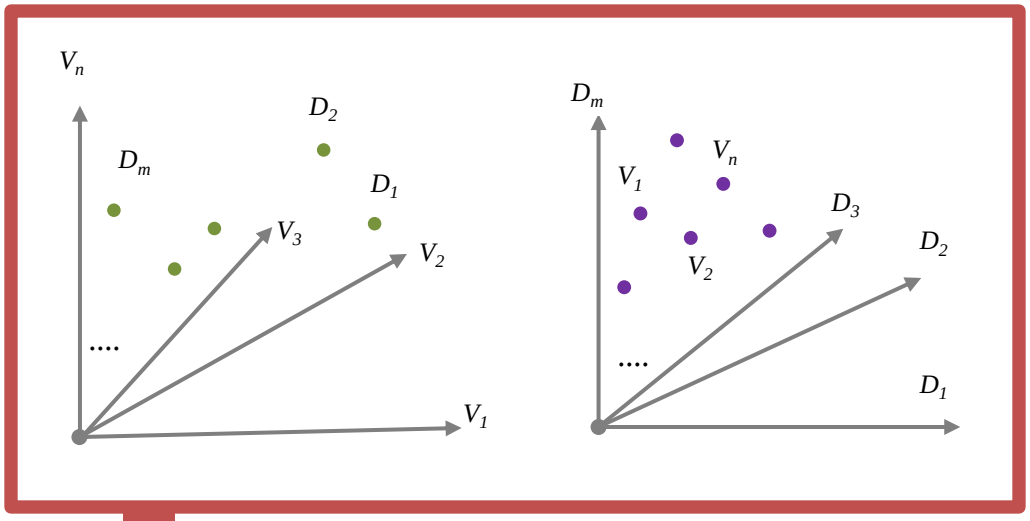
- similarity layout of the data based on vector similarity
- similarity layout of the attributes based on pairwise correlation

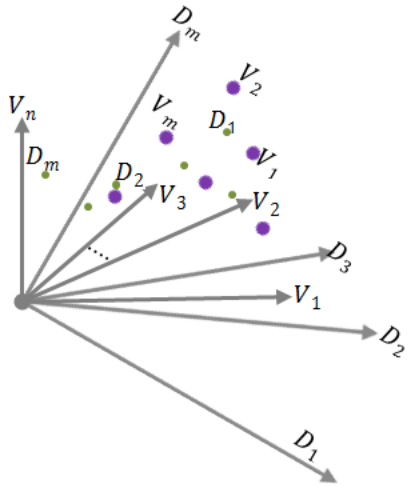


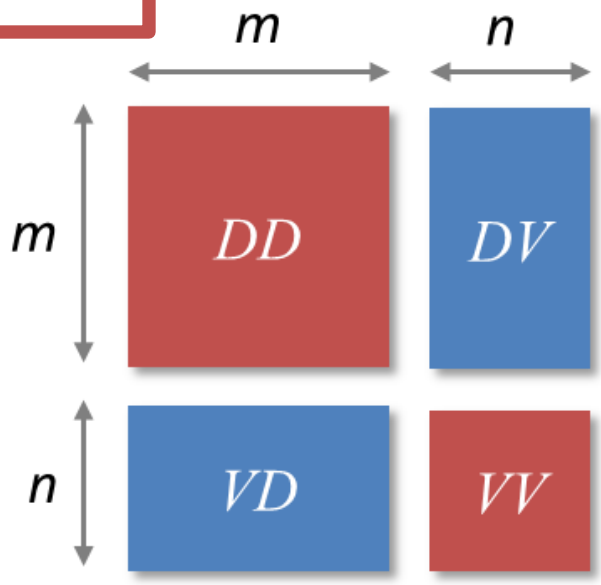
data attributes data + attributes



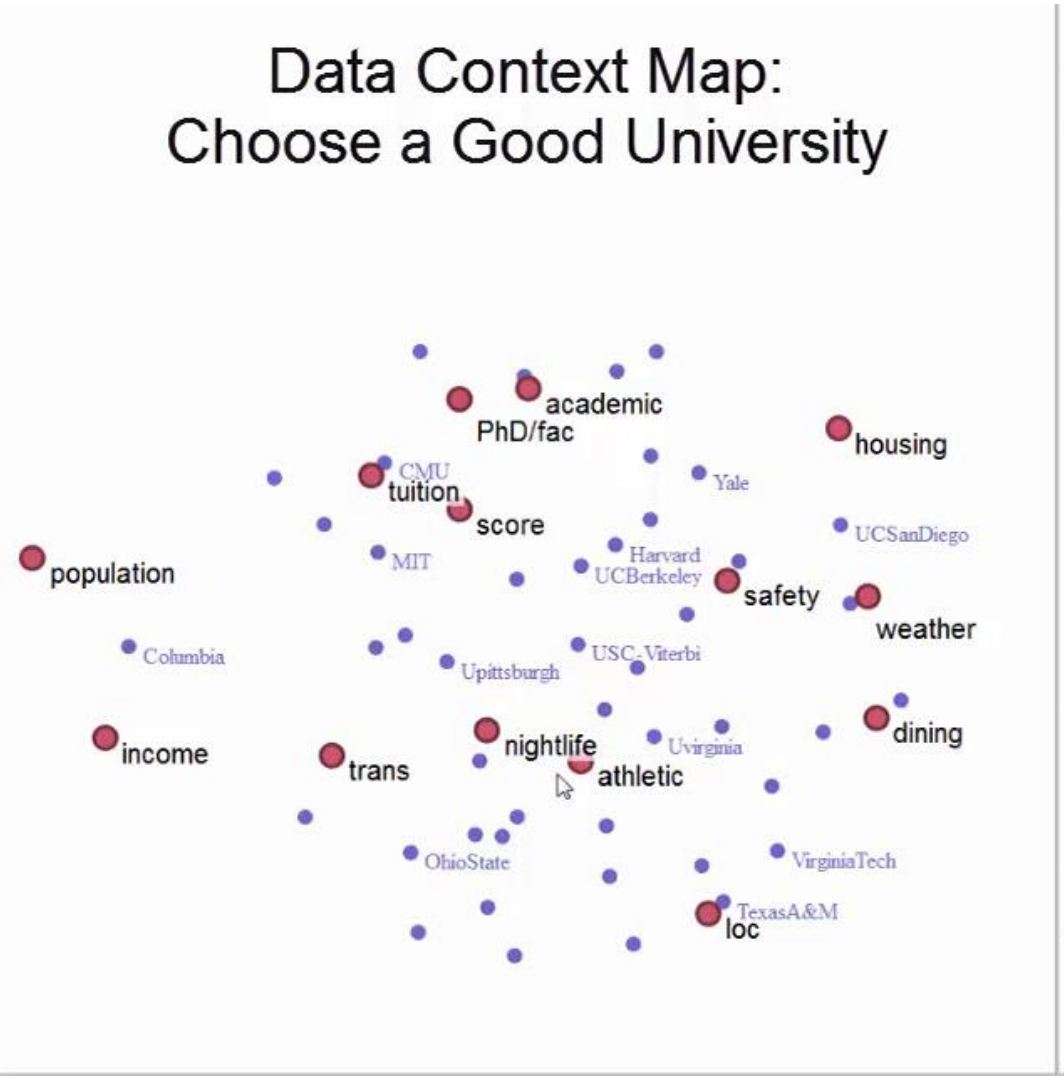
Achieved by Joint Matrix Optimization







The Data Context Map

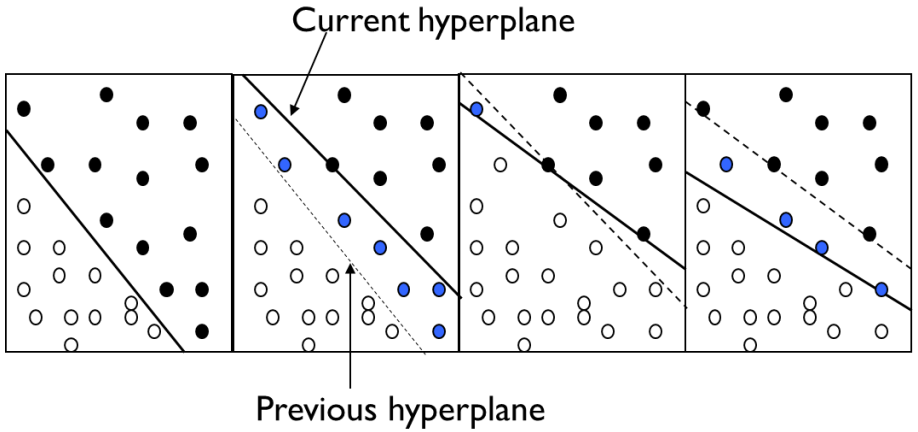


Streaming Data

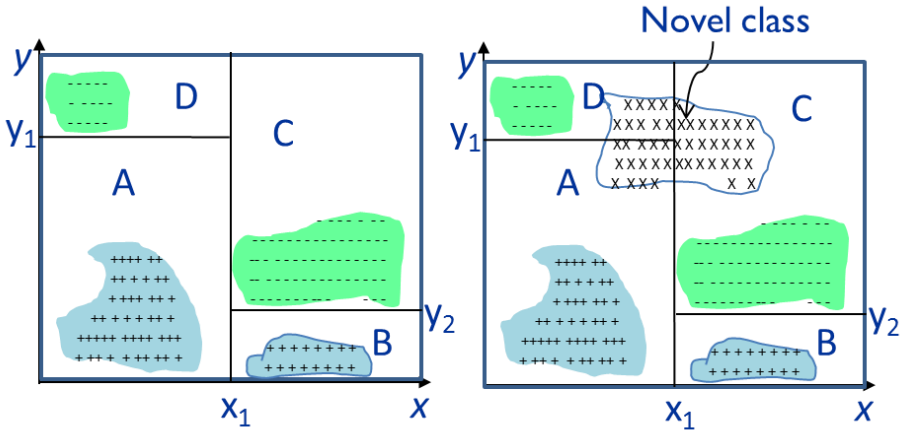
One-pass constraint

- data can be processed only once and not all be stored

Concept drift



Concept evolution

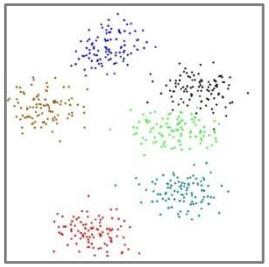


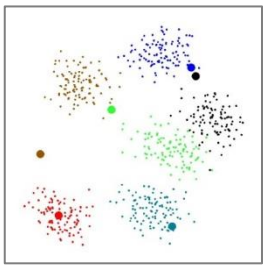
The Synopsis Map

Evolve clusters

- can keep up with concept drift and evolution
- add new samples and remove stale samples
- update clusters by merging, splitting, or removal
- maintain anomalies

Extend data context map into a *synopsis context map*



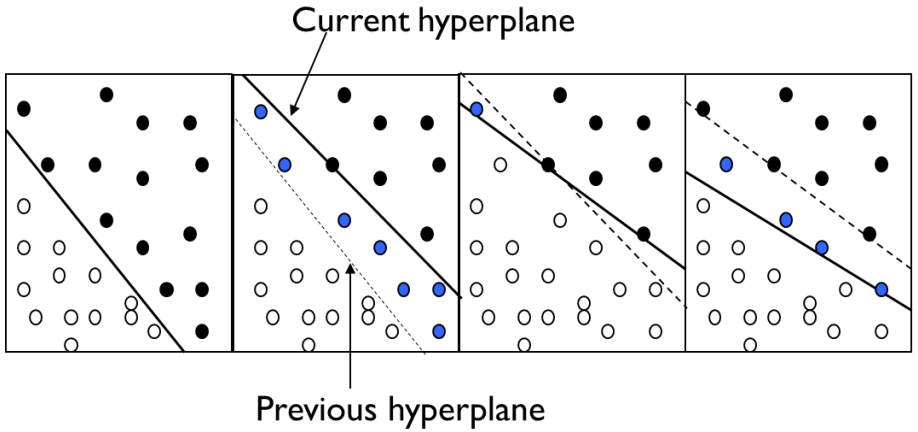


Streaming Data

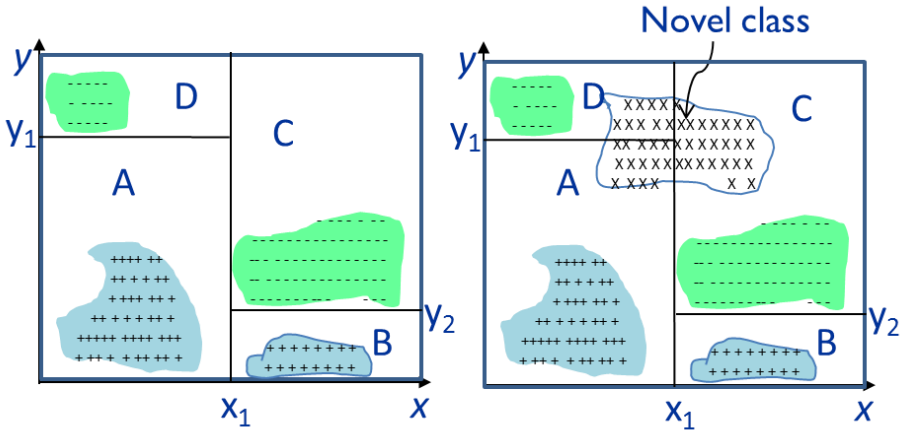
One-pass constraint

- data can be processed only once and not all be stored

Concept drift



Concept evolution

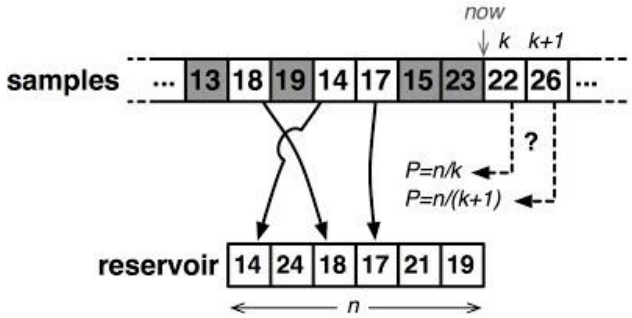


Synopsis

Keep representative samples as a synopsis

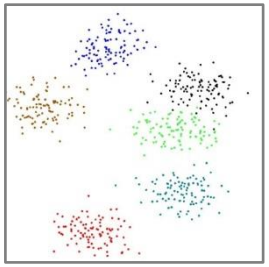
Simplest form is reservoir sampling

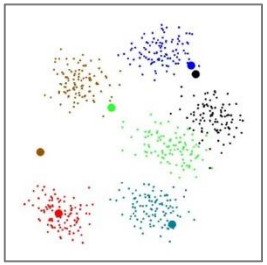
- purely probabilistic and sample-based
- p (sample in reservoir) is k/n



More informative is to evolve clusters (not samples)

- more apt to keep up with concept drift and evolution
- add new samples and remove stale samples
- update clusters by merging, splitting, or removal
- maintain anomalies





Extend data context map into a *synopsis context map*

Attribute Management

Distinguish between

- irrelevant vs. redundant vs. semantically similar variables

Irrelevant variables

- random relationships with all other variables (e.g. no correlations)
- data dependent

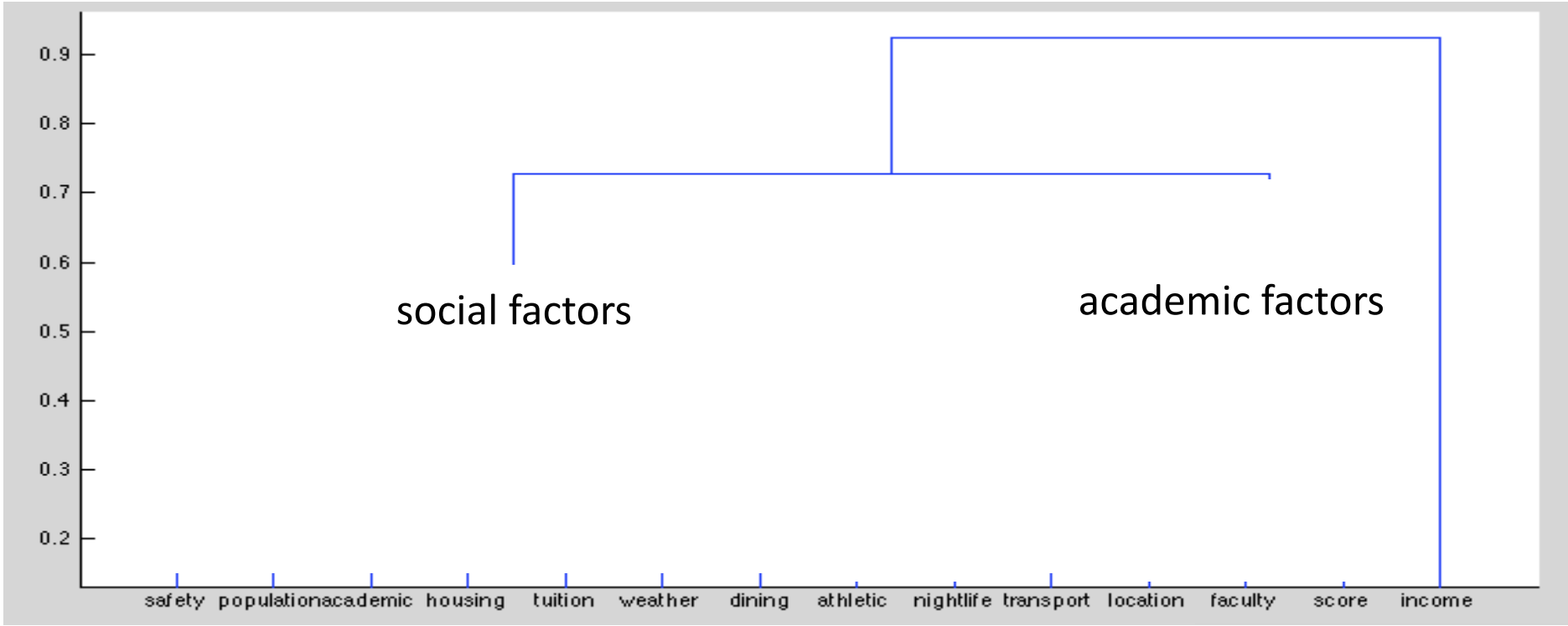
Redundant

- same relationships with all other variables (e.g., similar correlations)
- data dependent

Semantically similar variables

- similar meaning, class, category
- e.g. nightlife, transportation, housing, location → city

Ongoing Work: Attribute Dendrogram



But…

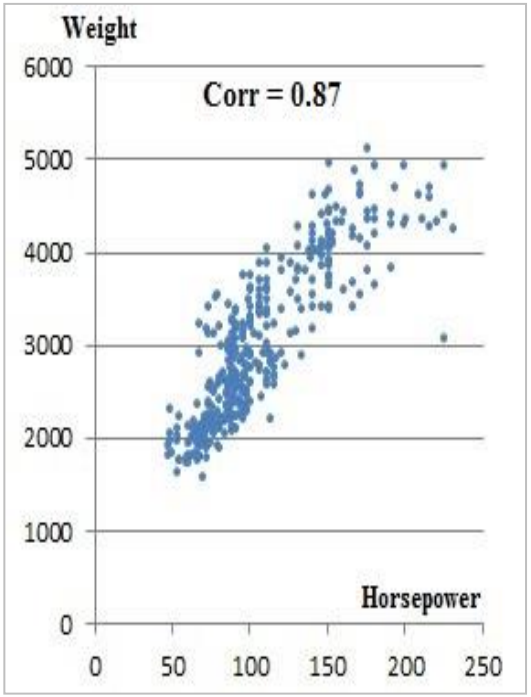
…do I need so much gas?

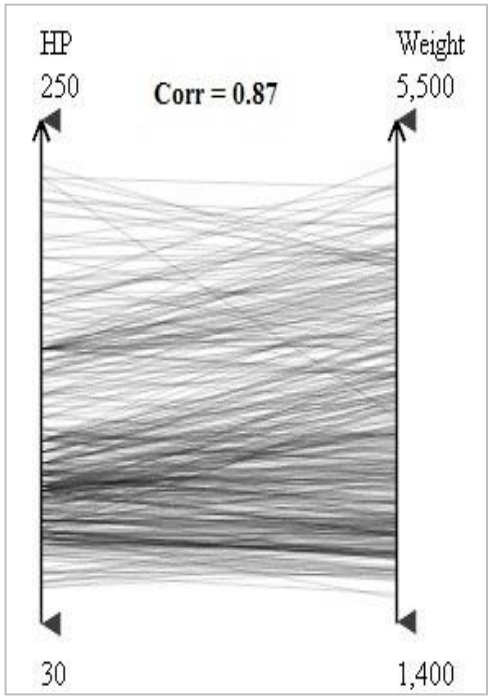
We Need a Measure for Relationships

Correlation

- a statistical measure that indicates the extent to which two or more variables fluctuate together

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{x})^2}}$$





Problems With These Visualizations?

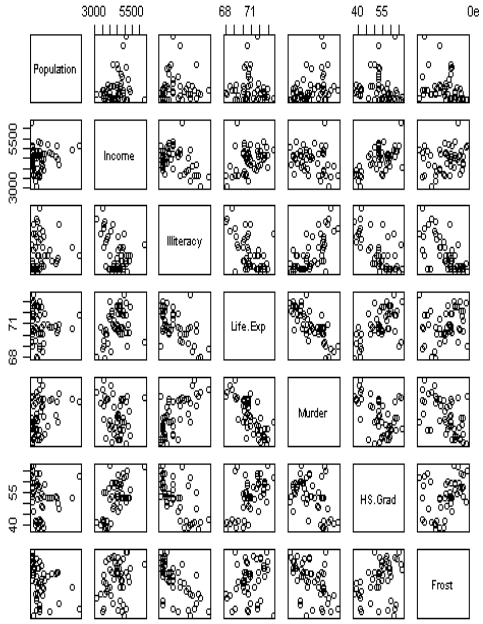
They don't scale well for large numbers of variables

- can you tell which variable is 2nd-most correlated with 'Income'?

Yes, we can use a correlation matrix heat map

- but brightness and color are poor visual variables to communicate quantitative

1	-0.78	-0.77	-0.83	0.42	0.58	0.36
-0.78	1	0.85	0.9	-0.51	-0.35	-0.57
-0.77	0.85	1	0.87	-0.69	-0.42	-0.28
-0.83	0.9	0.87	1	-0.42	-0.31	-0.45
0.42	-0.51	-0.69	-0.42	1	0.29	0.21
0.58	-0.35	-0.42	-0.31	0.29	1	0.18
0.36	-0.57	-0.28	-0.45	0.21	0.18	1



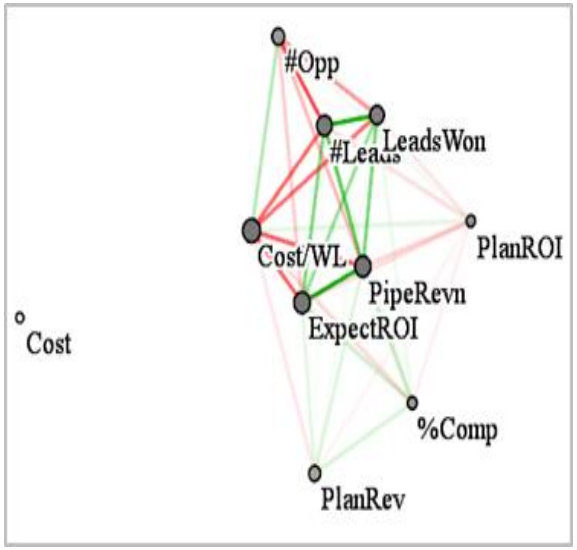
What's the #1 Visual Variable for QI?

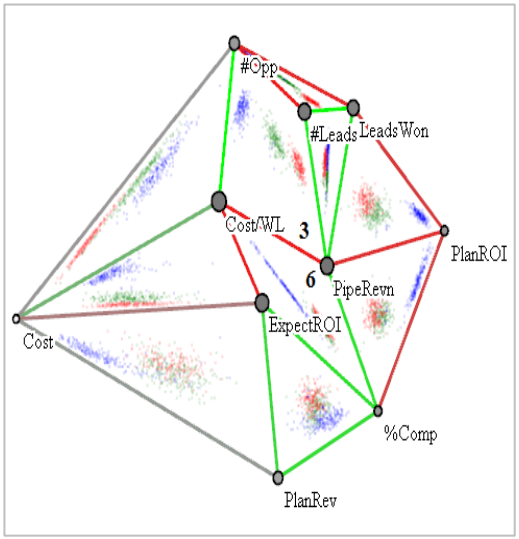
The spatial (planar) variables!! (J. Bertin, '67)

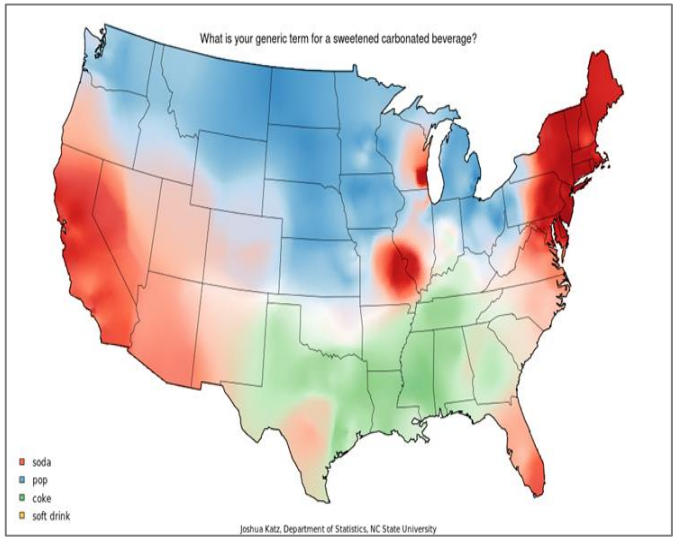
That's why geographic maps work so well

Can we build a *correlation* map?

You bet…







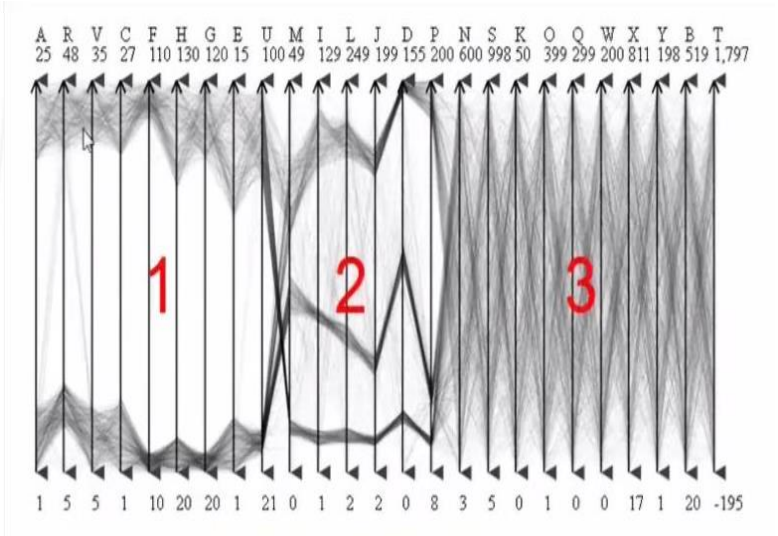
Building the Correlation Matrix

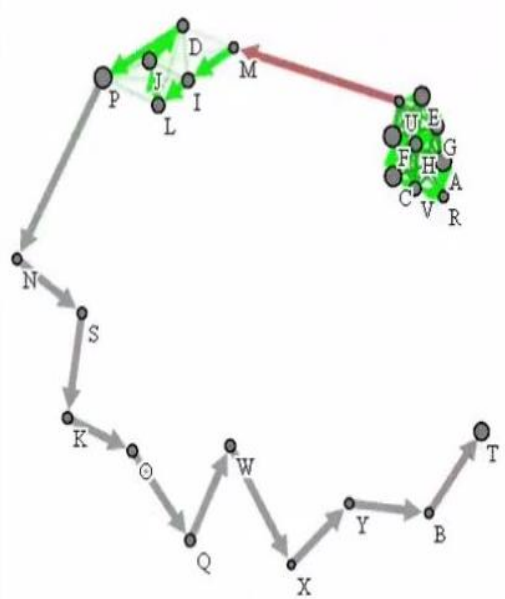
Create a correlation matrix

Run a mass-spring model

You can even use it to order your parallel coordinate axes via TSP

Run Traveling Salesman on the correlation nodes

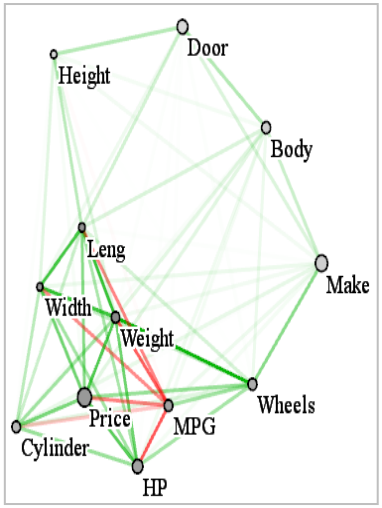




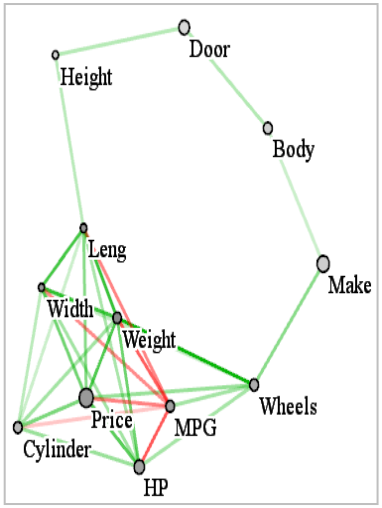
	MRK	MSFT	PFE	PG	T	TRV	UTX	VZ	WMT	XOM
MRK	1.	0.39	0.72	−0.43	0.57	0.031	−0.26	0.61	−0.11	−0.25
MSFT	0.39	1.	0.14	0.11	0.56	0.25	0.25	0.67	−0.074	0.24
PFE	0.72	0.14	1.	−0.77	0.08	−0.37	−0.65	0.19	−0.077	−0.72
PG	−0.43	0.11	−0.77	1.	0.25	0.68	0.92	0.086	0.072	0.9
T	0.57	0.56	0.08	0.25	1.	0.65	0.46	0.87	−0.059	0.54
TRV	0.031	0.25	−0.37	0.68	0.65	1.	0.83	0.43	−0.0067	0.81
UTX	−0.26	0.25	−0.65	0.92	0.46	0.83	1.	0.27	−0.033	0.93
VZ	0.61	0.67	0.19	0.086	0.87	0.43	0.27	1.	0.026	0.36
WMT	−0.11	−0.074	−0.077	0.072	−0.059	−0.0067	−0.033	0.026	1.	0.032
XOM	−0.25	0.24	−0.72	0.9	0.54	0.81	0.93	0.36	0.032	1.

Interaction with the Correlation Network

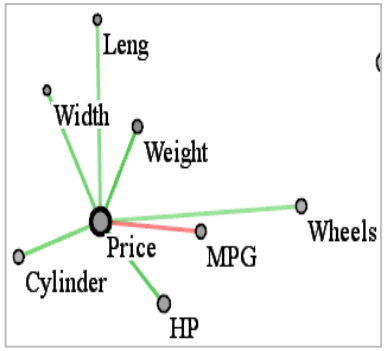
- Vertices are attributes, edges are correlations
  - vertex: size determined by $\sum_{j=0}^{D} \frac{|correlation(i,j)|}{D-1}$ $j \neq i$
  - edge: color/intensity → sign/strength of correlation



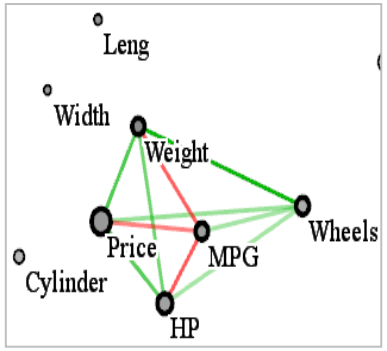
all edges



filtered by strength

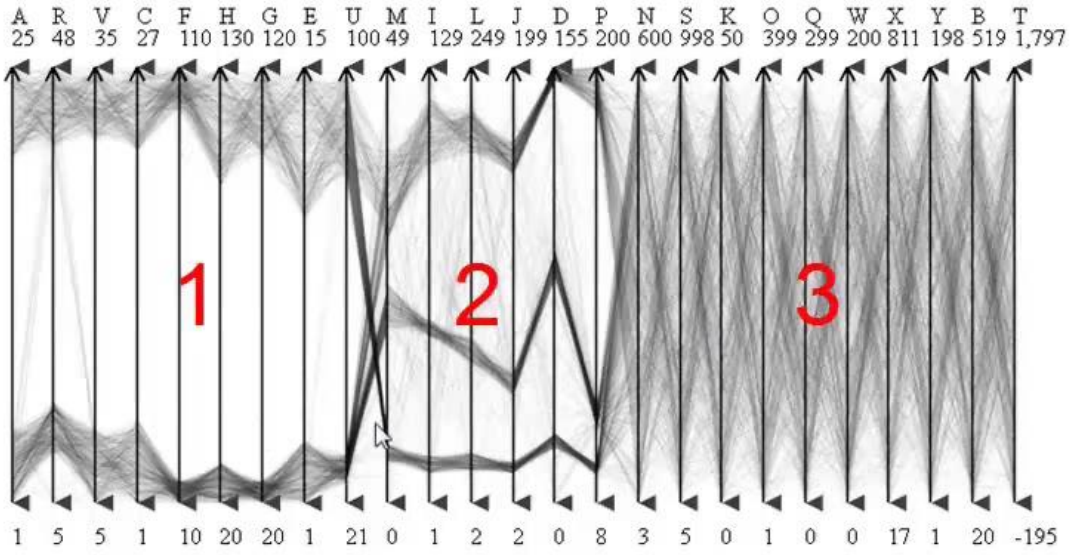


attribute centric

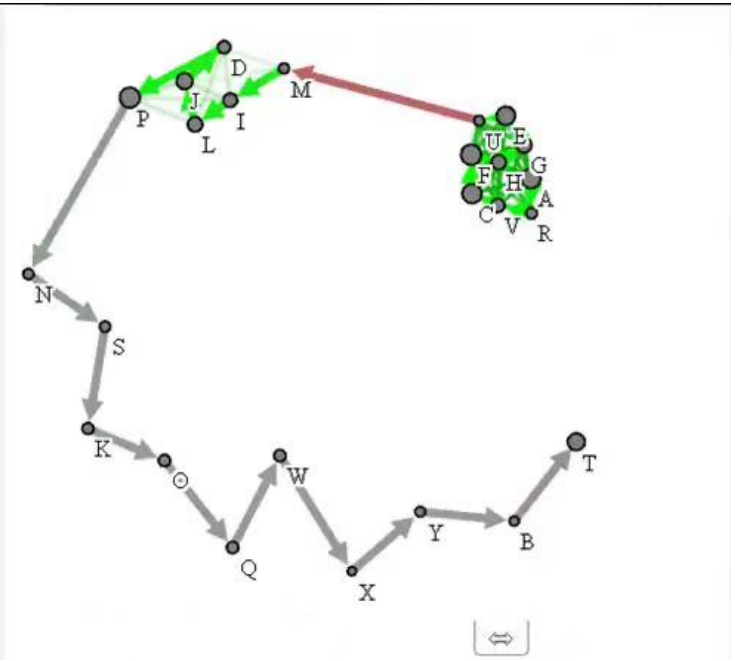


subset of attributes

Multiscale Zooming



3 subspaces are well seperated.

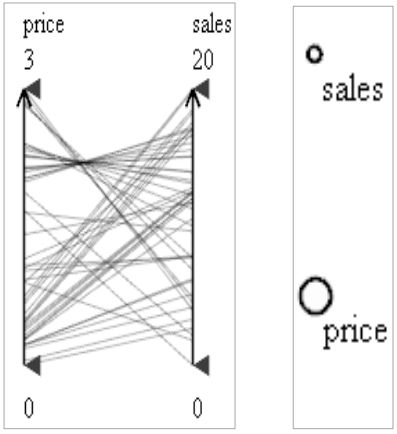


Exploring Correlation Sensitivity

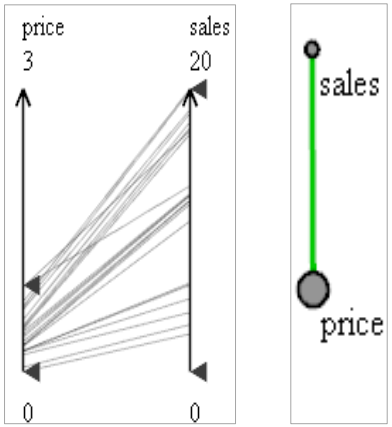
Correlation strength can often be improved by constraining a variable's value range (bracketing)

This limits the derived relationships to this value range

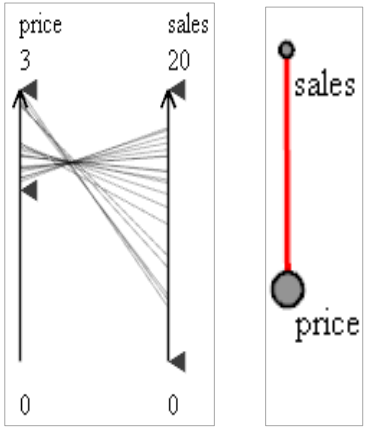
Such limits are commonplace in targeted marketing, etc.



no bracketing



lower price range

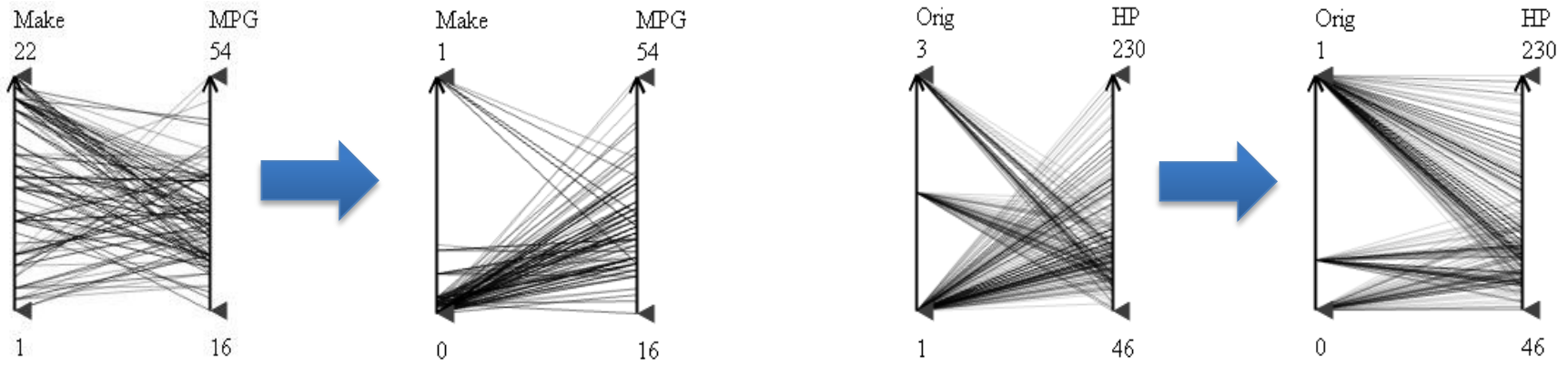


higher price range

Unifying Categorical and Numerical Variables

We transform categorical variables to numerical variables

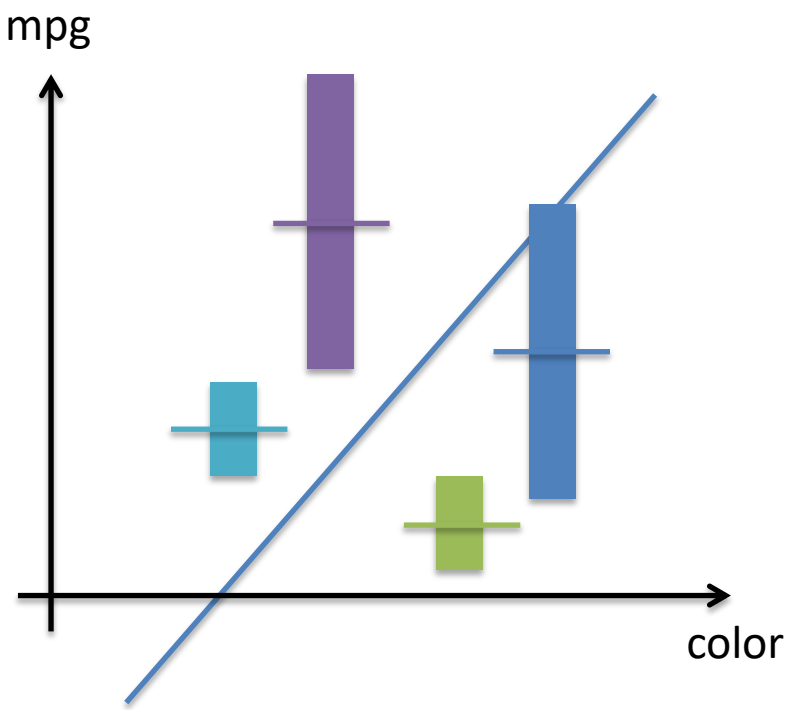
- use a pairwise correlation optimization approach



Correlations can be clearly better observed after transformations

Transformation Procedure

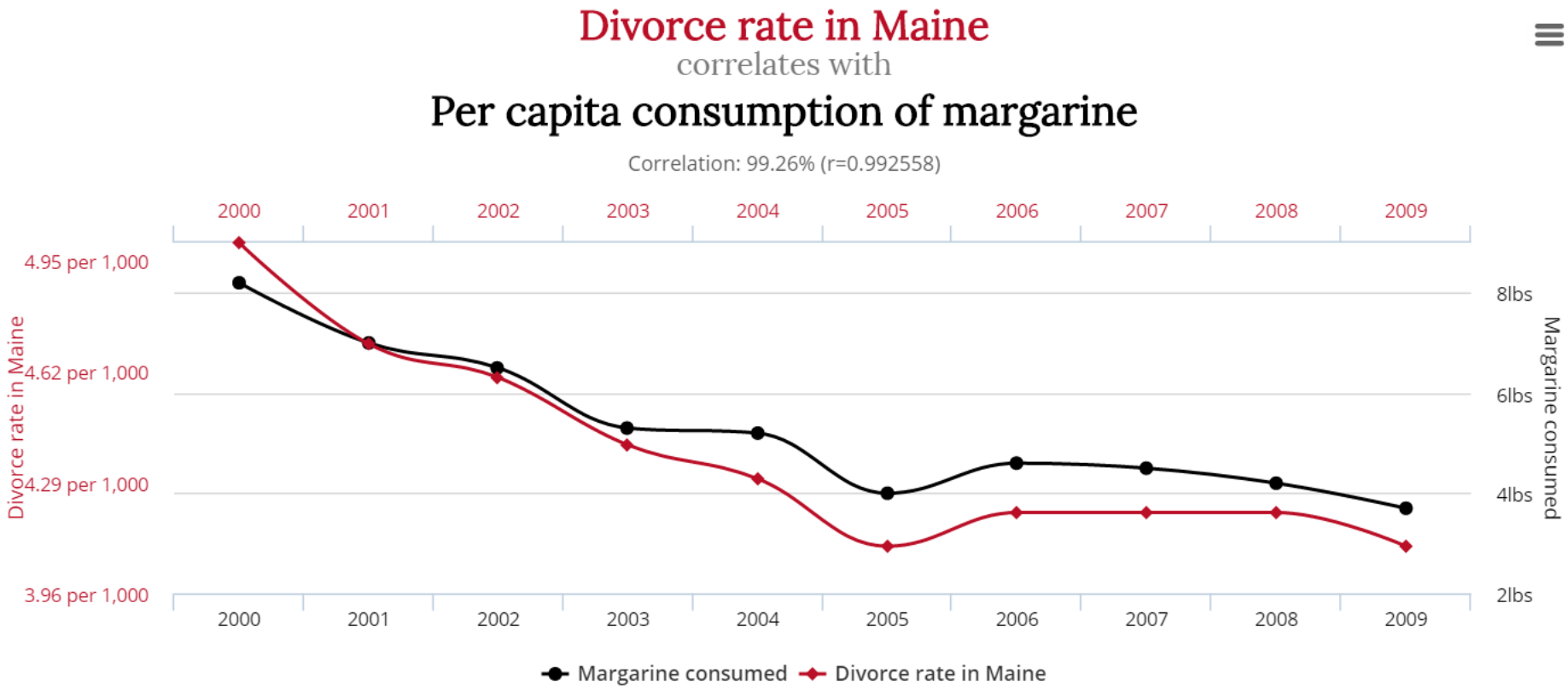
Applied to car data



But....

...do people talk about spurious correlations?

Spurious Correlations



Eat less margarine → save your marriage

Save your marriage → eat more butter

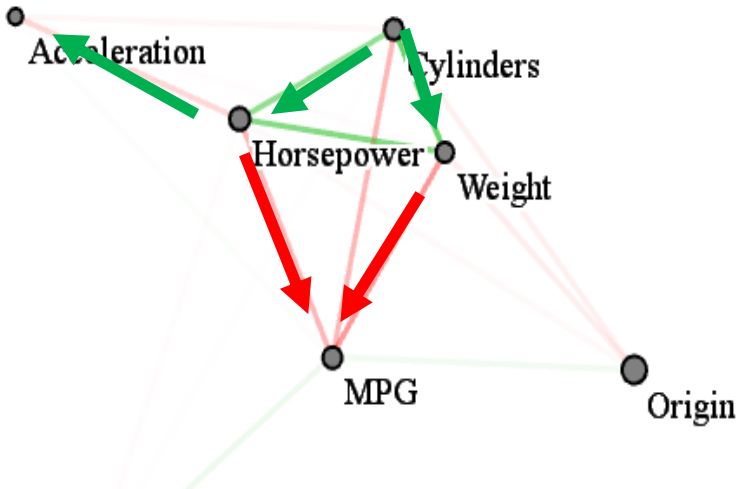
The Ultimate Goal: Causal Model

Controversial topic

- correlation ≠ causation
- promising casual inferencing algorithms exist
- but inferring causation from observational data remains shaky

Gives correlation links casual directions

- have the domain expert examine these and possible change them



THE VISUAL CAUSALITY ANALYST

The Algorithm

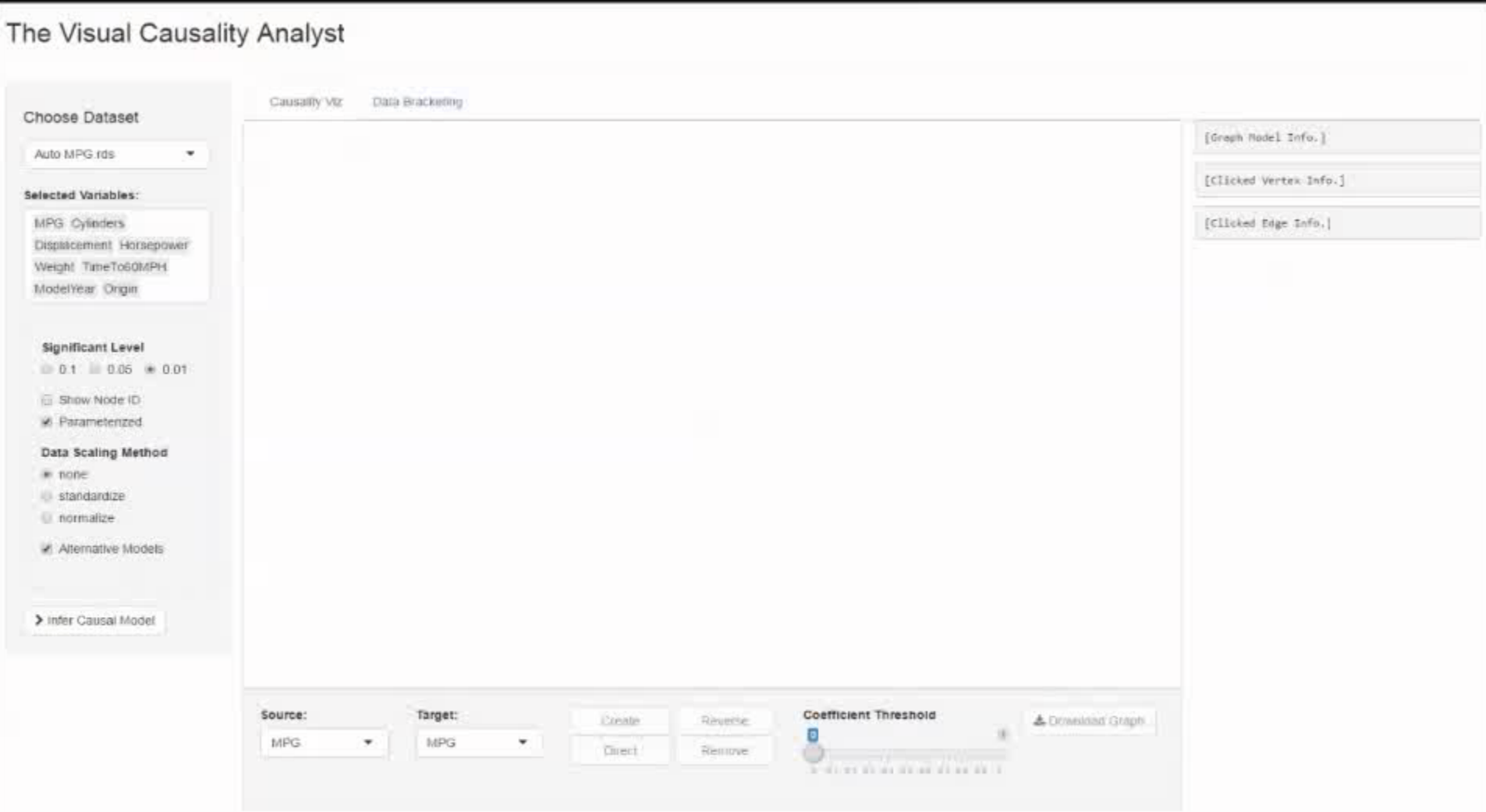
First, construct an initial DAG with a constraint-based scheme

- very fast (as opposed to slow exhaustive schemes)
- but incomplete

Then, allow the user to hypothesize edits to the DAG

- interactively add, delete, reverse casual edges
- check the score if the model is better or worse
- keep the edit DAG if score improves

The Visual Causality Analyst



Final Thoughts

Data too big?

- use clustering with stratified sampling
- abstract into model (correlation, casual, classification, HMM, etc..)

Data not numerical?

- images, video, text, etc.
- create feature vectors of numbers and you're ready to go

Data time-varying and streaming?

- cluster behaviors, not points
- perform subsequence discovery, clustering, and evolution

Want to try it?

- soon cloud-served for your favorite web browser

Credits

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

Faculty:

- Kevin T. McDonnell (Stony Brook), Wei Xu (BNL)

Domain scientists:

- Dr. Alla Zelenyuk, Dr. Dan Imre (PNL), Yangan Liu (BNL)

PhD students (at Stony Brook and SUNY Korea):

- Jenny Lee, Nafees Ahmed, Bing Wang, Puripant (Joe) Ruchikachorn, Sungsoo Ha, Jisung Kim, Jun Wang, Shenghui Cheng. Eric Papenhausen, Salman Mahmood, Ziyi Zheng (PhD, now at Google). Julia Nam (PhD, now at Microsoft), Zhiyuan Zhang (now at Facebook), Wen Zhong, Xie Cong, Darius Coelho

More Detail? Visit my Webpage…

http://www.cs.stonybrook.edu/~mueller

(for videos see dedicated paper web pages)

Any questions?