KMAP – A Visualizer for Kohonen Self-Organizing Map Data Mining Results

George S. Almasi
Staff Member Emeritus, IBM T. J. Watson Research Center
Vice President, Michael Rothman & Associates
almasi@gsalmasi.com, (914) 232-2378

Richard D. Lawrence
Staff Member, IBM T. J. Watson Research Center

Michael J. Rothman
President, Michael Rothman & Associates

DIMACS Workshop on Visualization and Data Mining, October 24, 2002

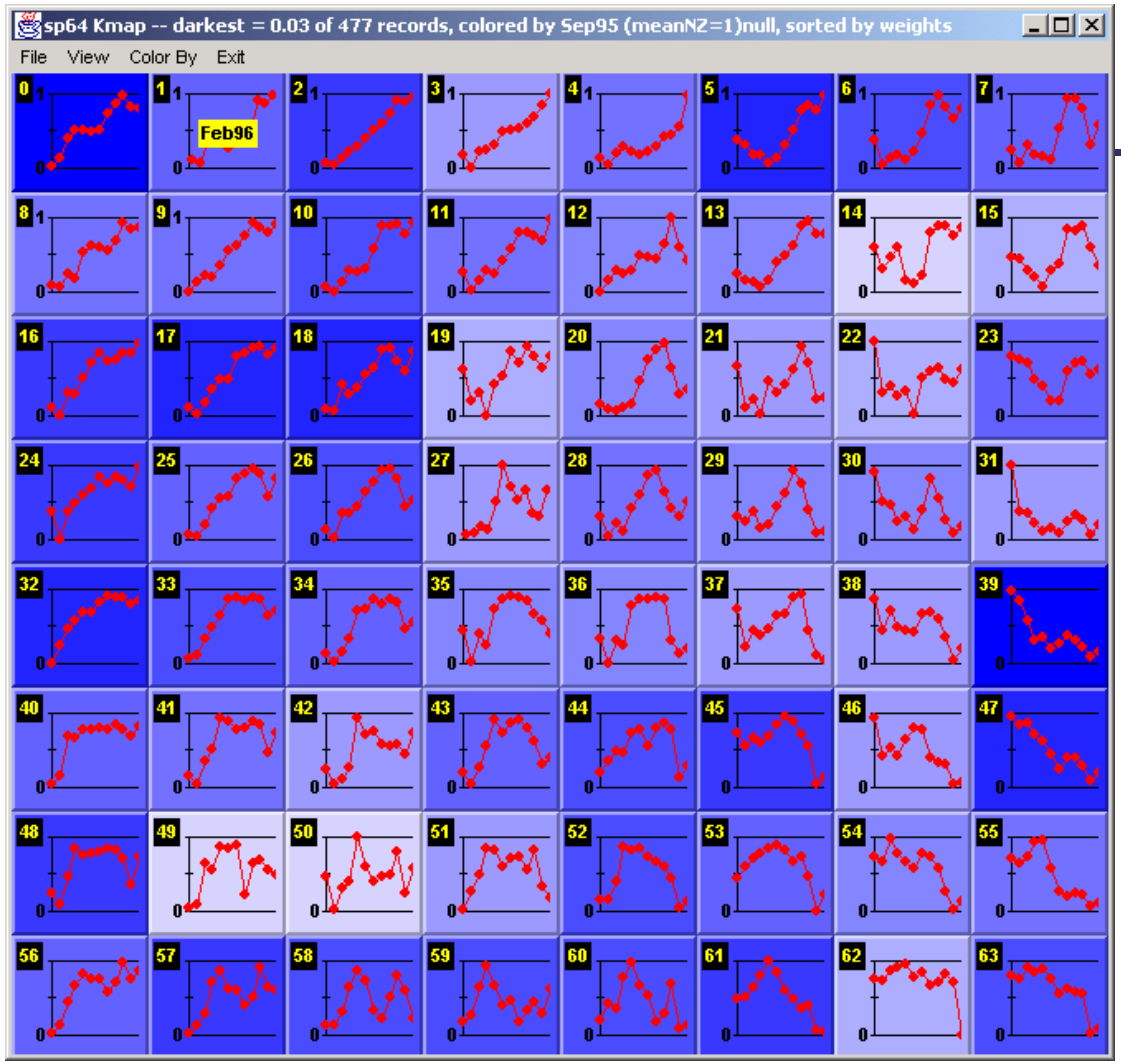
Kohonen Self-Organizing Map is a data mining technique for breaking a data set into clusters

- Similar records are clumped together
- Similar clumps are placed near each other in a 2-D map
- More precisely, similar records from a dataset are separated into clusters using a Euclidian metric, and then the multi-dimensional clusters are put in a 2-D map in such a way as to preserve the underlying topology

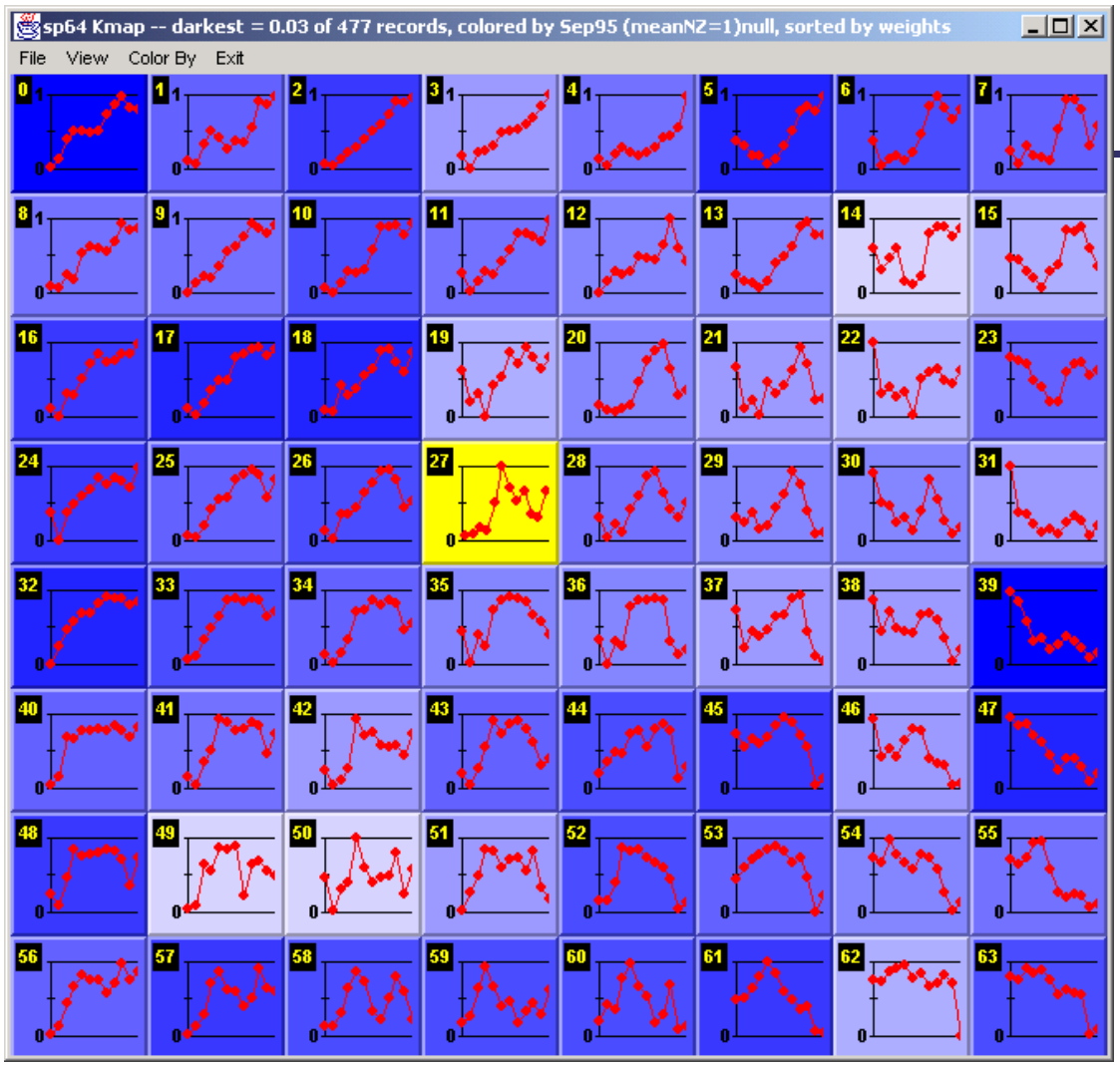
The KMAP visualizer shows Kohonen Map topology and allows interactive exploration of the clusters

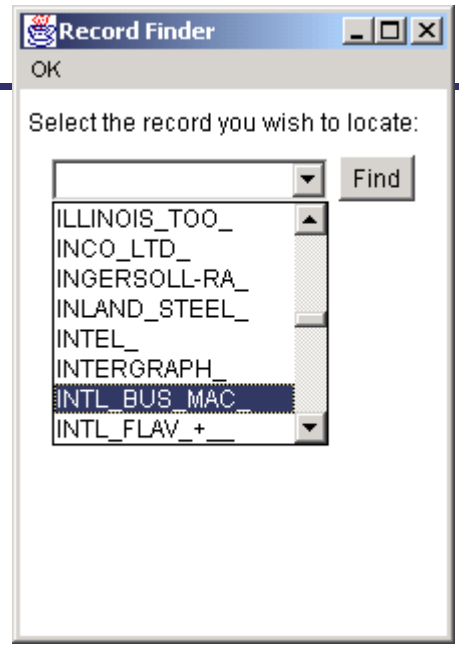
- Written in Java for portability
- Uses Sitraka JCChart Java Beans
- Can run as Web applet
- Several demos and view modes will be shown:
  1. Stock Market (S&P 500 “birds of a feather”)
     - The straightforward view, but with some surprises
  2. Supermarket Shoppers (for personalized recommendations)
     - The “top ten” view for data with many attributes
  3. Credit Bureau (bankruptcy prediction)
     1. Custom “patterns within patterns” view for “extracted scores”, a method for dealing with still more attributes and disparate data bases.

A years' comparison of S&P 500 month-end closing prices, broken into 64 clusters…

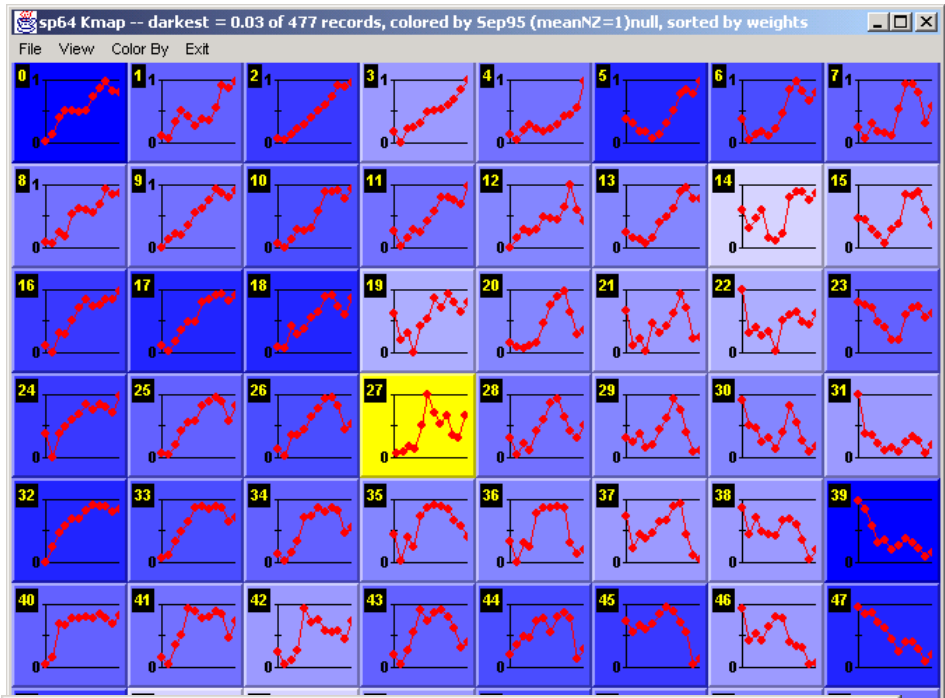


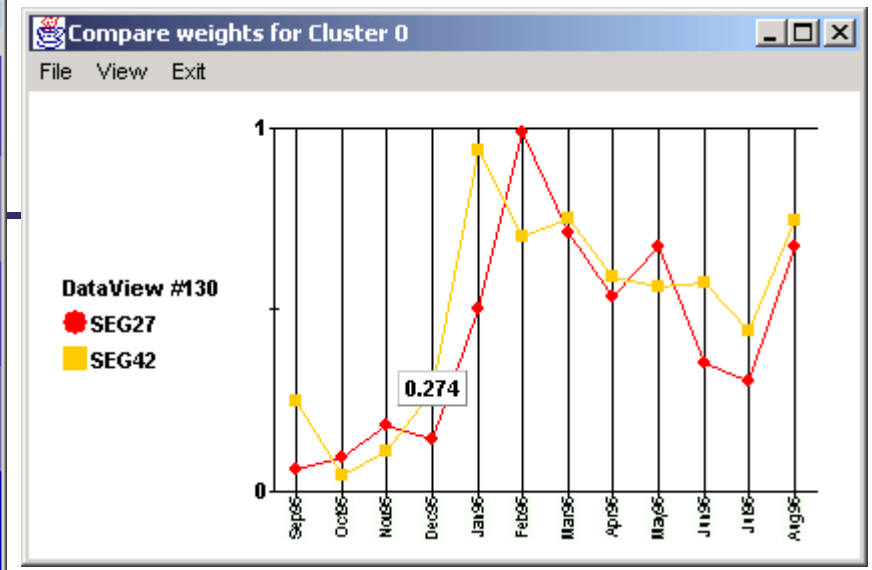
S&P 500 drilldown: where's IBM?



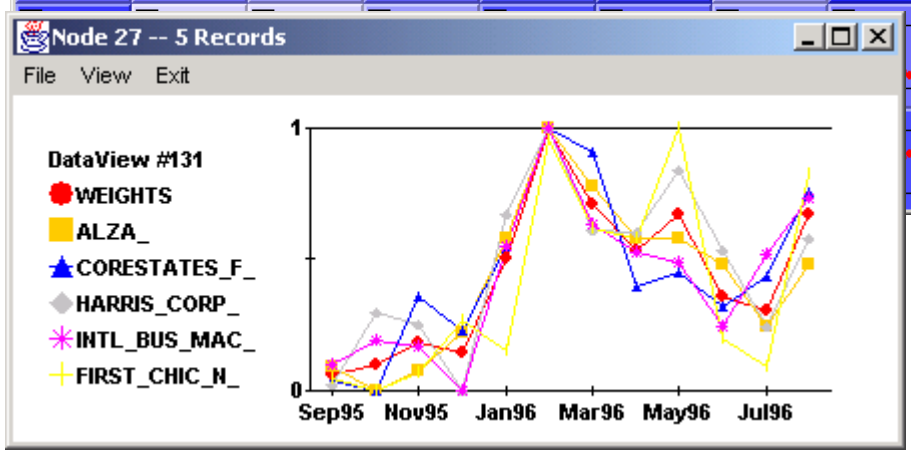


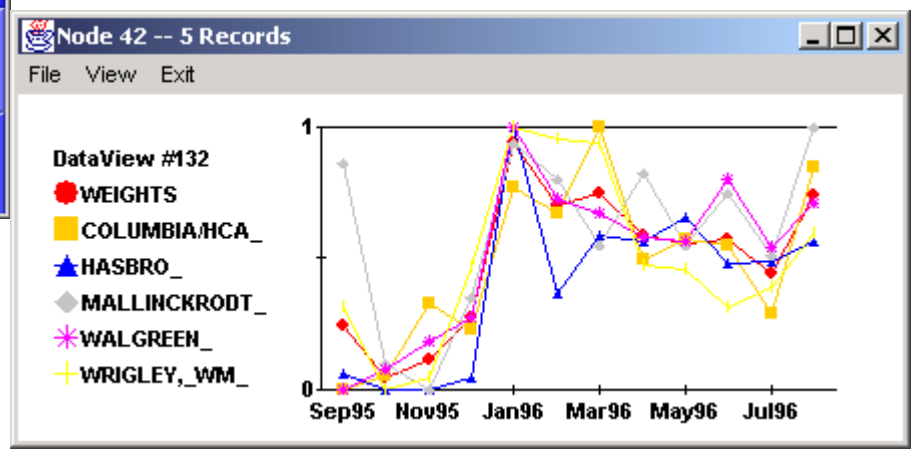
Can one stock predict another's behavior?





Does Hasbro predict IBM?





The KMAP visualizer shows Kohonen Map topology and allows interactive exploration of the clusters

- Written in Java for portability
- Uses Sitraka JCChart Java Beans
- Can run as Web applet
- Several demos and view modes will be shown:
  1. Stock Market (S&P 500 "birds of a feather")
     - The straightforward view, but with some surprises
  2. Supermarket Shoppers (for personalized recommendations)
     - The "top ten" view for data with many attributes
  3. Credit Bureau (bankruptcy prediction)
     1. Custom "patterns within patterns" view for "extracted scores", a method for dealing with still more attributes and disparate data bases.

Supermarket shoppers fall into distinct clusters

supermarket Kmap , colrd by Total_spend (mean=191.168) -- grayest = 5% of 35957 recs, sorted by weights

File View Color By Sort By Exit

0 33.93 2K(5%)	1 17.49 1K(3%)	2 45.39 865(2%)	3 73.96 986(3%)	4 111.03 90			
PETROL	MILK	INSTOR_BAKRY	CARRRS/GREEN	CARRRS/GRE			
BREAD/MORNING	INSTOR_BAKRY	CONFECTIONERY	CONFECTIONERY	BULK_SLC_M			
HEALTH/BEAUTY	BREAD/MORNING	CSP/SNAX/NUTS	CSP/SNAX/NUTS	INSTOR_BAKI			
6 29.57 1K(3%)	7 39.36 855(2%)	8 66.51 858(2%)	9 90.84 782(2%)	10 129.12 83			
CARRRS/GREEN	F/V_-_SALADS	MILK	CARRRS/GREEN	HOMEBAKER			
INSTOR_BAKRY	F/V_-_VEGGIES	BISCUITS	TEA	CARRRS/GRE			
MILK	F/V_-_FRUIT_1	BREAK-CEREALS	MARGARINE	SUGAR			
12 78.56 1K(3%)	13 65.68 920(3%)	14 67.09 737(2%)	15 102.48 861(2%	16 198.56 88			
TABLE_WINES	F/V-P_PRODUCE	CEREALS/PASTA	HSHOLD_POLISH	BUTTER/FATS			
BEERS/LAGER	F/V_-_SALADS	COOK_PRODS	HM_LNDRY/SOAP	SUGAR			
CARRRS/GREEN	F/V_-_VEGGIES	CAND/PKT_VEGS	PAPER_PRODS	TEA			
18 301.44 1K(3%)	19 181.81 1K(3%)	20 178.2 1K(3%)	21 178.28 950(3%	22 215.28 98			
SPIRITS	FRESH_BEEF	INTERNATIONAL	CANNED_FISH	CANNED_FRUIT	MILK/CREAM		
TABLE_WINES	PORK/LAMB	CEREALS/PASTA	CEREALS/PASTA	DESSERTS/PUDD	CANNED_FRUIT		
BEERS/LAGER	FRESH_POULTRY	COOK_PRODS	CAND/PKT_VEGS	BREAK-CEREALS	DESSERTS/PUDD		
24 345.65 902(3%	25 232.1 908(3%)	26 368.48 932(3%	27 269.06 1K(3%)	28 408.19 1K(3%)	29 204.56 1K(3%)		
TABLE_WINES	DELICATESSEN	INTERNATIONAL	HAIR_CARE	FRZ_POULTRY	CANNED_MEAT		
COOK_SALD_OIL	PRPK_SALADS	HAIR_CARE	HEALTH/BEAUTY	FRZ_MEAT	TEA		
F/V-P_PRODUCE	PRPK_SAUSAGES	CEREALS/PASTA	FRZ_PROC_MEAT	FRZ_SAUSAGE	MARGARINE		
30 265.12 987(3%	31 258.29 936(3%	32 379.03 992(3%	33 314.57 1K(3%)	34 255.41 1K(3%)	35 268.49 955(3%		
FRUIT_JCE-FR	F/V_-_HORTIC	CLOTHING	BABY_PRODS	FRZ_PROC_MEAT	CANNED_MEAT		
DELICATESSEN	NEWS/MAGS	LEISURE	CANND_PASTA	CSP/SNAX/NUTS	CANND_PASTA		
FISH	F/V-P_PRODUCE	HEALTH/BEAUTY	DAIRY-FRESH	PRPK_SLC_MEAT	CAND/PKT_VEGS		

Drilldown:

Node 33 -- 1163 Records, Total_spend=315, Age=32.055

File View Sort By Exit

Field	Wgt	mean	R	meanNZ	R	partcp	R
BABY_PRODS	2.909	34.16	12.90	34.22	2.40	1.00	5.3
CANND_PASTA	1.847	2.25	4.94	2.80	1.68	0.80	2.9
DAIRY-FRESH	1.413	10.76	2.55	11.59	1.73	0.93	1.4
CSP/SNAX/NUTS	1.205	7.93	2.05	8.70	1.39	0.91	1.4
HM_LNDRY/SOAP	1.186	10.99	1.97	12.18	1.34	0.90	1.4
FRZ_PROC_MEAT	1.092	8.64	1.93	10.96	1.22	0.79	1.5
MILK	1.087	9.16	1.82	9.75	1.55	0.94	1.1
BREAK-CEREALS	1.079	7.11	1.83	8.15	1.29	0.87	1.4
DESSERTS/PUDD	1.068	1.94	2.24	3.10	1.29	0.63	1.7
CONFECTIONERY	1.036	7.03	1.85	8.13	1.28	0.86	1.4

P. S.: This cluster's favorite candy bar is quite different from the overall favorite!

“Wallet share” offers a second view…

BABY_PRODS Kmap; darkest shade = 42% market share of 95,238

File View Color By Sort By Exit

	0..weight 0.	1..weight 0.	2..weight 0.	3..weight 0.	4..weight 0.	5..weight 0.
	02% Val	01% Val	01% Val	01% Val	01% Val	02% Val
m	1 0.33	0 0.17	1 0.38	1 0.32	1 0.23	2 0.67
mNZ	12 0.85	9 0.63	10 0.67	7 0.46	6 0.45	10 0.67
ptp	0.07 0.3	0.05 0.2	0.11 0.5	0.13 0.7	0.1 0.51	0.18 0.9

	6..weight 0.	7..weight 0.	8..weight 0.	9..weight 0.	10..weight 0.	11..weight 0.
	01% Val	00% Val	01% Val	00% Val	01% Val	02% Val
m	1 0.34	0 0.13	1 0.43	0 0.17	1 0.37	1 0.52
mNZ	12 0.83	10 0.72	10 0.73	6 0.45	8 0.57	8 0.56
ptp	0.08 0.4	0.03 0.1	0.11 0.5	0.07 0.3	0.12 0.6	0.17 0.9

	12..weight 0.	13..weight 0.	14..weight 0.	15..weight 0.	16..weight 0.	17..weight 0.
	01% Val	01% Val	00% Val	02% Val	01% Val	01% Val
m	1 0.25	1 0.29	1 0.23	2 0.73	1 0.49	1 0.31
mNZ	9 0.63	11 0.78	7 0.52	12 0.84	7 0.5	6 0.41
ptp	0.07 0.4	0.07 0.3	0.08 0.4	0.16 0.8	0.18 0.9	0.14 0.7

	18..weight 0.	19..weight 0.	20..weight 0.	21..weight 0.	22..weight 0.	23..weight 0.
	01% Val	01% Val	02% Val	01% Val	01% Val	01% Val
m	1 0.43	1 0.41	2 0.58	1 0.49	1 0.31	1 0.31
mNZ	8 0.56	8 0.59	11 0.76	9 0.62	6 0.39	6 0.4
ptp	0.14 0.7	0.13 0.6	0.14 0.7	0.15 0.7	0.15 0.7	0.15 0.8

	24..weight 0.	25..weight 0.	26..weight 0.	27..weight 0.	28..weight 0.	29..weight 0.
	03% Val	02% Val	04% Val	01% Val	04% Val	01% Val
m	4 1.35	2 0.68	4 1.63	1 0.48	4 1.49	1 0.21
mNZ	16 1.11	11 0.78	13 0.93	5 0.34	12 0.85	5 0.35
ptp	0.23 1.2	0.16 0.8	0.33 1.7	0.27 1.4	0.33 1.7	0.11 0.6

	30..weight 0.	31..weight 0.	32..weight 0.	33..weight 2.	34..weight 0.	35..weight 0.
	03% Val	02% Val	09% Val	42% Val	01% Val	03% Val
m	3 1.14	2 0.68	8 3.11	34 12.9	1 0.26	3 1.28
mNZ	13 0.91	9 0.61	19 1.32	34 2.4	4 0.28	11 0.77
ptp	0.23 1.2	0.21 1.1	0.44 2.3	1 5.36	0.17 0.9	0.31 1.6

SPIRITS Kmap; darkest shade = 47% market share of 113,756

File View Color By Sort By Exit

	0..weight 0.	1..weight 0.	2..weight 0.	3..weight 0.	4..weight 0.	5..weight 0.
	02% Val	01% Val	00% Val	00% Val	01% Val	02% Val
m	2 0.53	1 0.21	0 0.11	1 0.17	1 0.26	2 0.55
mNZ	33 1.25	19 0.72	14 0.53	13 0.49	19 0.71	15 0.58
ptp	0.05 0.4	0.03 0.2	0.02 0.2	0.04 0.3	0.04 0.3	0.11 0.9

	6..weight 0.	7..weight 0.	8..weight 0.	9..weight 0.	10..weight 0.	11..weight 0.
	01% Val	00% Val	01% Val	01% Val	01% Val	04% Val
m	1 0.44	0 0.11	1 0.25	1 0.23	1 0.25	4 1.11
mNZ	30 1.14	16 0.62	20 0.76	13 0.51	14 0.52	23 0.89
ptp	0.05 0.3	0.02 0.1	0.04 0.3	0.05 0.4	0.06 0.4	0.15 1.2

	12..weight 0.	13..weight 0.	14..weight 0.	15..weight 0.	16..weight 0.	17..weight 0.
	03% Val	00% Val	00% Val	01% Val	02% Val	02% Val
m	4 1.18	1 0.16	0 0.09	1 0.28	3 0.83	3 0.92
mNZ	27 1.02	19 0.72	10 0.39	18 0.68	21 0.79	19 0.7
ptp	0.14 1.1	0.03 0.2	0.03 0.2	0.05 0.4	0.13 1.0	0.16 1.3

	18..weight 2.	19..weight 0.	20..weight 0.	21..weight 0.	22..weight 0.	23..weight 0.
	47% Val	01% Val	01% Val	01% Val	01% Val	02% Val
m	44 14	1 0.37	1 0.41	1 0.28	1 0.27	2 0.7
mNZ	44 1.68	20 0.76	13 0.47	14 0.52	15 0.58	21 0.8
ptp	1 8.34	0.06 0.4	0.1 0.87	0.06 0.5	0.06 0.4	0.1 0.87

	24..weight 0.	25..weight 0.	26..weight 0.	27..weight 0.	28..weight 0.	29..weight 0.
	05% Val	01% Val	02% Val	01% Val	02% Val	01% Val
m	7 2.09	2 0.53	3 0.85	2 0.49	2 0.76	1 0.4
mNZ	25 0.96	15 0.58	15 0.58	16 0.59	20 0.75	18 0.68
ptp	0.26 2.1	0.11 0.9	0.18 1.4	0.1 0.83	0.12 1.0	0.07 0.9

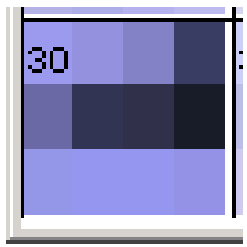
	30..weight 0.	31..weight 0.	32..weight 0.	33..weight 0.	34..weight 0.	35..weight 0.
	01% Val	02% Val	03% Val	01% Val	02% Val	02% Val
m	2 0.52	2 0.69	3 1.04	1 0.38	2 0.48	2 0.7
mNZ	22 0.81	17 0.66	20 0.76	13 0.5	16 0.62	19 0.71
ptp	0.08 0.6	0.13 1.0	0.16 1.3	0.09 0.7	0.09 0.7	0.12 1

The KMAP visualizer shows Kohonen Map topology and allows interactive exploration of the clusters

- Written in Java for portability
- Uses Sitraka JCChart Java Beans
- Can run as Web applet
- Several demos and view modes will be shown:
  1. Stock Market (S&P 500 "birds of a feather")
     - The straightforward view, but with some surprises
  2. Supermarket Shoppers (for personalized recommendations)
     - The "top ten" view for data with many attributes
  3. Credit Bureau (bankruptcy prediction)
     1. Custom "patterns within patterns" view for "extracted scores", a method for dealing with still more attributes and disparate data bases.

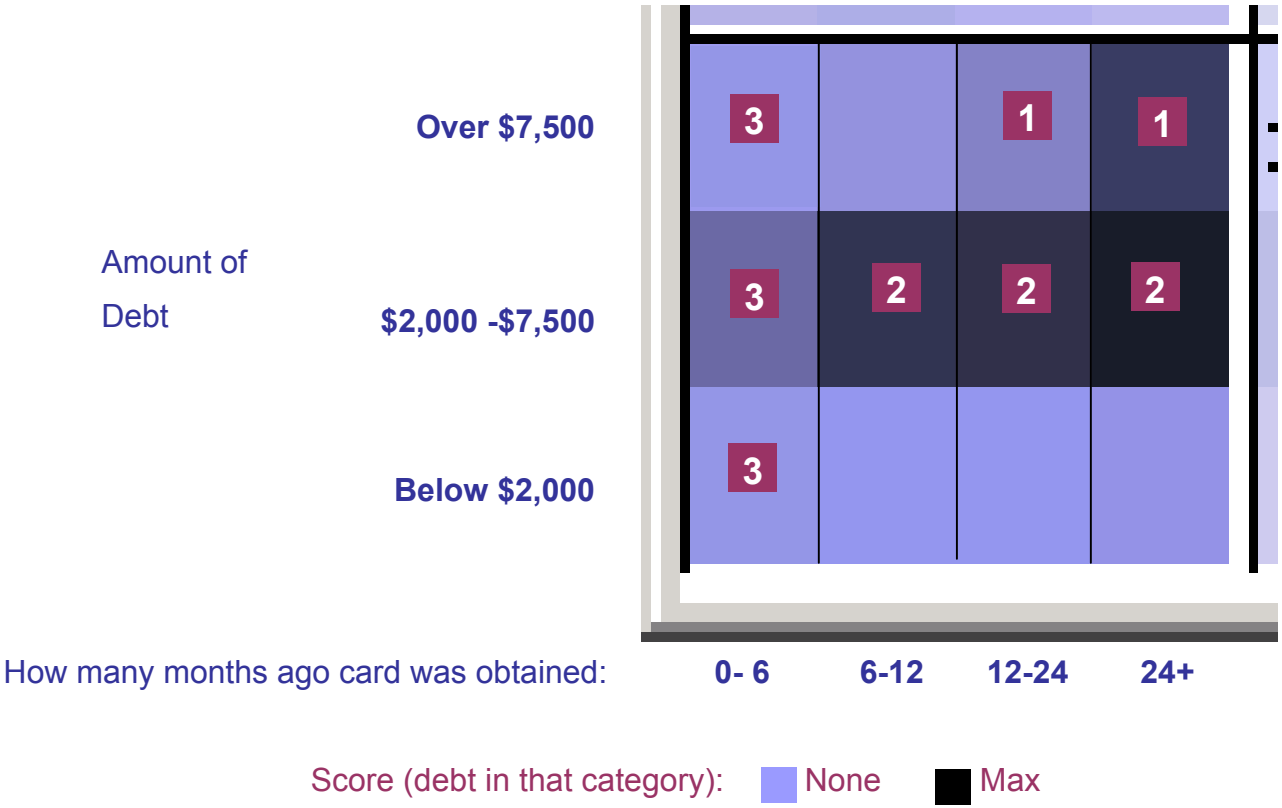
Example: Bankruptcy Prediction

- **We applied Score Extraction to bankruptcy prediction:**
  1. Complex Credit Bureau data (200+ items per record) was boiled down by dividing each household's total debt into 12 categories, based on the amount of each debt and how long ago it was incurred.
  2. A data mining program divided these households into 36 groups, based on these new categories.
  3. Checking subsequent bankruptcy history of each group revealed one group with unusually high bankruptcy rate
  4. Placing each group's debt attributes in a small 3x4 checkerboard pattern shows the characteristic "signature" below for the high bankruptcy group:



This is explained in the next chart…

The Analyst can see a Bankruptcy Signature

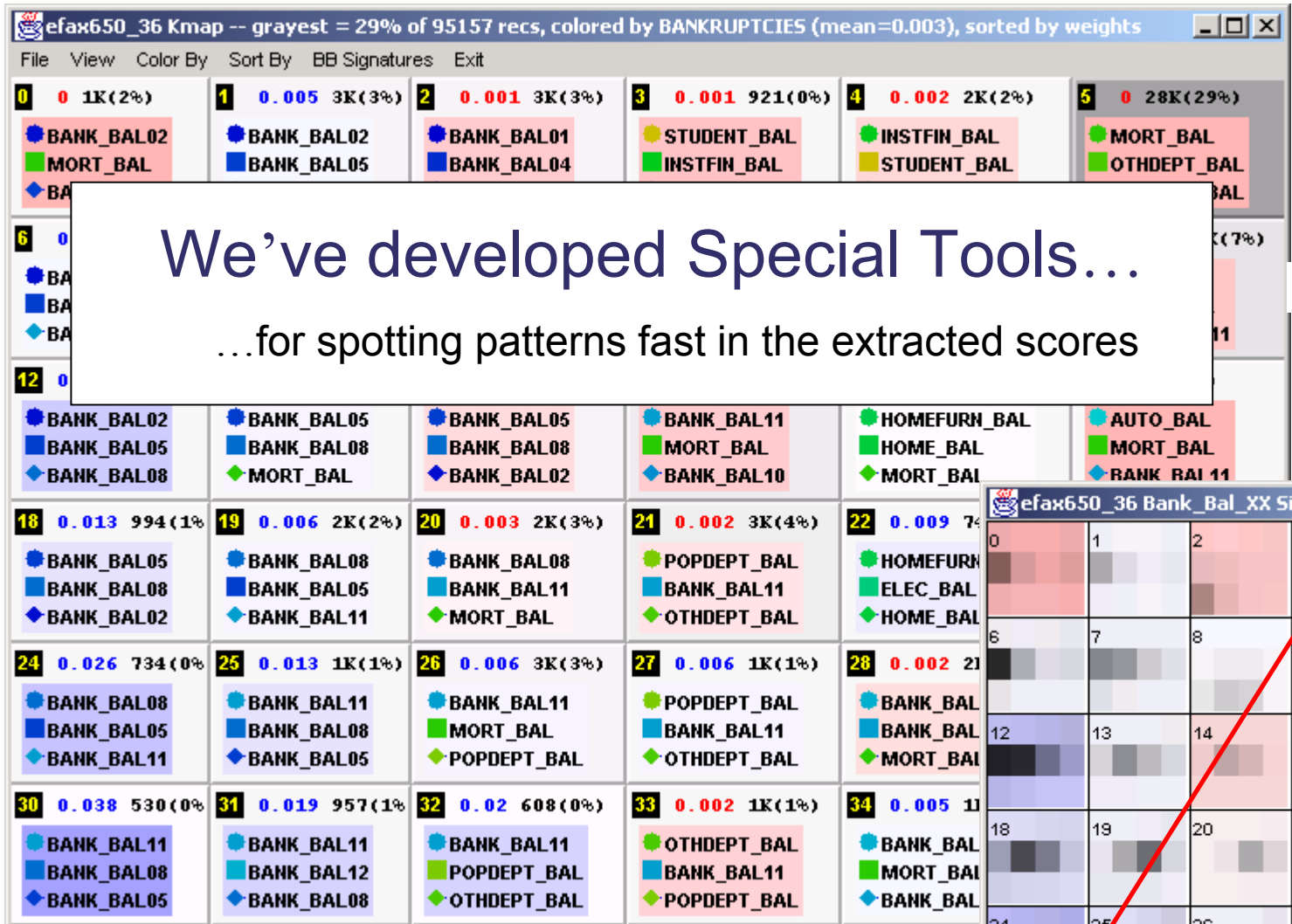


What an analyst sees here:

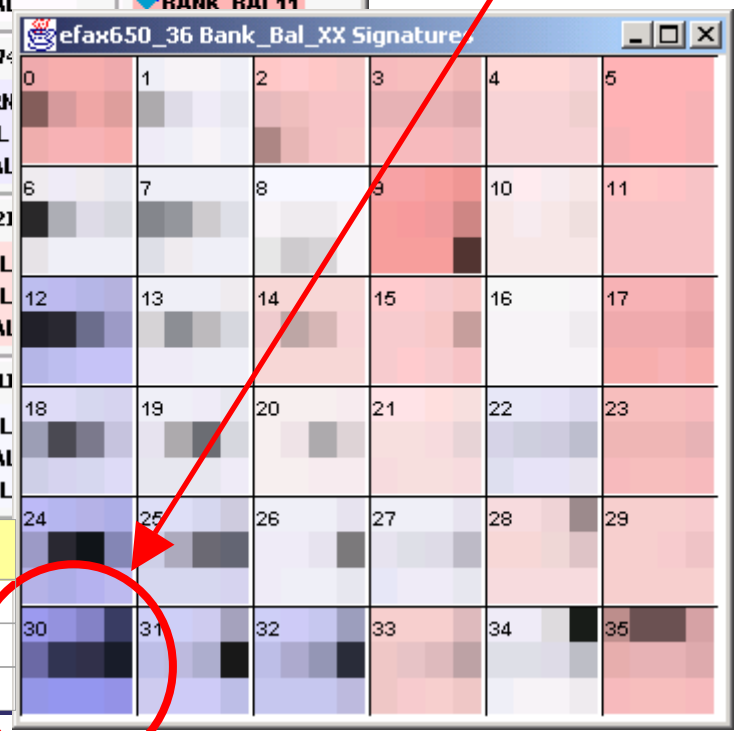
1. they used to be able to get credit lines over $7500, but not in the last year
2. a "sweet spot" in credit lines between $2000 & $7500… easy to get and they quickly add up to a lot of money
3. have been unable to get new cards in the last 6 months as their "risk" score elevates

We've developed Special Tools…

…for spotting patterns fast in the extracted scores



	under 6 mo.	6 to 12 mo.	12 to 24	over 24
over $7500	Bal_03	Bal_06	Bal_09	Bal_12
$2000 to $7500	Bal_02	Bal_05	Bal_08	Bal_11
under $2000	Bal_01	Bal_04	Bal_07	Bal_10



These 530 households (0.6% of the total) Have high Probability Of bankruptcy

Michael Rothman & Associates, LLC
mjrothman@preference-engine.com

Summary: KMAP is a versatile tool for visualizing data mining results

For more information:

1. Demos & documents at our web site: www.preference-engine.com/
2. Credit bureau data study: www.twocrows.com/largedb.pdf
3. Personalized recommendations: R. D. Lawrence, G. S. Almasi et al, "Personalization of Supermarket Product Recommendations", *Data Mining and Knowledge Discovery* 5(1/2): 11-32 (2001) …
4. … and: Almasi et al, U. S. Patent 6,260,036 (2001)
5. High-speed Kohonen clustering program that made Kmap necessary: R. D. Lawrence, G. S. Almasi, H. E. Rushmeier, "A Scalable Parallel Algorithm for Self-Organizing Maps with Applications to Sparse Data Mining Problems", *Data Mining and Knowledge Discovery* 3(2):171-195 (1999).
6. George S. Almasi, almasi@gsalmasi.com, 914-232-2378