

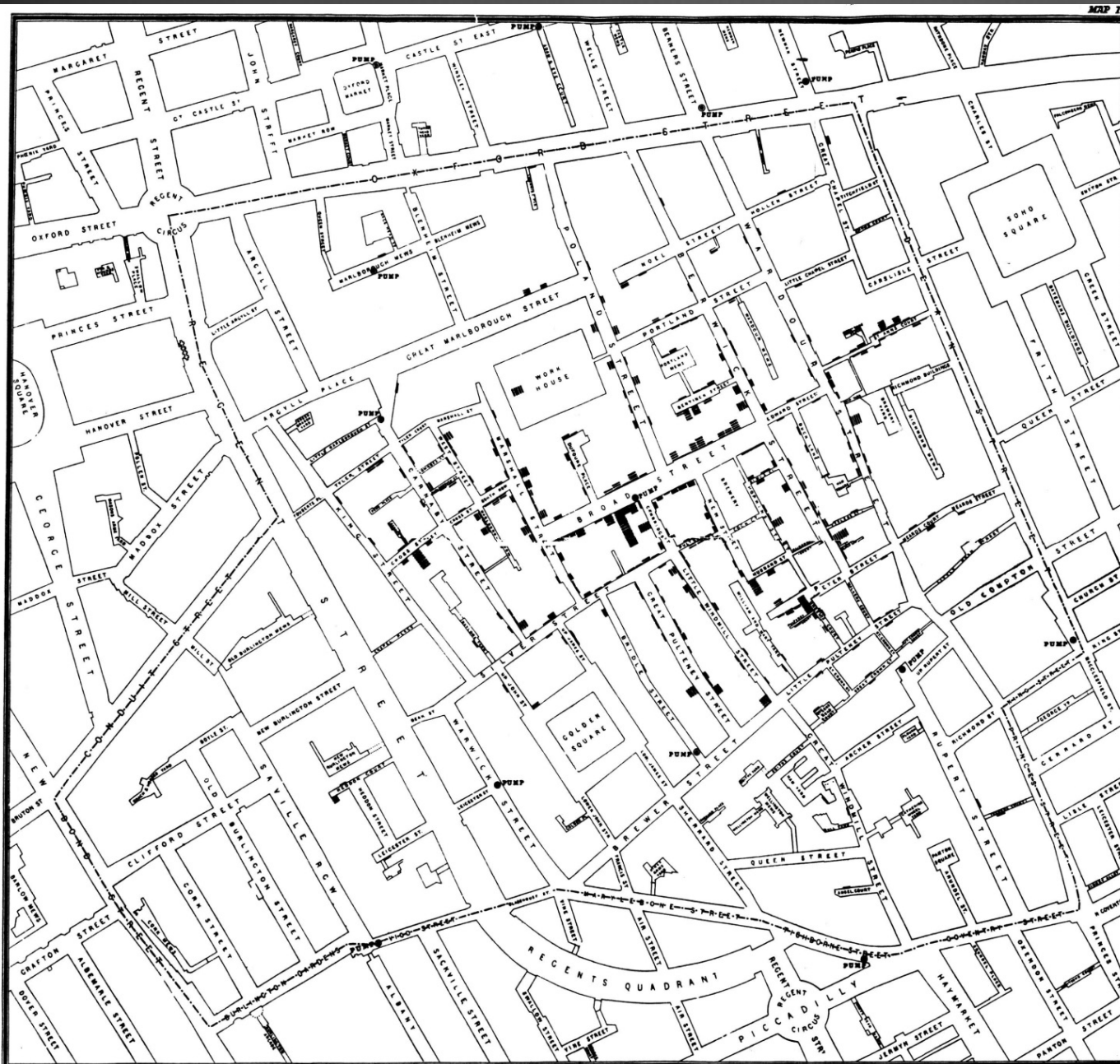
Biostat 202:Lecture 5

Analytics & Visualization: Telling Stories with Data



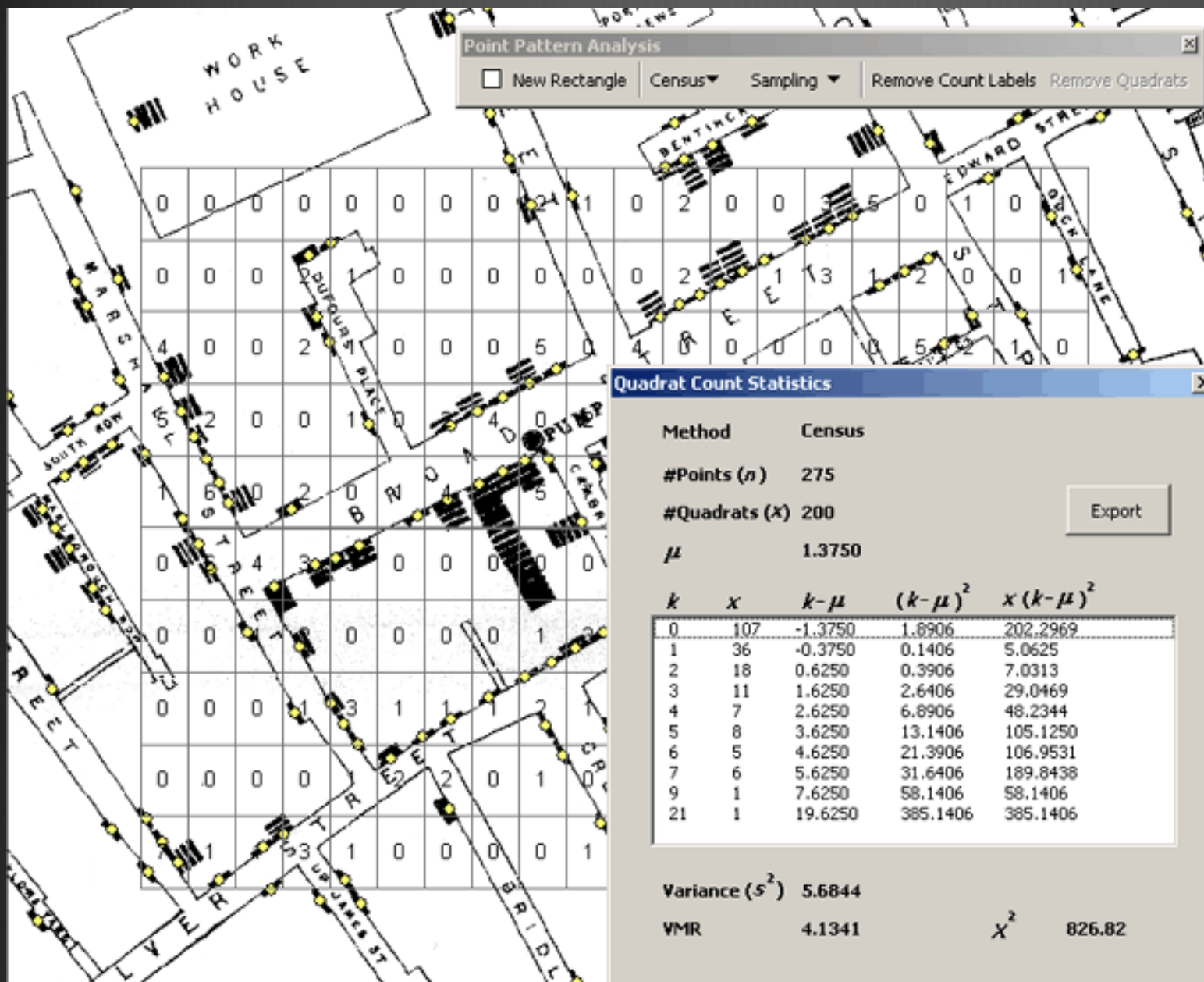
Isabel Elaine Allen, PhD
Professor, Epidemiology & Biostatistics
isabel.allen@ucsf.edu

...where am I and what is missing in this picture?



Decoding the Outliers:
John Snow's Cholera
Map, 1854

**"Know There Are Unknowns: How suspected confounding variables influence models," I. Elaine Allen & Julia E. Seaman, Journal of Quality Progress, February, 2011*



Creating the grid to digitize John Snow's original map*.

*CartoDB & Digitizelt software: https://docs.cartodb.com/tutorials/conditional_styling/



John Snow's Cholera Map,
recreated from data
extracted by the London
Daily Examiner, 2013

*What is unusual
about the pattern of
Cholera cases?*



John Snow's Cholera Map, recreated from data extracted by the London Daily Examiner, 2013

What is unusual about the pattern of Cholera cases?

Uncovering the message: Emergency Room visits*

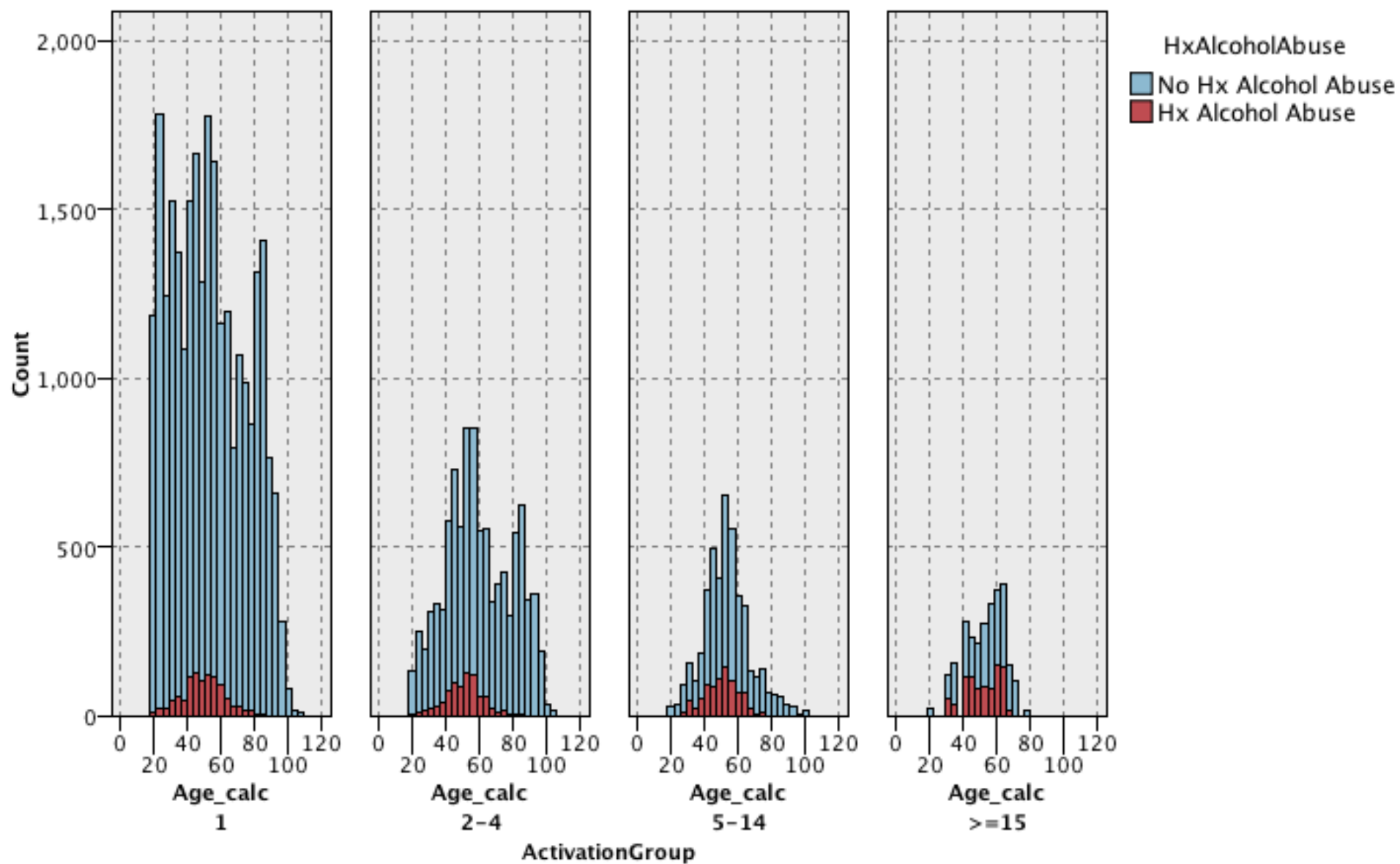
Frequency groups	1	2-4	5-14	15+	Total
Number (Percent) of people	26682 (84.8%)	4075 (13.0%)	605 (1.9%)	100 (0.3%)	31462 (100%)
Number (Percent) of Activations	26682 (61.2%)	9773 (22.4%)	4436 (10.2%)	2668 (6.1%)	43559 (100%)
Number (Percent) of Transports	23311 (59.6%)	8998 (23.0%)	4224 (10.8%)	2574 (6.6%)	39107 (100%)

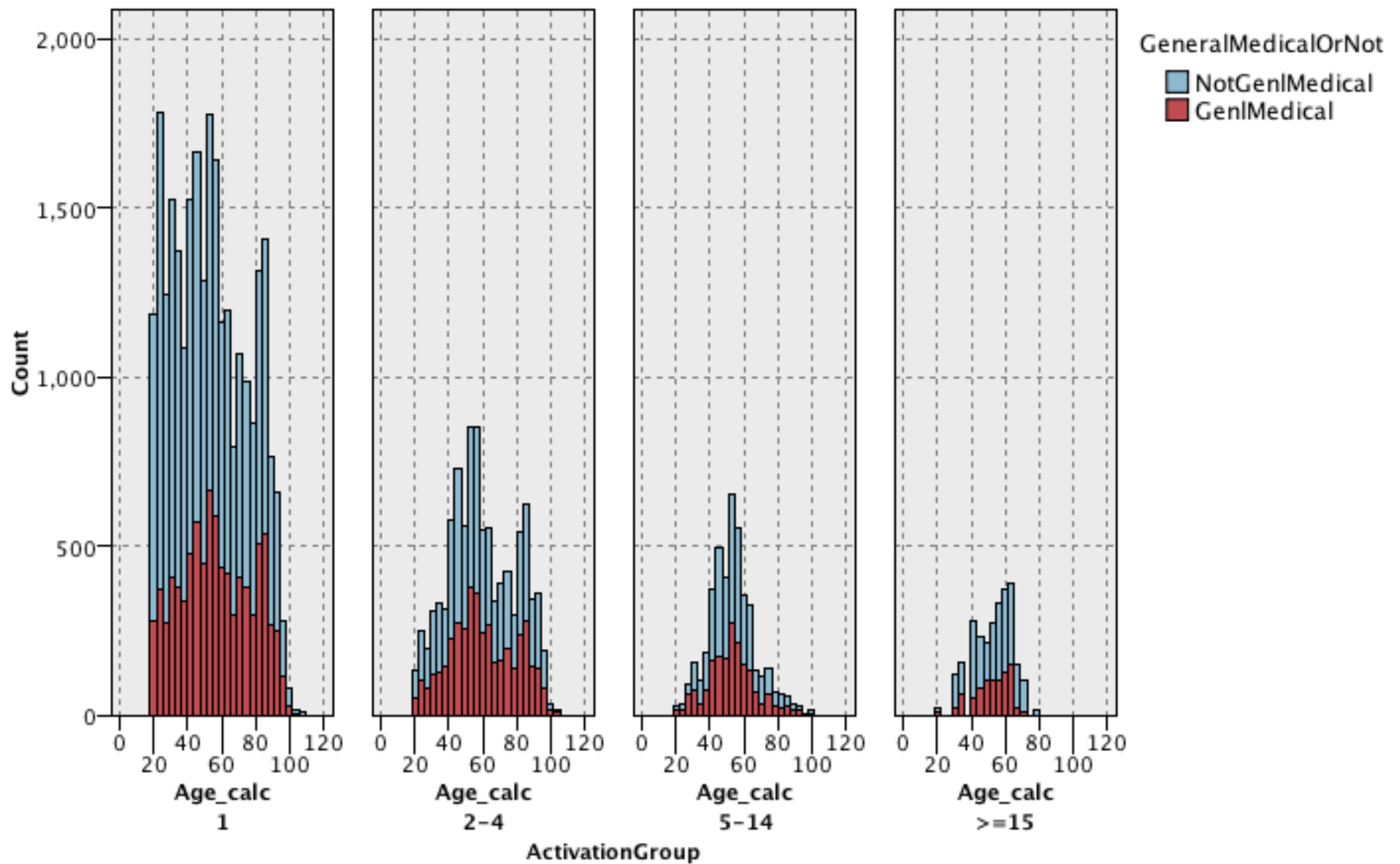
	1	2-4	5-14	15+	Total
Average Age (Mean, 95% CI)	53 ($\pm$ 0.26)	60 ($\pm$ 0.61)	52 ($\pm$ 1.15)	49 ($\pm$ 2.22)	55 ($\pm$ 0.24)

The table of mean age shows no significant difference between age groups – can we refine this result? Who are the frequent flyers?

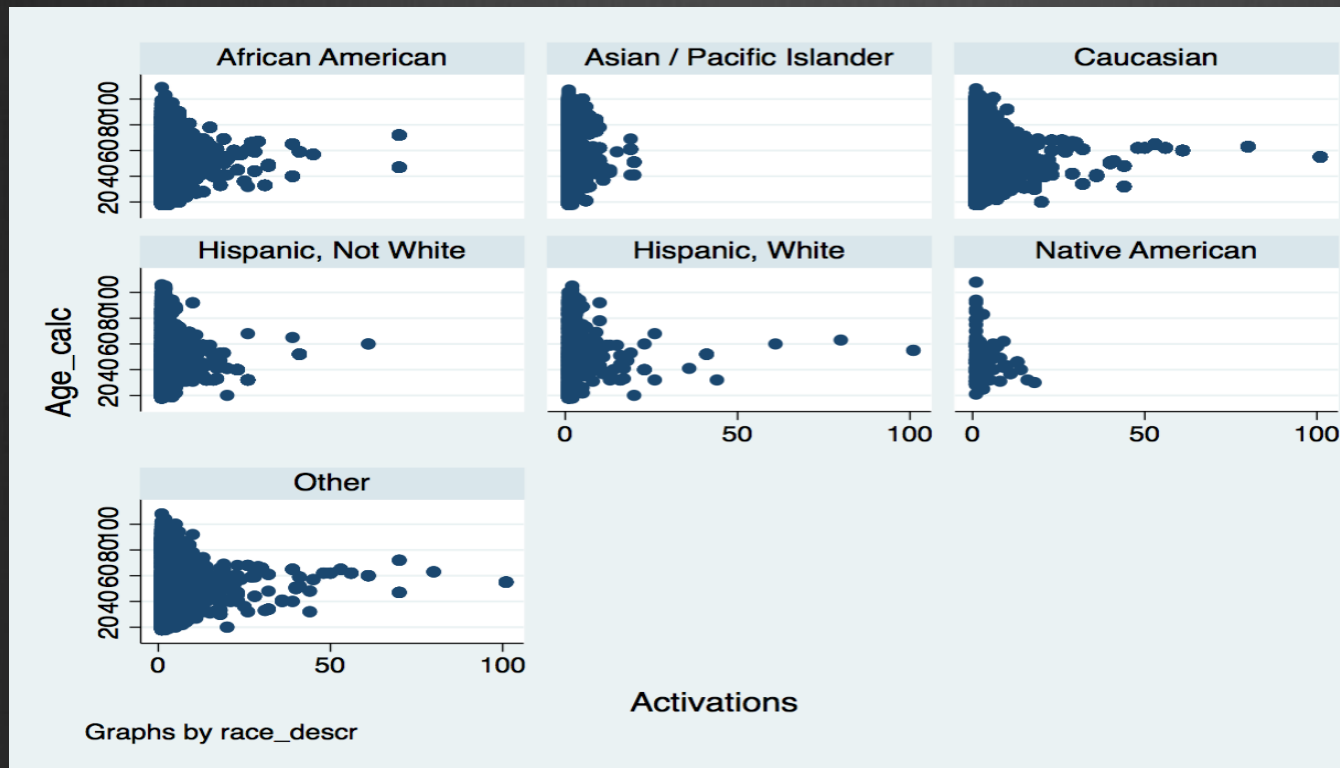
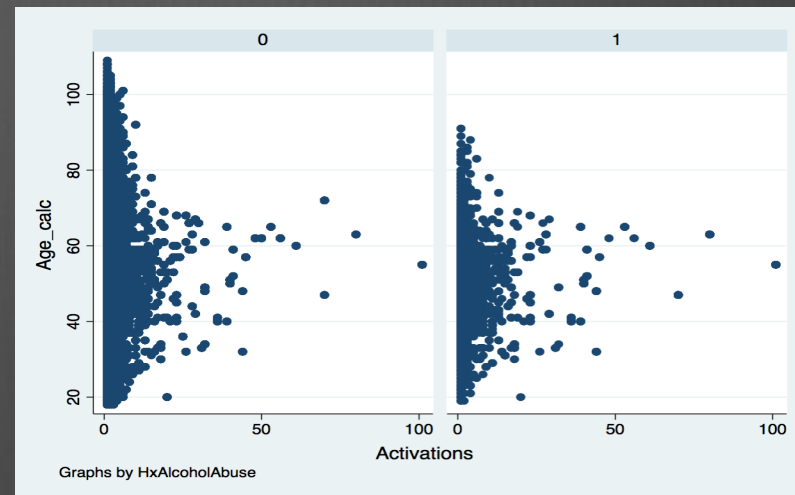
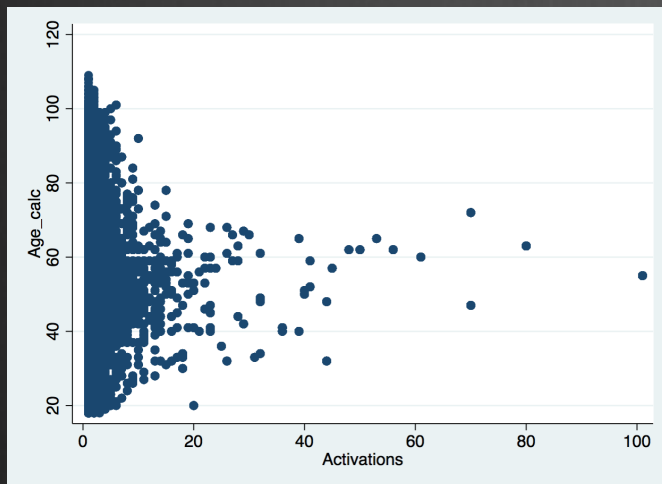
Better ways to tell the story? What factors do you think we should examine?

"EMS-STARS: Emergency Medical Services 'Superuser' Transport Associations, a Retrospective Study" M. Kennedy Hall MD, Maria C. Raven MD MPH MSc FACEP*, Jane Yu, PhD**, Clement Yeh MD*, Elaine Allen PhD**, Robert M. Rodriguez MD FAAEM*, Niels L Tangherlini NREMTA BA+, Karl A. Sporer MD***, John F. Brown MD MPA FACEP* Annals of Emergency Medicine, 2013





Stata Graphics – successive panels of scatter plots



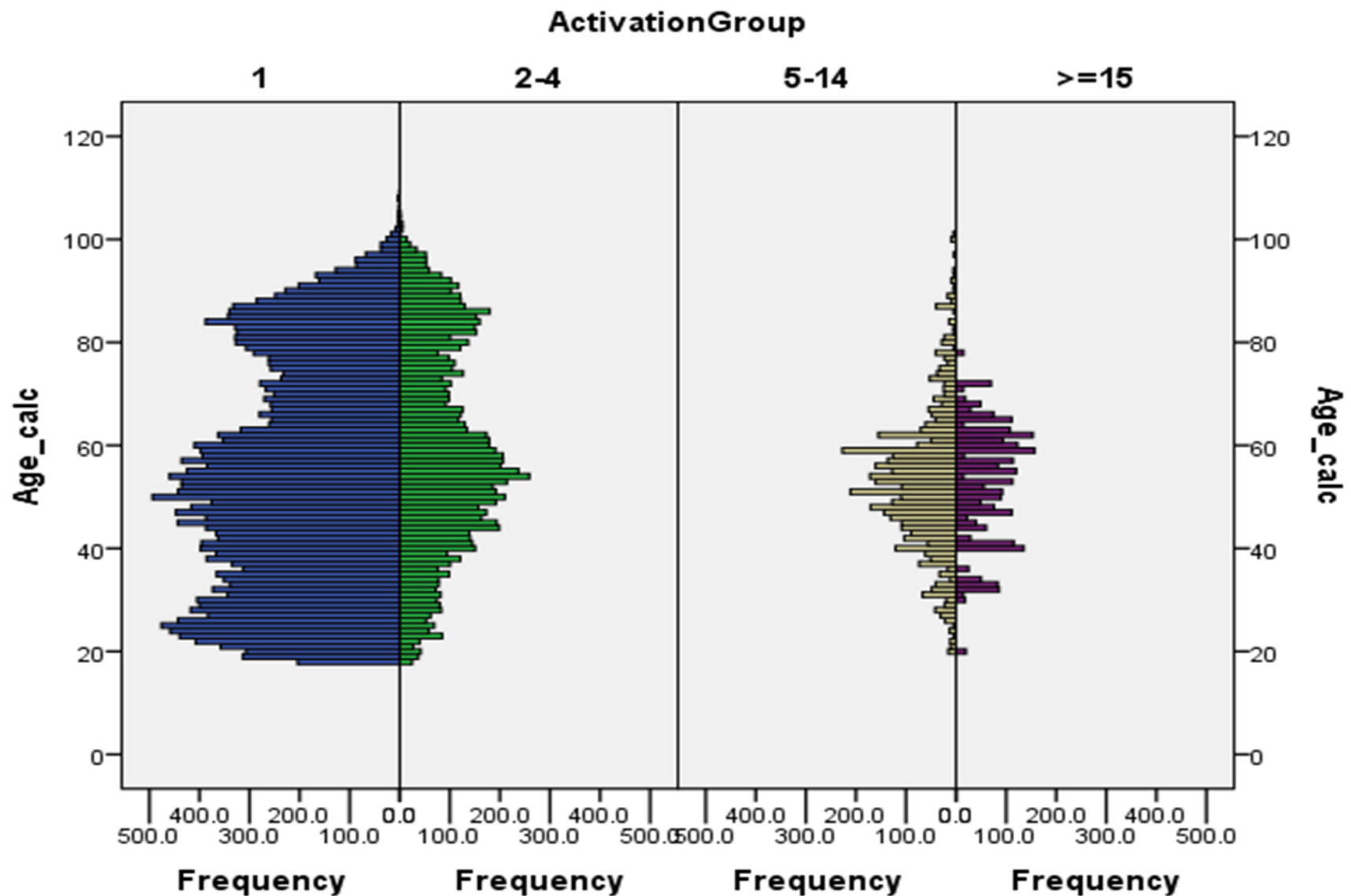
Age_calc



Activations

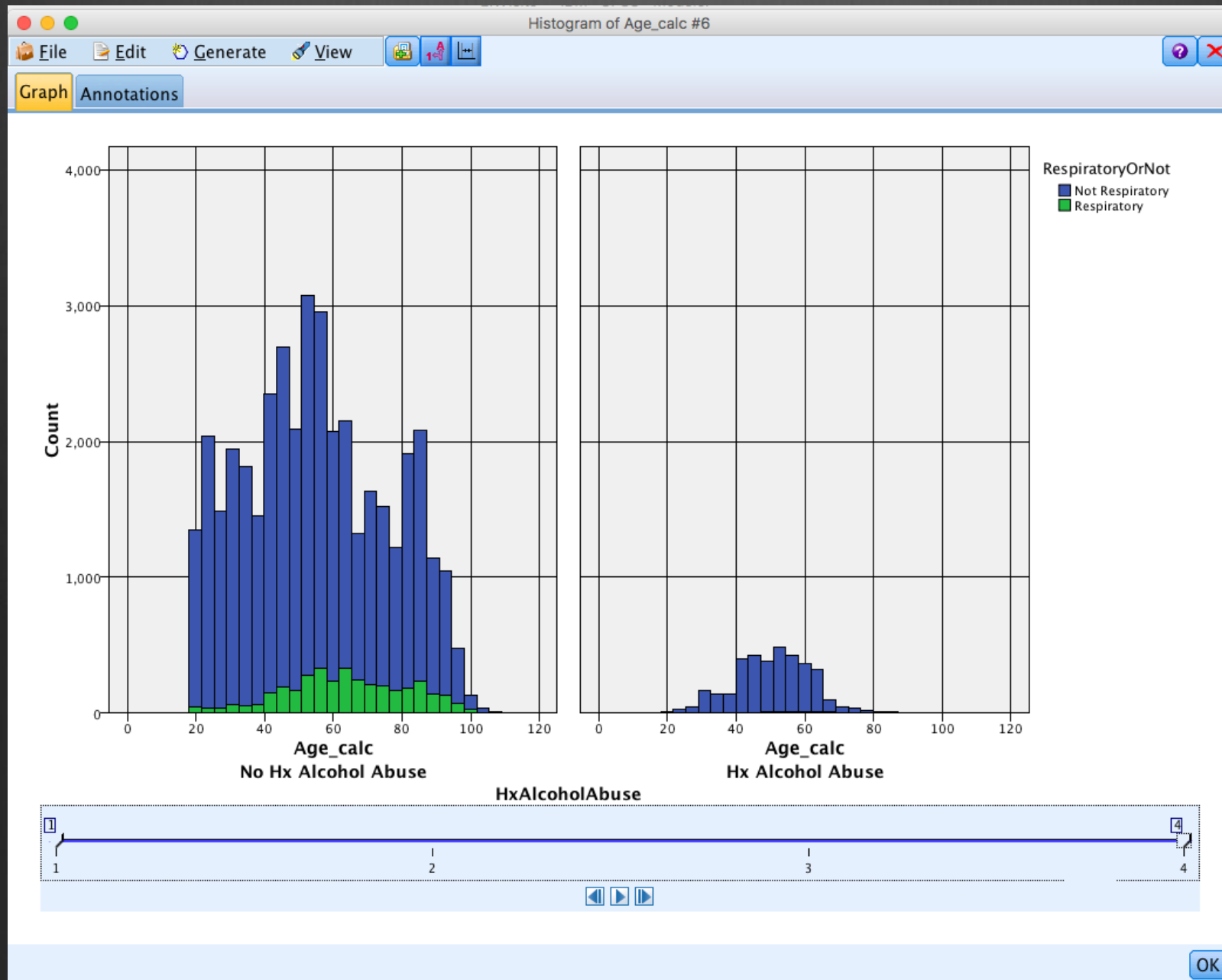
Graphs by MedicalHx

Examining the patterns of age was important*

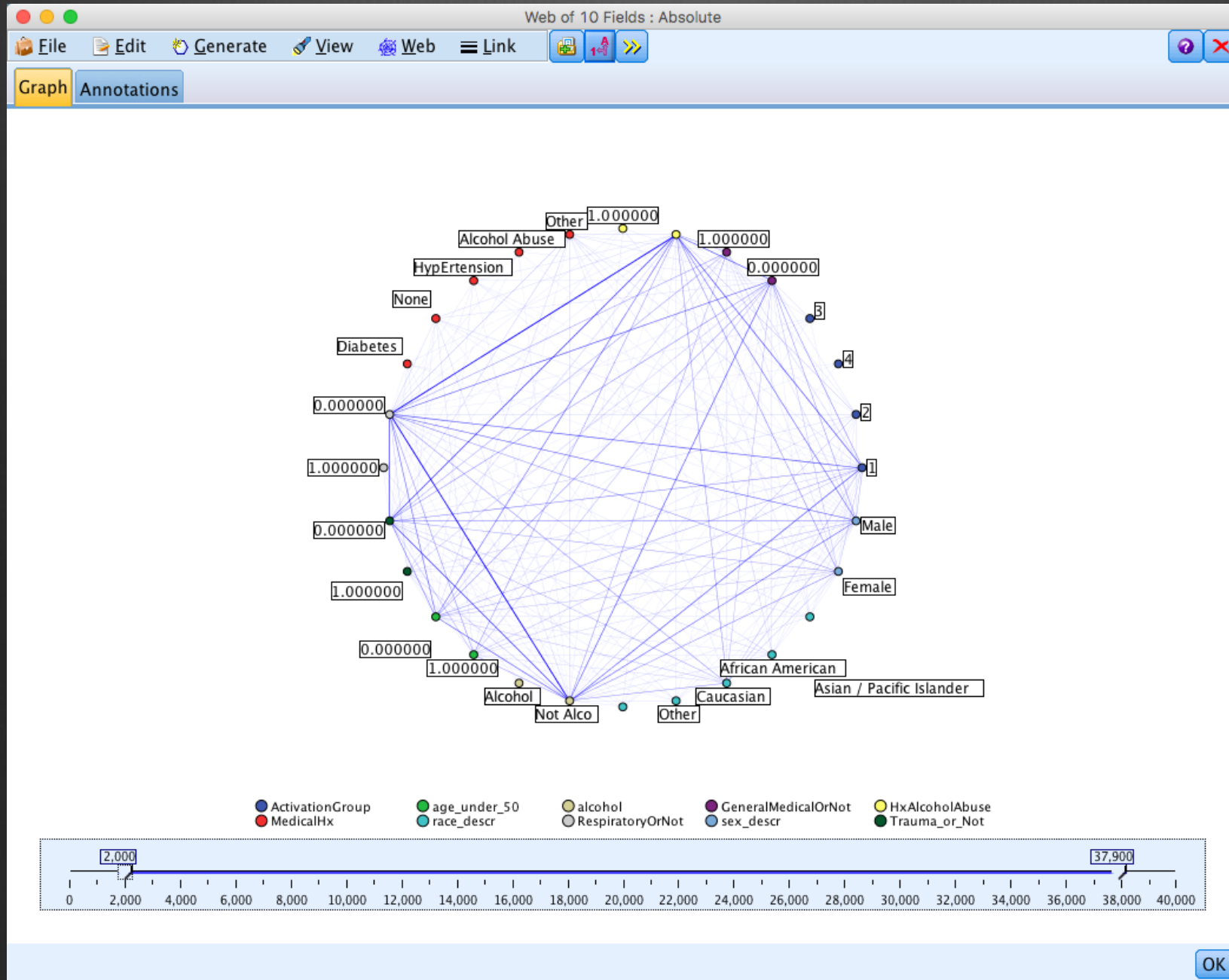


**SPSS Graphics*

Animation Plot from Modeler – let's look at more detail...



Web Plot from Modeler – let's look at more detail...



Animating your figures: Kohonen Analysis

Osteoarthritis data

age	race	spurious1	spurious2	spurious3	BMI	hispanic	hx_knee_inj	freq_knee_pain	KellgrenL	lateral_JS	medial_JS	min_medial_JSW	female	prog_cohort	white	Side	FT_cart_loss
61	Black	0.09	0.56	0.13	36.50	Non-Hispanic	No	Both	Extreme	0	3.0	1.6	Male	1	Non-White	Left	Yes
73	White	0.90	0.21	0.53	25.60	Non-Hispanic	No	Both	Extreme	0	3.0	0.6	Male	0	White	Left	Yes
66	White	0.91	0.18	0.97	30.70	Non-Hispanic	No	Both	Extreme	0	3.0	1.1	Female	0	White	Left	Yes
59	White	0.73	0.95	0.40	26.80	Non-Hispanic	No	Both	Extreme	0	3.0	0.0	Female	1	White	Right	Yes
57	White	0.11	0.24	0.25	26.30	Non-Hispanic	No	Both	Extreme	0	3.0	0.7	Male	0	White	Left	Yes
77	White	0.94	0.29	0.30	30.10	Non-Hispanic	No	Both	Extreme	0	3.0	0.5	Female	1	White	Right	Yes
53	White	0.28	0.13	0.32	34.50	Non-Hispanic	No	Both	Extreme	0	3.0	0.0	Male	1	White	Right	Yes
79	White	0.89	0.24	0.32	23.60	Non-Hispanic	No	Both	Extreme	0	3.0	0.0	Female	0	White	Right	Yes
58	White	0.82	0.92	0.61	27.00	Non-Hispanic	No	Both	Extreme	0	3.0	1.6	Male	1	White	Left	Yes
58	Black	0.02	0.81	0.50	28.80	Non-Hispanic	No	Both	Extreme	0	3.0	0.0	Female	1	Non-White	Right	Yes
60	White	0.78	0.85	0.96	27.90	Non-Hispanic	No	Both	Extreme	0	3.0	1.0	Female	1	White	Left	Yes

Going to use SPSS Modeler to cluster the data using a Kohonen network: Self Organizing Map

Created by Teuvo Kohonen ~ 20 years ago

Great for visualisation of high dimensional data

The clustering is an extension of a neural net

Consists of neurons arranged in a grid

Unsupervised analyses, usually 2 or 3 dimensions

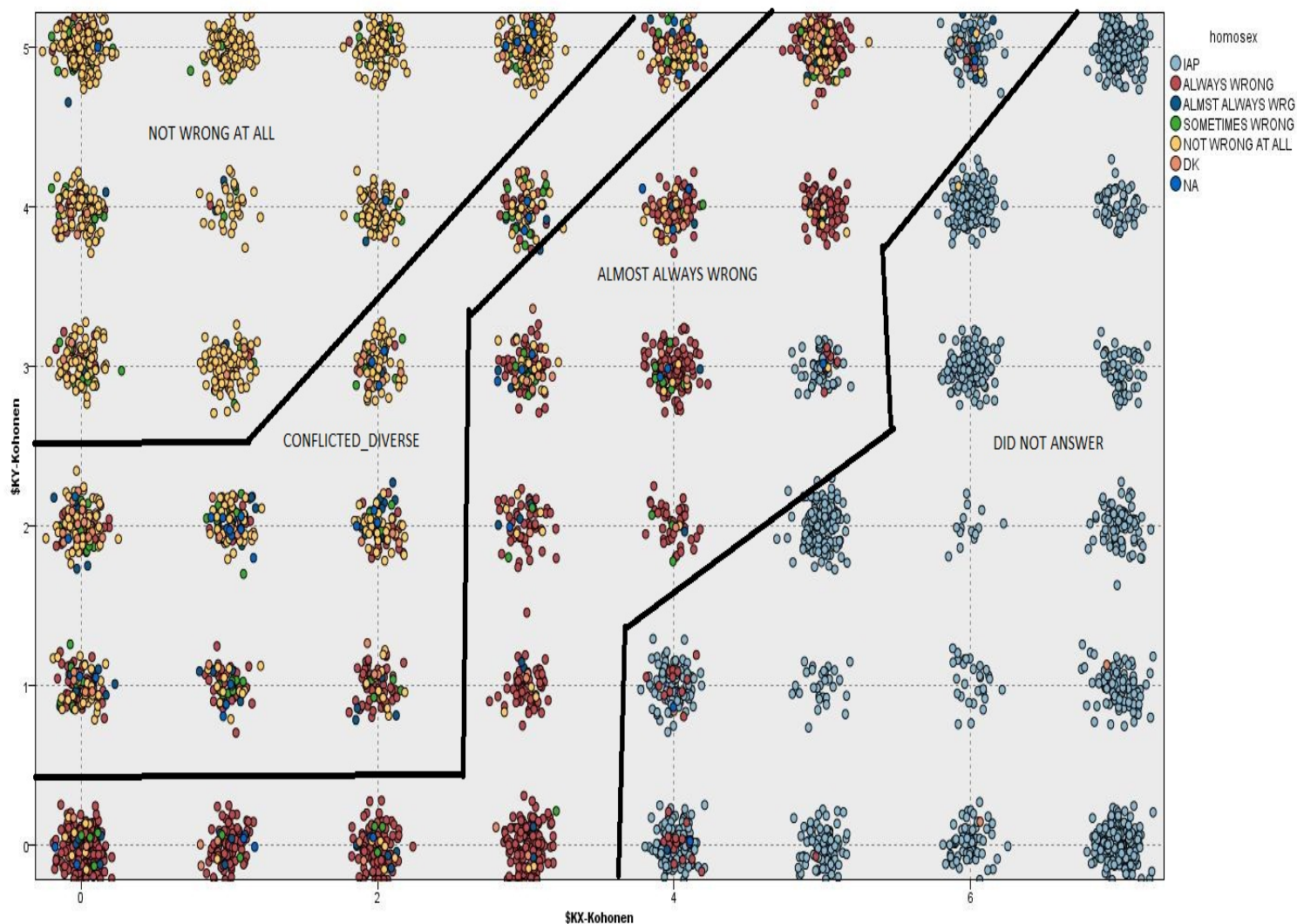
What's a good example of a self organizing map?

<http://joneisen.me/d3-tsp-demo/>

Animating your figures: Kohonen Analysis

- Opening the Modeler stream
- Examining the data
- Simple X / Y plot
- Web plot
- Kohonen analysis
- Kohonen Plots

Kohonen classifying respondents on the General Social Survey on Attitudes towards Homosexuality



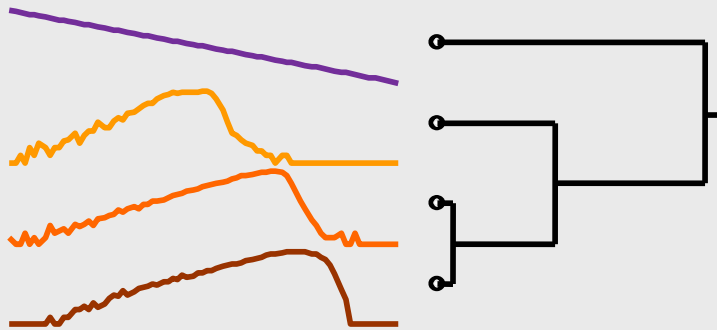
60 variables used for clustering including demographics & social & behavioral activities & habits –

The Kohonen network identified 48 groups and these groups are plotted using the attitude toward homosexuality variable.

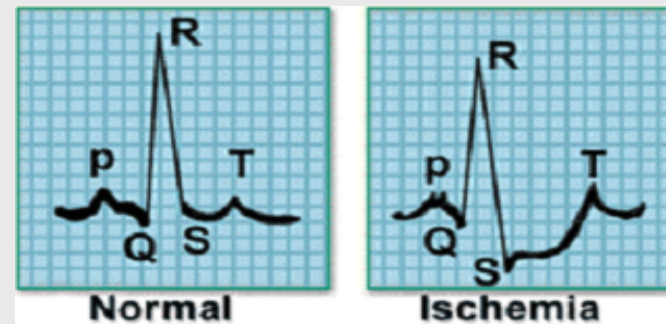
The General Social Survey (GSS) is a sociological survey used to collect information and keep a historical record of the concerns, experiences, attitudes, and practices of residents of the United States. Started in 1972 with annual surveys.

Time Series Data: What do we want to do with it?

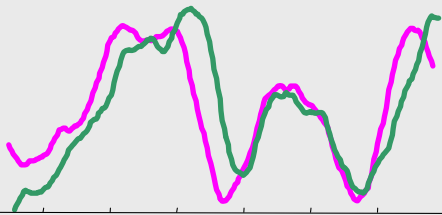
Cluster



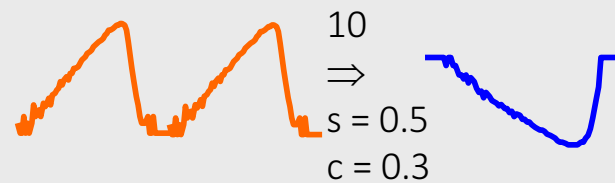
Classification



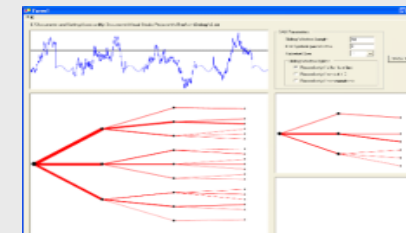
Motif Discovery



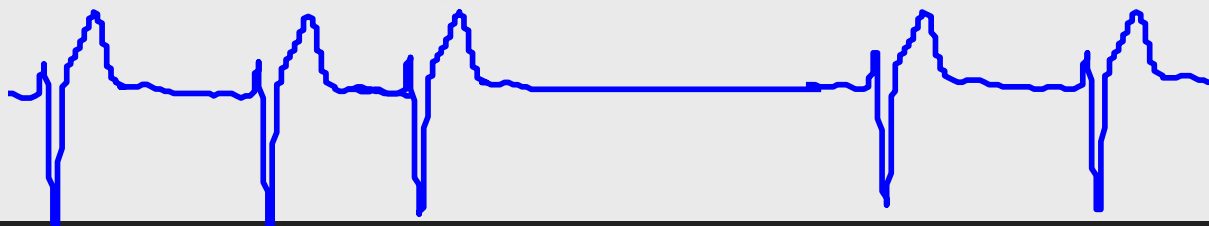
Rule Discovery



Visualization



Novelty Detection



Why is analyzing time series data hard?

1. Time series data are large databases
2. All of these techniques require **similarity** matching
3. Most data require a degree of subjectivity
4. Data formats differ, sampling rates differ, there is often noise
5. Need algorithms that can search for patterns quickly

...How do we define **similar**???

What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features*.



Similarity is hard to define, but...
“We know it when we see it”

The real meaning of similarity is a philosophical question.

We need a more objective approach.

**Webster's Dictionary*

Preventing Birth-Related Injuries:

Maternal Antecedents of Encephalopathy in Singleton Term Infants (MAESTRA)

To study the relationship of neonatal encephalopathy to overly frequent uterine contractions and fetal heart rate abnormalities within a large birth cohort. The cohort will be delivered by Kaiser Permanente Northern California OB physicians.

Tracings of the last 4 hours before birth to identify artifacts in the

- Fetal heart rate
- Maternal heart rate
- Uterine contractions
- Match to birth and maternal health records

UCSF_sample_Artifacts		
Field Name	Data Type	Description
SiteId	Number	Site identifier
PatientId	Number	Unique record identifier
ArtifactId	Number	An artifact is an efm feature . Each artifact has a unique identifier
Category	Number	Category refers to the type of artifact 1=baseline,2= acceleration 3=deceleration 4=contraction
StartTime	Date/Time	Refers to the time at the beginning of the artifact
Duration	Number	duration of the artifact in seconds
Y1	Number	baseline value at the beginning
Y2	Number	baseline value at the end
BaselineVariability	Number	variability in bpm
SecondsToPeak	Number	number of seconds between onset and the peak value
PeakValue	Number	highest FHR point in an acceleration or lowest point in a deceleration
Confidence	Number	a number from 0 to 100 representing the degree of matching between this acceleration or deceleration to its label ie late
Repair	Number	How many point in this feature are missing or artifactual. A second has 4 points
Height	Number	The difference between the peak value and the baseline preceding it
IsNonInterpretable	Yes/No	yes if the amount of repair exceeds a predetermined threshold, the threshold varies depending upon the length of the feature and the continuity of the ends
DecelerationCategory	Number	A number indicating the type of deceleration 1=Early, 2=Variable,4=Late 5=non associated
ContractionStart	Date/Time	A feature of decelerations giving the start time of an associated contraction, It will be blank if there is no contraction associated with the deceleration.
Area	Number	Area of decels and accels

TracingDuration.txt

SiteId	PatientId	ITCStartTime	StartTime	ITCEndTime	EndTime	SegmentCount	HrsOfTracing
1,1	2008-07-27	23:29:28	2008-07-27	16:29:28	2008-07-28	02:52:19	2008-07-27 19:52:19,1,3.380833
1,2	2010-02-12	09:56:21	2010-02-12	01:56:21	2010-02-12	16:49:54	2010-02-12 08:49:54,1,6.892500
1,3	2010-12-11	14:21:52	2010-12-11	06:21:52	2010-12-11	15:34:58	2010-12-11 07:34:58,1,1.218333
1,4	2008-03-31	00:34:51	2008-03-30	17:34:51	2008-03-31	11:42:29	2008-03-31 04:42:29,1,11.127222
1,5	2009-01-16	10:58:42	2009-01-16	02:58:42	2009-01-16	22:29:44	2009-01-16 14:29:44,1,11.517222
1,6	2010-03-01	00:27:11	2010-02-28	16:27:11	2010-03-01	14:26:58	2010-03-01 06:26:58,1,13.996388
1,7	2006-06-25	04:24:19	2006-06-24	21:24:19	2006-06-25	06:04:53	2006-06-24 23:04:53,1,1.676111
1,8	2009-04-24	07:21:14	2009-04-24	00:21:14	2009-04-24	09:28:38	2009-04-24 02:28:38,1,2.123333
1,9	2007-03-04	10:36:35	2007-03-04	02:36:35	2007-03-05	10:45:06	2007-03-05 02:45:06,1,24.141944
1,10	2011-01-08	05:10:21	2011-01-07	21:10:21	2011-01-08	15:24:44	2011-01-08 07:24:44,1,10.239722
1,11	2010-08-21	02:18:06	2010-08-20	19:18:06	2010-08-22	02:26:37	2010-08-21 19:26:37,1,24.141944
1,12	2006-08-25	22:00:41	2006-08-25	15:00:41	2006-08-25	23:13:06	2006-08-25 16:13:06,1,1.206944
1,13	2005-11-22	06:44:11	2005-11-21	22:44:11	2005-11-22	08:12:53	2005-11-22 00:12:53,1,1.478333
1,14	2010-12-23	12:43:51	2010-12-23	04:43:51	2010-12-23	14:16:21	2010-12-23 06:16:21,1,1.541666
1,15	2010-08-14	04:13:45	2010-08-13	21:13:45	2010-08-15	04:22:16	2010-08-14 21:22:16,1,24.141944
1,16	2006-11-30	20:08:46	2006-11-30	12:08:46	2006-11-30	20:54:28	2006-11-30 12:54:28,1,.761666
1,17	2006-07-16	10:30:30	2006-07-16	03:30:30	2006-07-16	19:25:48	2006-07-16 12:25:48,1,8.921666
1,18	2009-01-28	21:32:47	2009-01-28	13:32:47	2009-01-28	23:04:38	2009-01-28 15:04:38,1,1.530833
1,19	2008-01-17	21:40:16	2008-01-17	13:40:16	2008-01-18	00:58:12	2008-01-17 16:58:12,1,3.298888
1,20	2006-07-03	12:37:51	2006-07-03	05:37:51	2006-07-03	13:32:49	2006-07-03 06:32:49,1,.916111
1,21	2009-11-25	21:30:30	2009-11-25	13:30:30	2009-11-25	23:15:37	2009-11-25 15:15:37,1,1.751944
1,22	2010-03-24	20:55:27	2010-03-24	13:55:27	2010-03-24	21:34:39	2010-03-24 14:34:39,1,.653333
1,23	2010-01-11	22:45:10	2010-01-11	14:45:10	2010-01-12	01:27:35	2010-01-11 17:27:35,1,2.706944
1,24	2010-12-01	01:27:05	2010-11-30	17:27:05	2010-12-01	22:05:59	2010-12-01 14:05:59,1,20.648333
1,25	2007-04-30	20:10:29	2007-04-30	13:10:29	2007-04-30	20:51:12	2007-04-30 13:51:12,1,.678611
1,26	2006-12-08	18:22:25	2006-12-08	10:22:25	2006-12-09	09:36:32	2006-12-09 01:36:32,1,15.235277
1,27	2009-09-13	10:51:56	2009-09-13	03:51:56	2009-09-13	16:18:46	2009-09-13 09:18:46,1,5.447222
1,28	2009-04-03	23:33:59	2009-04-03	16:33:59	2009-04-04	17:40:36	2009-04-04 10:40:36,1,18.110277
1,29	2008-10-21	14:54:06	2008-10-21	07:54:06	2008-10-21	19:48:41	2008-10-21 12:48:41,1,4.909722
1,30	2008-06-04	18:11:10	2008-06-04	11:11:10	2008-06-05	18:19:41	2008-06-05 11:19:41,1,24.141944
1,31	2006-12-28	01:19:57	2006-12-27	17:19:57	2006-12-28	02:15:31	2006-12-27 18:15:31,1,.926111
1,32	2005-11-02	13:53:16	2005-11-02	06:53:16	2005-11-02	14:34:26	2005-11-02 07:34:26,1,.686111
1,33	2007-12-29	08:07:50	2007-12-29	00:07:50	2007-12-30	02:39:54	2007-12-29 18:39:54,1,18.534444
1,34	2009-08-09	18:40:34	2009-08-09	11:40:34	2009-08-10	18:49:05	2009-08-10 11:49:05,1,24.141944
1,35	2009-09-28	13:07:46	2009-09-28	06:07:46	2009-09-28	14:00:24	2009-09-28 07:00:24,1,.877222
1,36	2009-08-19	23:33:36	2009-08-19	16:33:36	2009-08-20	23:42:07	2009-08-20 16:42:07,1,24.141944

SiteId	PatientId	ArtifactId	Category	StartTime	MinBeforeTracingEnd	epoch	changebytime	Duration	Y1	Y2	BaselineVariability	SecondsToPeak	PeakValue	Confidence	Repair	Height	IsNonInterpretable	DecelerationCategory	NonReassuringFeatures	ContractionStart	DecelOnsetDelay	Area
1	1	13	1	36:50.5	195.483333	195		74.5	156	141	22											0
1	1	14	1	38:40.2	193.05	193	1.833333	151	142	150	14											0
1	1	15	1	41:11.5	191.133333	191	2.516667	153.75	150	151	18											0
1	1	16	2	43:45.5	188.566666	188	2.566667	39.25				17.5	169	93	0	20	FALSE					1812
1	1	17	1	7/27/08 16:44	187.9	187	0.666666	67	147	154	12											0
1	1	18	1	45:32.2	186.783333	186	1.116667	260.75	148	150	8											0
1	1	19	1	49:53.3	182.433333	182	4.35	260.25	152	163	17											0
1	1	20	1	54:13.7	178.1	178	4.333333	93.5	160	160	16											0
1	1	21	2	55:47.5	176.533333	176	1.566667	83.5				66.5	176	60	0	18	FALSE					2934.5
1	1	22	1	57:11.2	175.133333	175	1.4	252	159	152	17											0
1	1	23	1	01:23.5	170.933333	170	4.2	111.75	157	149	16											0
1	1	1	4	7/27/08 17:03	169.183333	169	1.75	123				67										0
1	1	24	2	03:15.5	169.066666	169	0.116667	47				17.25	173	79	0	20	FALSE					1877.5
1	1	25	1	04:02.7	168.283333	168	0.783333	37	155	149	11											0
1	1	26	2	7/27/08 17:04	167.65	167	0.633333	33.5				8.5	171	89	0	15	FALSE					1147
1	1	27	1	05:13.7	167.1	167	0.55	73.75	153	152	18											0
1	1	28	2	7/27/08 17:06	165.65	165	1.45	50.25				13.75	170	91	0	16	FALSE					1951
1	1	29	1	07:30.5	164.816666	164	0.833334	243	150	169	15											0
1	1	2	4	7/27/08 17:14	158.133333	158	6.683333	52				28										0
1	1	3	4	7/27/08 17:25	146.683333	146	11.45	64				32										0
1	1	4	4	7/27/08 17:31	140.566666	140	6.116667	70				28										0
1	1	5	4	7/27/08 18:09	102.733333	102	37.833333	157				65										0
1	1	6	4	7/27/08 18:14	97.433333	97	5.3	112				67										0
1	1	7	4	7/27/08 18:20	92.266666	92	5.166667	124				64										0
1	1	8	4	7/27/08 18:23	89.266666	89	3	95				42										0
1	1	9	4	7/27/08 18:26	86.2	86	3.066666	108				63										0
1	1	10	4	7/27/08 18:28	83.833333	83	2.366667	83				42										0
1	1	11	4	7/27/08 18:31	80.9	80	2.923333	67				28										0
1	1	12	4	7/27/08 18:54	57.65	57	23.25	85				61										0
1	1	30	1	55:01.7	57.3	57	0.35	174.75	151	151	18											0
1	1	31	2	57:56.7	54.383333	54	2.916667	49.75				24.75	168	93	0	18	FALSE					2341
1	1	32	1	58:47.7	53.533333	53	0.85	225.25	154	153	18											0
1	1	33	1	02:33.2	49.766666	49	3.766667	220.5	161	152	18											0
1	1	34	2	7/27/08 19:06	46.083333	46	3.683333	71.25				12	170	93	6	18	FALSE					3801.214286
1	1	35	1	07:25.5	44.9	44	1.183333	158.5	151	150	8											0
1	1	36	1	14:27.5	37.866666	37	7.033334	447.25	151	152	6											0
1	1	37	3	21:56.2	30.383333	30	7.483333	42				24.5	100	98	46	53	TRUE	2	0			5142.756098
1	1	38	1	22:38.5	29.683333	29	0.7	100.5	151	147	9											0

	PatientId	ArtifactId	MinBeforeT~d	Duration	SecondsToP~k
1	2	148	33.783333	68	30
2	2	149	30.916666	90	53
3	2	150	29.083333	55	31
4	2	151	27.116666	85	41
5	2	152	25.5	83	47
6	2	153	23.833333	82	41
7	2	154	21.216666	50	28
8	2	155	17.633333	69	30
9	2	156	15.033333	81	34
10	2	157	10.683333	105	62
11	3	399	35.166666	69	36
12	3	400	31.733333	83	40
13	3	401	27.583333	86	47
14	3	402	24.25	89	45
15	3	403	22.533333	82	42
16	3	404	19.883333	75	36

Interpretable Decelerations

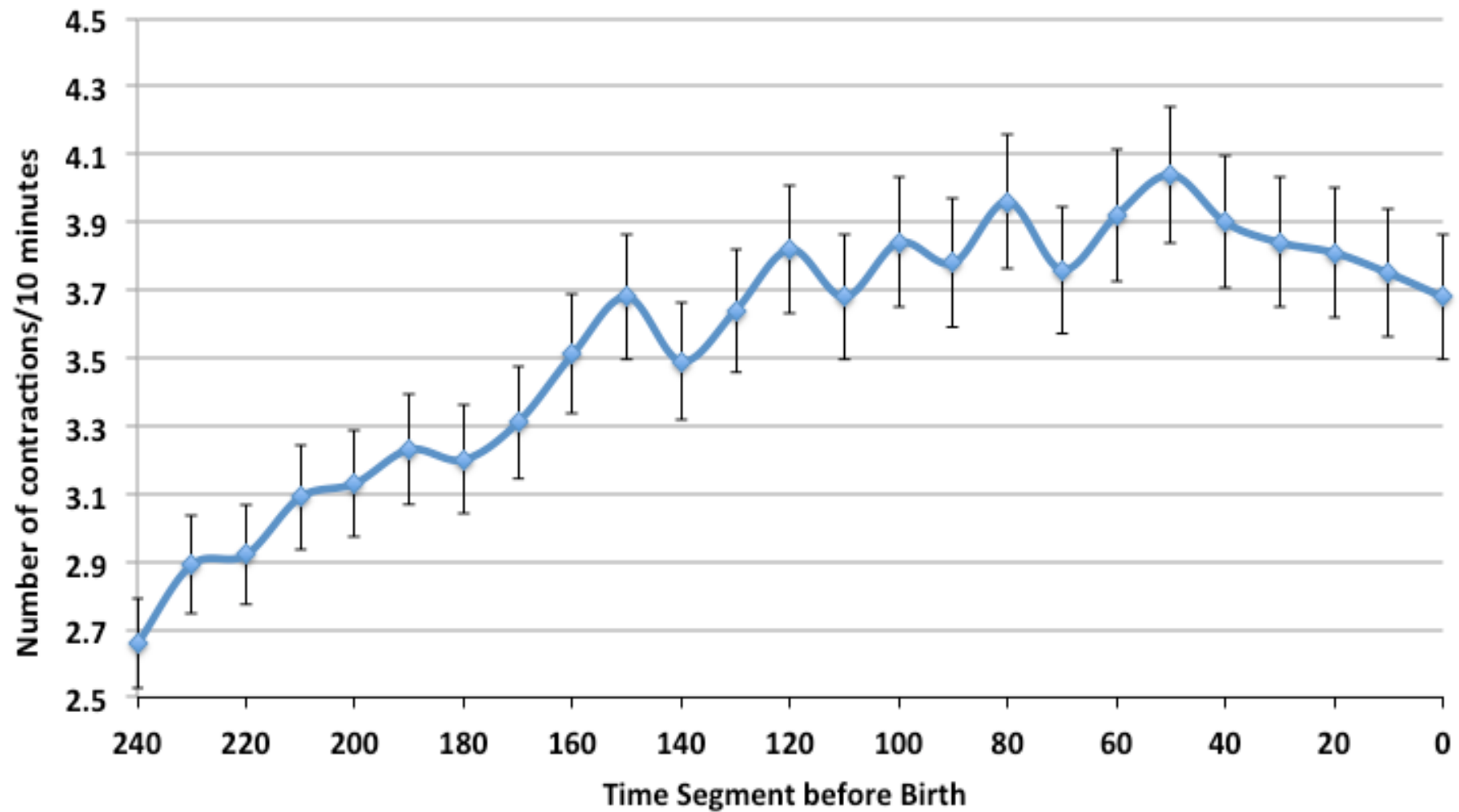
Deceleration						
Category	Freq.	Percent	Mean Area	Std. Dev.	Min	Max
Early	429	9.33	2174.637	1388.192	419	9062.5
Late	891	19.38	2650.345	1731.03	142	13053
Non-Associated	395	8.59	3171.894	2008.815	179.5	11624
Variable	2,883	62.7	4133.22	4348.57	10.5	77327.4
Decelerations	4,605					

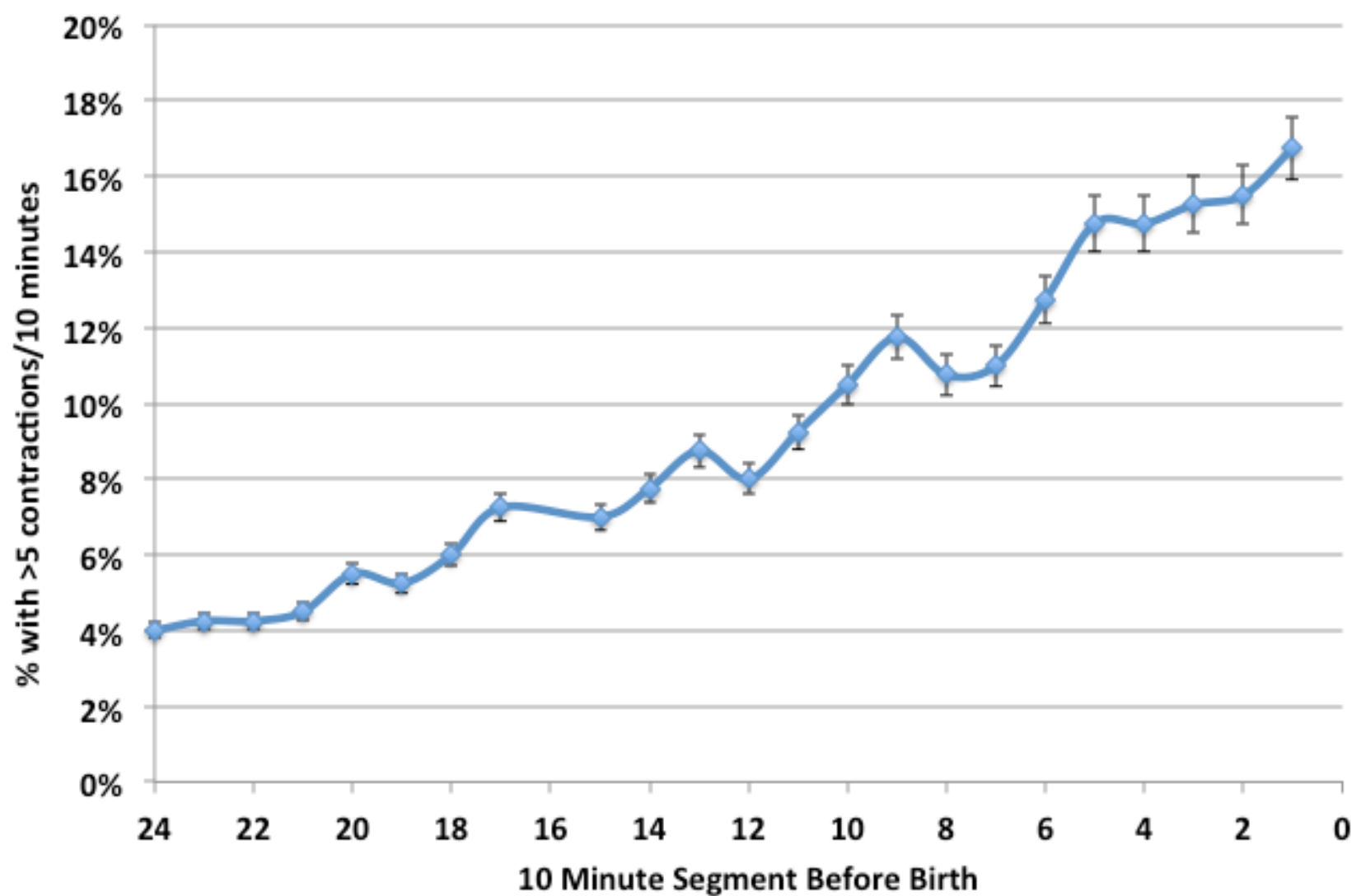
Late Decelerations	
>10 Decelerations	34
1-9 Decelerations	68
None	50
Total Patients	152

All Decelerations

Deceleration						
Category	Freq.	Percent	Mean Area	Std. Dev.	Min	Max
Early	517	7.38	2267.22	1428.58	231.42	9062.50
Late	1,124	16.04	2850.94	1932.11	-1001.50	14303.00
Non-Associated	582	8.31	3202.28	2106.41	179.50	12152.77
Variable	4,784	68.27	5467.93	6736.75	-2953.50	158273.80
Total	7,007	100				

Late Decelerations	
>20 Decelerations	17
10-19 Decelerations	24
1-9 Decelerations	78
None	52
Total Patients	171





Other types of visualizations? When to use a graphic?

You should already know:

Scatter plots

Bar charts

Histograms

Box plots

Pie charts

Classification trees

Advice:

If the graphic makes it easier to understand, use it

Don't use a table and a graphic showing the same thing