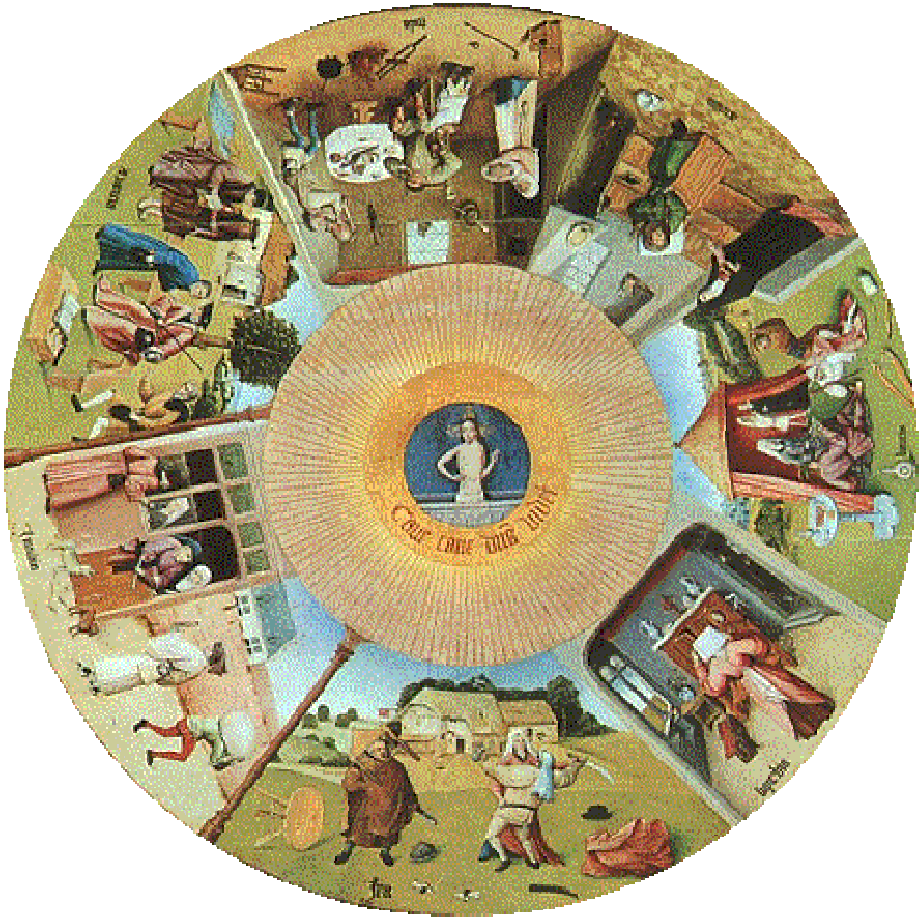
Seven Deadly Sins of Data Mining



Richard De Veaux

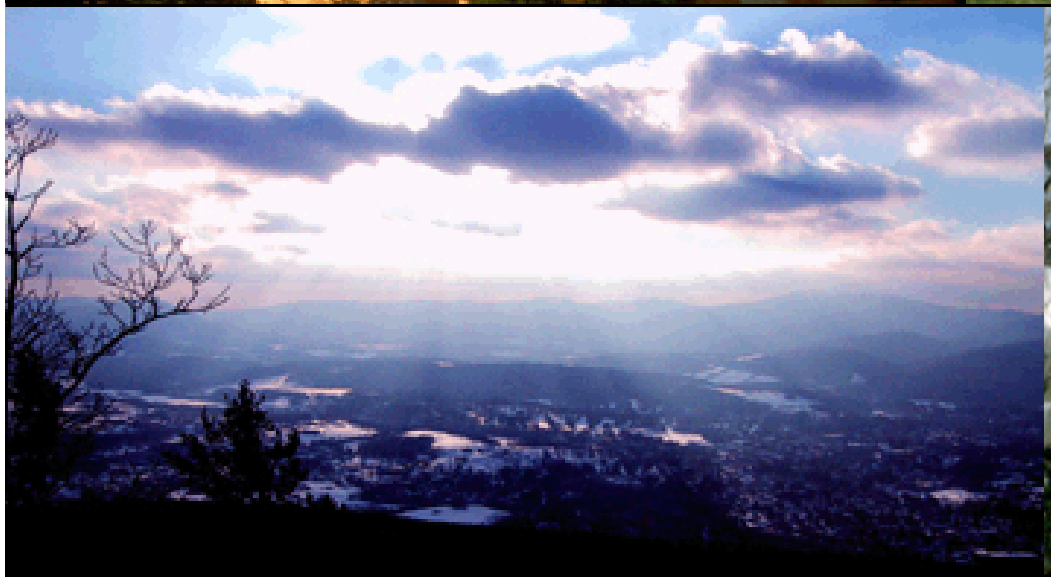
Williams College

Williams College

Where is Massachusetts?

Where in Massachusetts?

Williams College



Polls

Special Report

America's Best Colleges

Edited by David M. Ewalt, 08.11.10, 06:00 PM EDT

Forbes' list of public and private colleges and universities ranks the best schools--from the students' point of view.

The best college in America isn't in Cambridge or Princeton, West Point or Annapolis. It's nestled in the Berkshire Mountains. Williams College, a 217-year-old private liberal arts school, tops our third annual ranking of America's Best Colleges. Our list of more than 600 undergraduate institutions is based on the quality of the education they provide, the experiences of the students and how much they achieve. Read More »

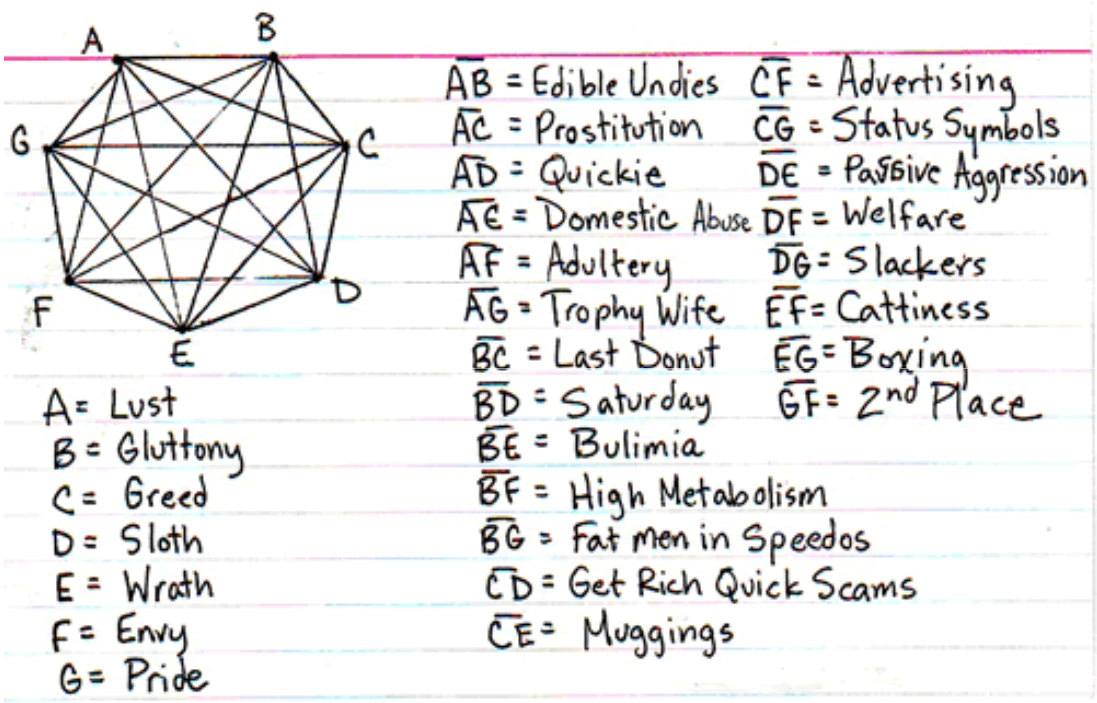
Forbes America's Best Colleges 2010

Download This Logo

1. Williams College
2. Princeton University
3. Amherst College
4. United States Military Academy
5. Massachusetts Institute of Technology
6. Stanford University
7. Swarthmore College
8. Harvard University
9. Claremont McKenna College
10. Yale University
11. United States Air Force Academy
12. Wellesley College
13. Columbia University

The Original Seven Sins

- Lust
- Gluttony
- Greed
- Sloth
- Wrath
- Envy
- Pride



Thanks to Ewan's Corner, Source Unknown

What is Data Mining?

- Lots of definitions
  - http://en.wikipedia.org/wiki/Data_mining
- Finding interesting structure in data
  - *Structure:* refers to statistical patterns, predictive models, hidden relationships
  - KDD

- Examples of tasks addressed by Data Mining
  - Predictive Modeling (classification, regression)
  - Segmentation (Data Clustering)
  - Summarization
  - Visualization

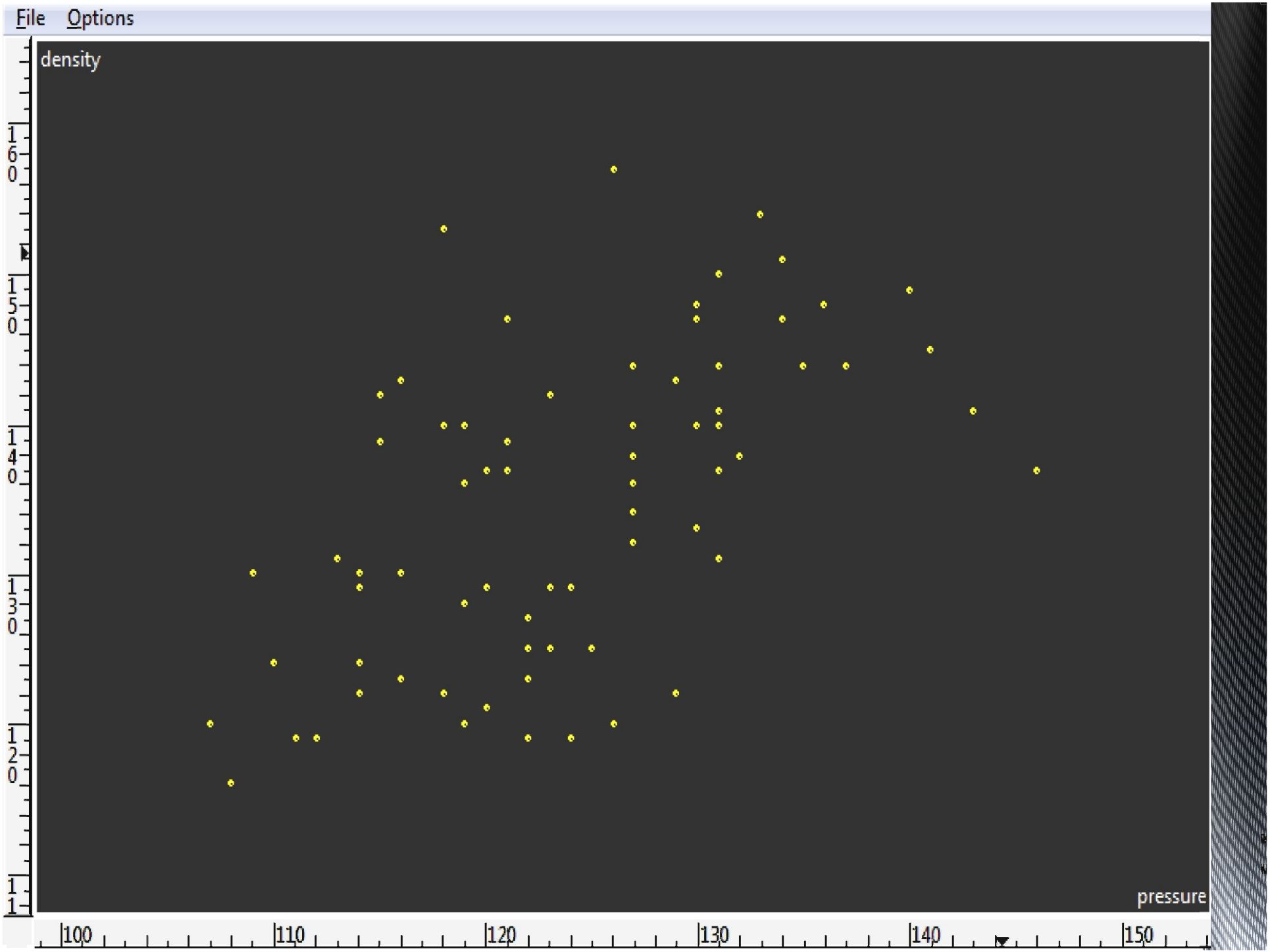
What isn't Data Mining?

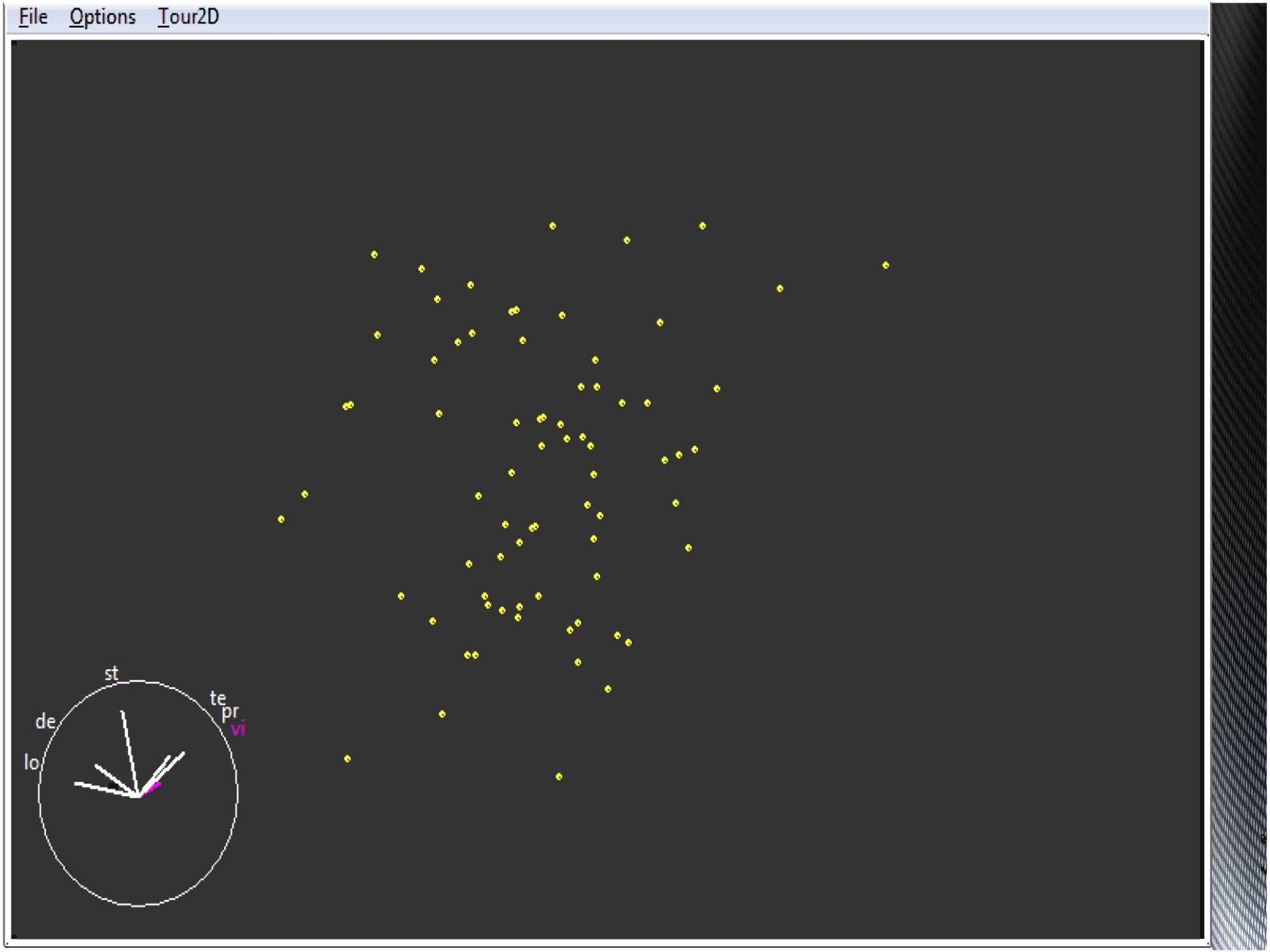
- A magic wand
  - Answering questions your haven't asked
  - Automatically reporting interesting findings
  - Replacing thinking about your business
  - Replacing smart data analysis
  - Enabling bad data to provide useful information
- So, what are the Seven Deadly Sins of DM?

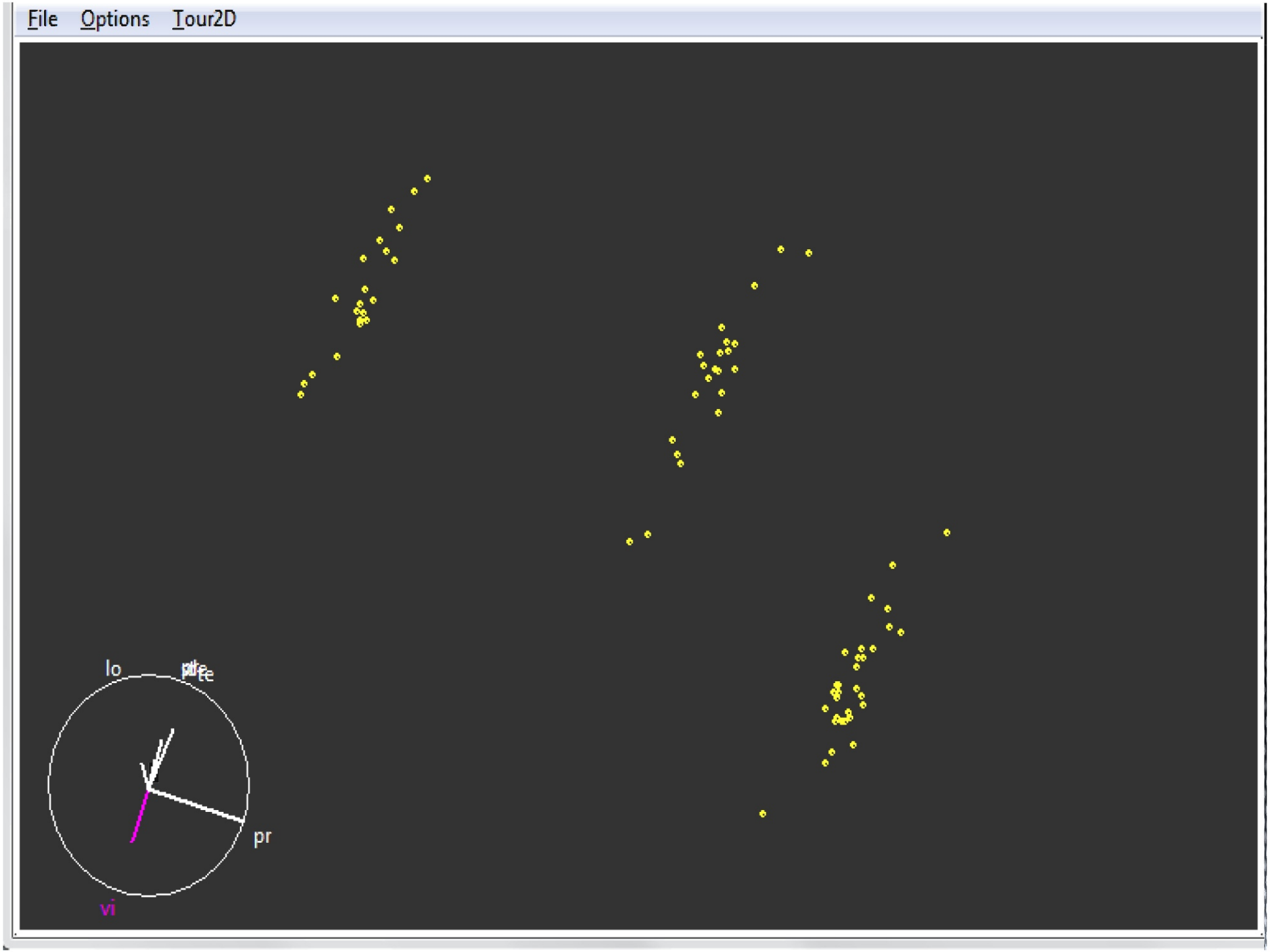
Failing to Define the Problem

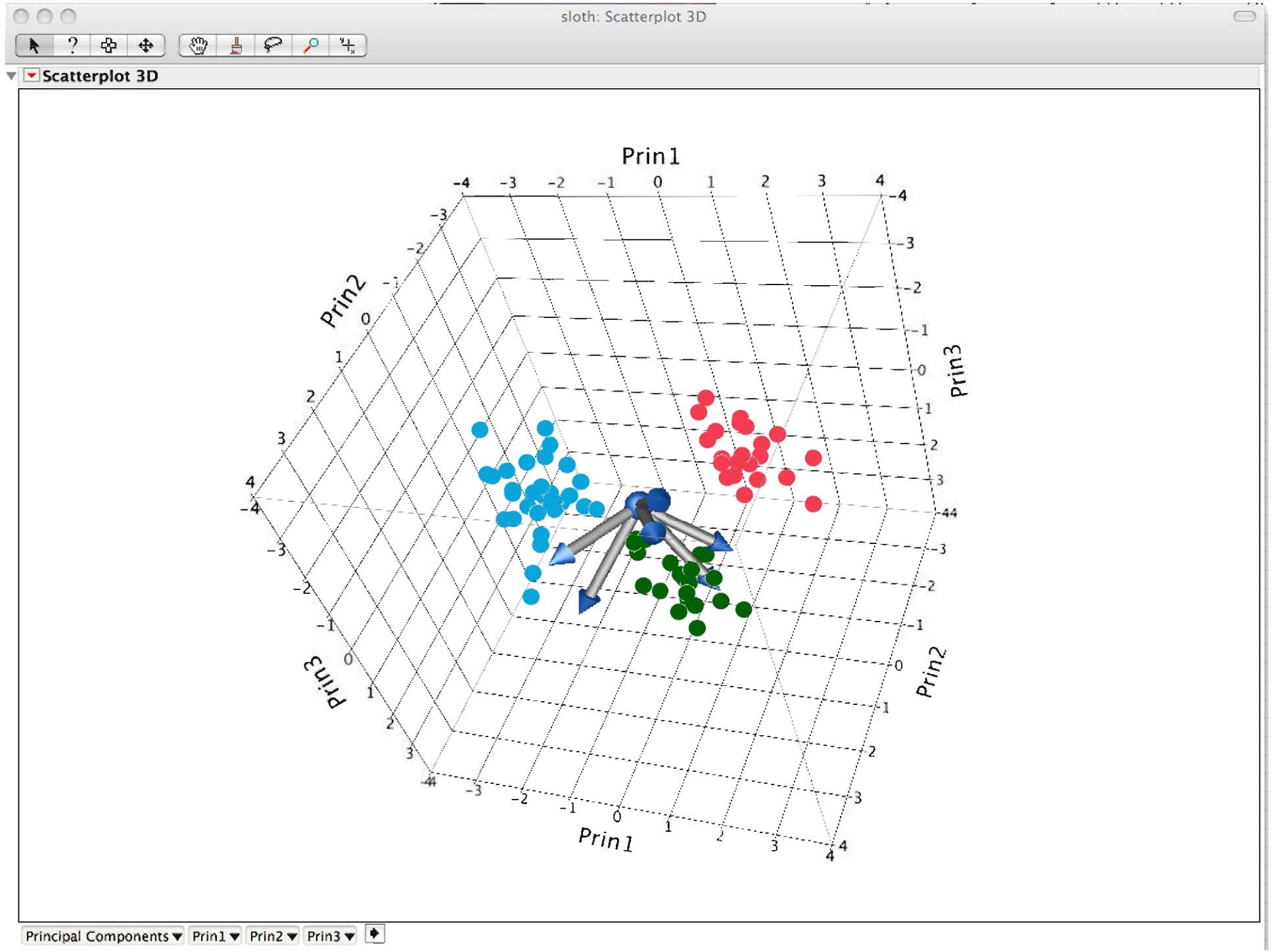
Production Analysis

- Statistician studies 6 properties for 74 recent samples from production
  - Viscosity
  - Max temperature to failure
  - Max pressure to failure
  - Density
  - Load at failure
  - Stress at failure


File
Options
density
160
150
140
130
120
110
100
pressure


File
Options
Tour2D
st
te
pr
vi
de
lo


File
Options
Tour2D
lo
pe
pr
vi


sloth: Scatterplot 3D
Scatterplot 3D
Prin1
Prin2
Prin3
Principal Components
Prin1
Prin2
Prin3

Three Groups!!!

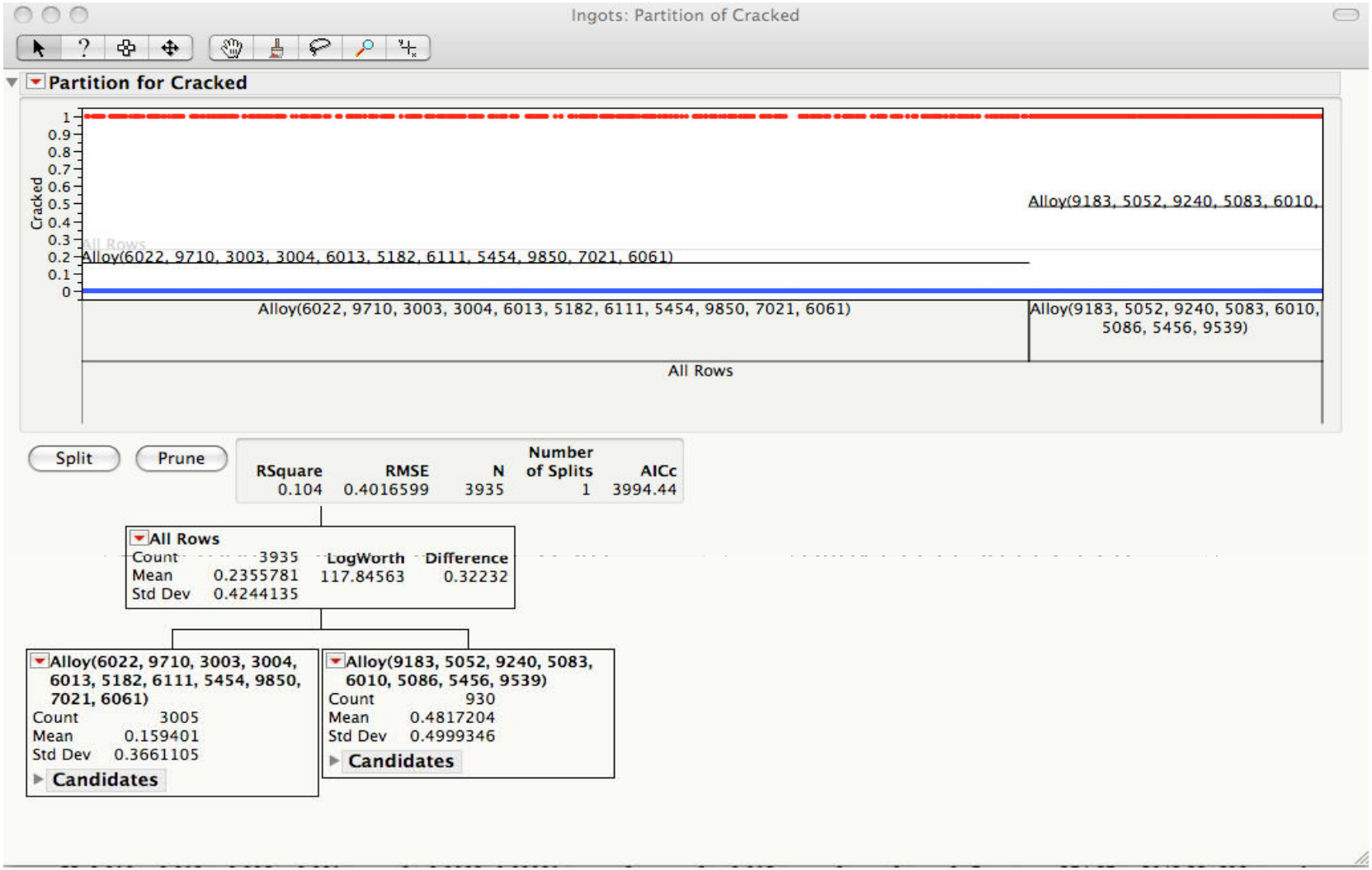
“Do you know we make 3 products?”

Not *Fully* Understanding the Problem

Ingot cracking

- 3935 30,000 lb. Ingots
- Up to 25% cracking rate
- $30,000 per recast
- 90 potential explanatory variables
  - Water composition (reduced)
  - Metal composition
  - Process variables
  - Other environmental variables

ALCOA

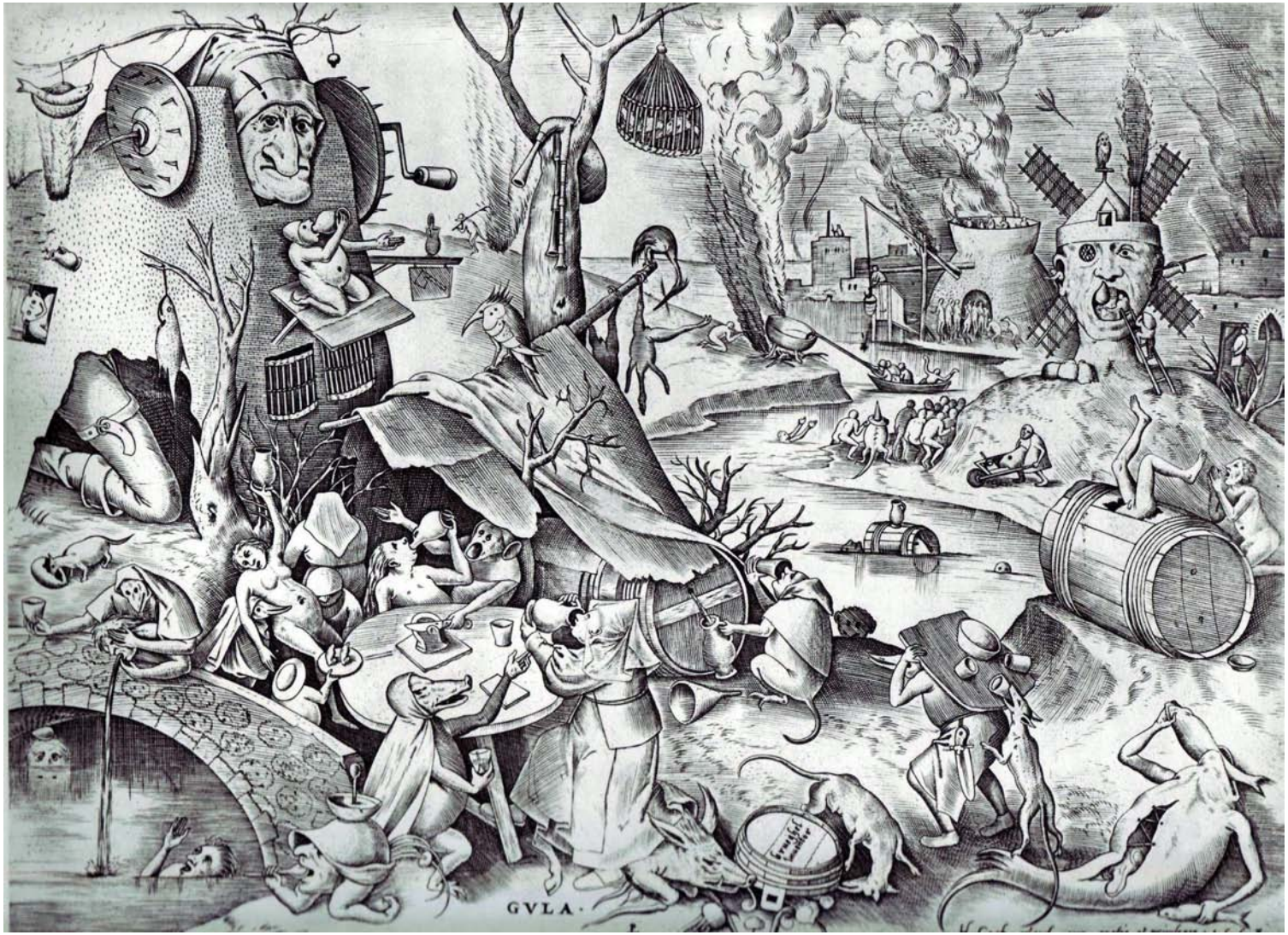


"We know that – some alloys are hard to make.

That's why we gave you the data in the first place."

ITI Cavtat

Underestimating Data Preparation



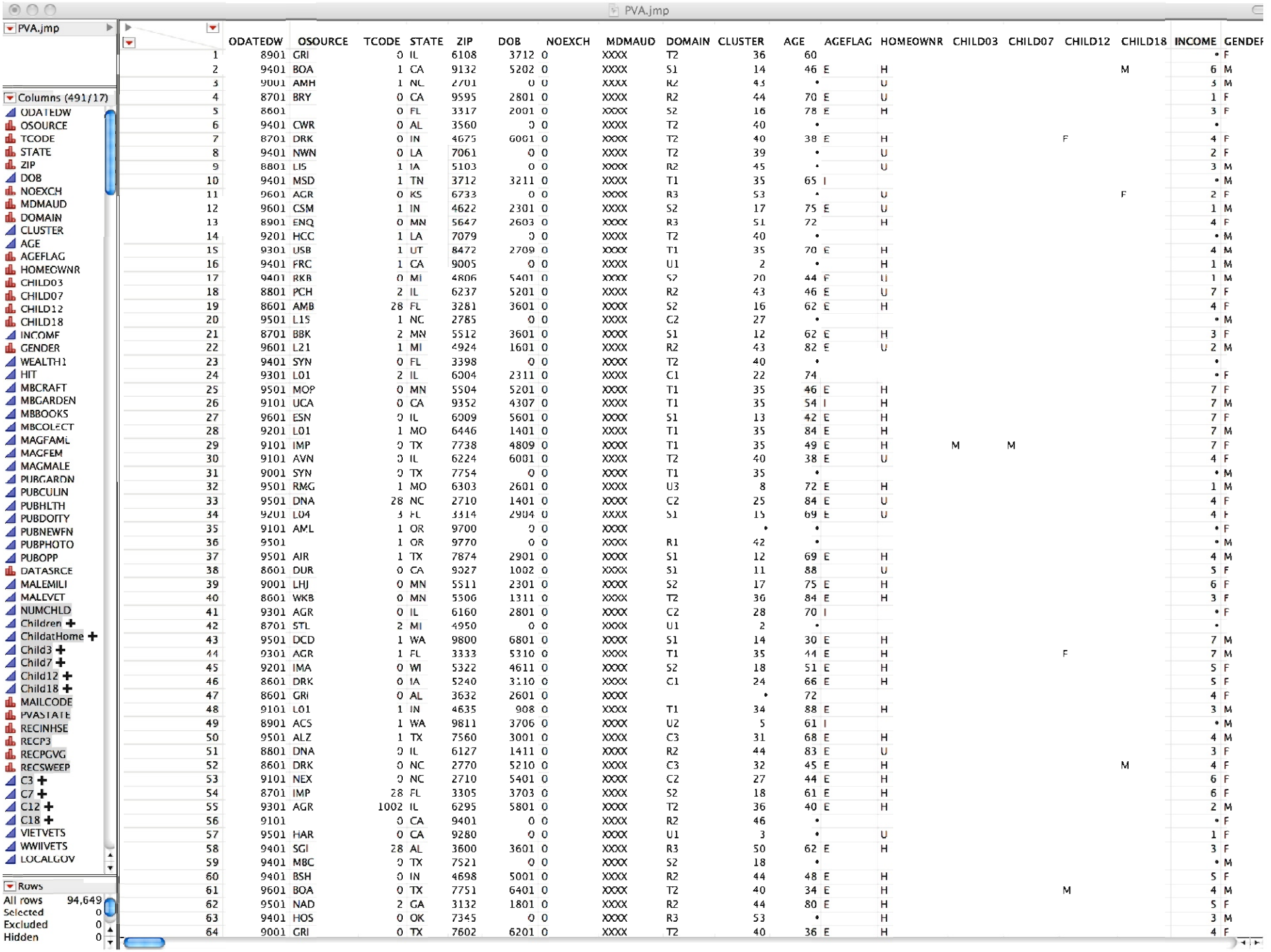
Data Preparation

60% to 95% of the time is spent preparing the data

87.1% of all statistics are made up

Paralyzed Veterans of America

- PVA is a philanthropic organization,
  - -Sanctioned by the US Govt to represent the disabled veterans
- They send out 4 million “free gifts” , every 6 weeks
  - -And hope for donations
- Data were used for the KDD 1998 cup
  - -200,000 donors
    - -(100,000 training, 100,000 test)
  - -481 demographic variables
    - -Past giving, income, age etc etc etc
  - -Recent campaign (only for training set)
    - -Did they give? (Target B)
    - -How much did they give (Target D)
- To optimize profit, who should receive the current solicitation?
- What is the most cost effective strategy?

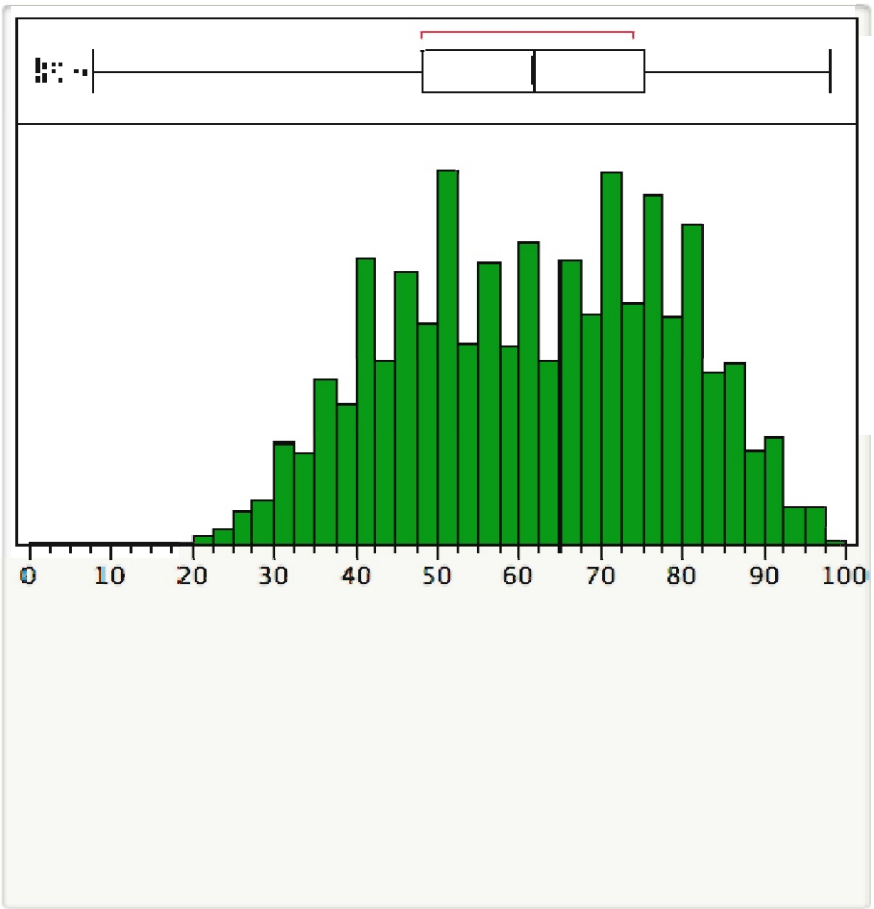


	ODATEDW	OSOURCE	TCODE	STATE	ZIP	DOB	NOEXCH	MDMAUD	DOMAIN	CLUSTER	AGE	AGEFLAG	HOMEOWNR	CHILD03	CHILD07	CHILD12	CHILD18	INCOME	GENDER
1	8901	GRI	0	IL	6108	3712	0	XXXX	T2	36	60							•	F
2	9401	BOA	1	CA	9132	5202	0	XXXX	S1	14	46	E	H				M	6	M
3	9001	AMH	1	NC	2701	0	0	XXXX	R2	45	•		U					3	M
4	8701	BRY	0	CA	9595	2801	0	XXXX	R2	44	70	E	U					1	F
5	8601		0	FL	3317	2001	0	XXXX	S2	16	78	E	H					3	F
6	9401	CWR	0	AL	3560	0	0	XXXX	T2	40	•							•	
7	8701	DRK	0	IN	4675	6001	0	XXXX	T2	40	38	E	H			F		4	F
8	9401	NWN	0	LA	7061	0	0	XXXX	T2	39	•		U					2	F
9	8801	LIS	1	IA	5103	0	0	XXXX	R2	45	•		U					3	M
10	9401	MSD	1	TN	3712	3211	0	XXXX	T1	35	65	I						•	M
11	9601	AGR	0	KS	6733	0	0	XXXX	R3	53	•		U				F	2	F
12	9601	CSM	1	IN	4622	2301	0	XXXX	S2	17	75	E	U					1	M
13	8901	ENQ	0	MN	5647	2603	0	XXXX	R3	51	72		H					4	F
14	9201	HCC	1	LA	7079	0	0	XXXX	T2	40	•							•	M
15	9301	USB	1	UT	8472	2709	0	XXXX	T1	35	70	E	H					4	M
16	9401	FRC	1	CA	9005	0	0	XXXX	U1	2	•		H					1	M
17	9401	RKB	0	MI	4806	5401	0	XXXX	S2	20	44	E	U					1	M
18	8801	PCH	2	IL	6237	5201	0	XXXX	R2	43	46	E	U					7	F
19	8601	AMB	28	FL	3281	3601	0	XXXX	S2	16	62	E	H					4	F
20	9501	L15	1	NC	2785	0	0	XXXX	C2	27	•							•	M
21	8701	BBK	2	MN	5512	3601	0	XXXX	S1	12	62	E	H					3	F
22	9601	L21	1	MI	4924	1601	0	XXXX	R2	43	82	E	U					2	M
23	9401	SYN	0	FL	3398	0	0	XXXX	T2	40	•							•	
24	9301	L01	2	IL	6004	2311	0	XXXX	C1	22	74							•	F
25	9501	MOP	0	MN	5504	5201	0	XXXX	T1	35	46	E	H					7	F
26	9101	UCA	0	CA	9352	4307	0	XXXX	T1	35	54	I	H					7	M
27	9601	ESN	0	IL	6009	5601	0	XXXX	S1	13	42	E	H					7	F
28	9201	L01	1	MO	6446	1401	0	XXXX	T1	35	84	E	H					7	M
29	9101	IMP	0	TX	7738	4809	0	XXXX	T1	35	49	E	H	M	M			7	F
30	9101	AVN	0	IL	6224	6001	0	XXXX	T2	40	38	E	U					4	F
31	9001	SYN	0	TX	7754	0	0	XXXX	T1	35	•							•	M
32	9501	RMG	1	MO	6303	2601	0	XXXX	U3	8	72	E	H					1	M
33	9501	DNA	28	NC	2710	1401	0	XXXX	C2	25	84	E	U					4	F
34	9201	L04	3	FL	3314	2904	0	XXXX	S1	15	69	E	U					4	F
35	9101	AML	1	OR	9700	0	0	XXXX		•	•							•	F
36	9501		1	OR	9770	0	0	XXXX	R1	42	•							•	M
37	9501	AIR	1	TX	7874	2901	0	XXXX	S1	12	69	E	H					4	M
38	8601	DUR	0	CA	9027	1002	0	XXXX	S1	11	88		U					5	F
39	9001	LHJ	0	MN	5511	2301	0	XXXX	S2	17	75	E	H					6	F
40	8601	WKB	0	MN	5506	1311	0	XXXX	T2	36	84	E	H					3	F
41	9301	AGR	0	IL	6160	2801	0	XXXX	C2	28	70	I						•	F
42	8701	STL	2	MI	4950	0	0	XXXX	U1	2	•							•	
43	9501	DCD	1	WA	9800	6801	0	XXXX	S1	14	30	E	H					7	M
44	9301	AGR	1	FL	3333	5310	0	XXXX	T1	35	44	E	H			F		7	M
45	9201	IMA	0	WI	5322	4611	0	XXXX	S2	18	51	E	H					5	F
46	8601	DRK	0	IA	5240	3110	0	XXXX	C1	24	66	E	H					5	F
47	8601	GRI	0	AL	3632	2601	0	XXXX		•	72							4	F
48	9101	L01	1	IN	4635	908	0	XXXX	T1	34	88	E	H					3	M
49	8901	ACS	1	WA	9811	3706	0	XXXX	U2	5	61	I						•	M
50	9501	ALZ	1	TX	7560	3001	0	XXXX	C3	31	68	E	H					4	M
51	8801	DNA	0	IL	6127	1411	0	XXXX	R2	44	83	E	U					3	F
52	8601	DRK	0	NC	2770	5210	0	XXXX	C3	32	45	E	H				M	4	F
53	9101	NEX	0	NC	2710	5401	0	XXXX	C2	27	44	E	H					6	F
54	8701	IMP	28	FL	3305	3703	0	XXXX	S2	18	61	E	H					6	F
55	9301	AGR	1002	IL	6295	5801	0	XXXX	T2	36	40	E	H					2	M
56	9101		0	CA	9401	0	0	XXXX	R2	46	•							•	F
57	9501	HAR	0	CA	9280	0	0	XXXX	U1	3	•		U					1	F
58	9401	SGI	28	AL	3600	3601	0	XXXX	R3	50	62	E	H					3	F
59	9401	MBC	0	TX	7521	0	0	XXXX	S2	18	•							•	M
60	9401	BSH	0	IN	4698	5001	0	XXXX	R2	44	48	E	H					5	F
61	9601	BOA	0	TX	7751	6401	0	XXXX	T2	40	34	E	H			M		4	M
62	9501	NAD	2	GA	3132	1801	0	XXXX	R2	44	80	E	H					5	F
63	9401	HOS	0	OK	7345	0	0	XXXX	R3	53	•		H					3	M
64	9001	GRI	0	TX	7602	6201	0	XXXX	T2	40	36	E	H					4	F

PVA: Distribution of AGE

Distributions

AGE



Quantiles

100.0%	maximum	98
99.5%		95
97.5%		90
90.0%		83
75.0%	quartile	75
50.0%	median	62
25.0%	quartile	48
10.0%		39
2.5%		31
0.5%		26
0.0%	minimum	1

Moments

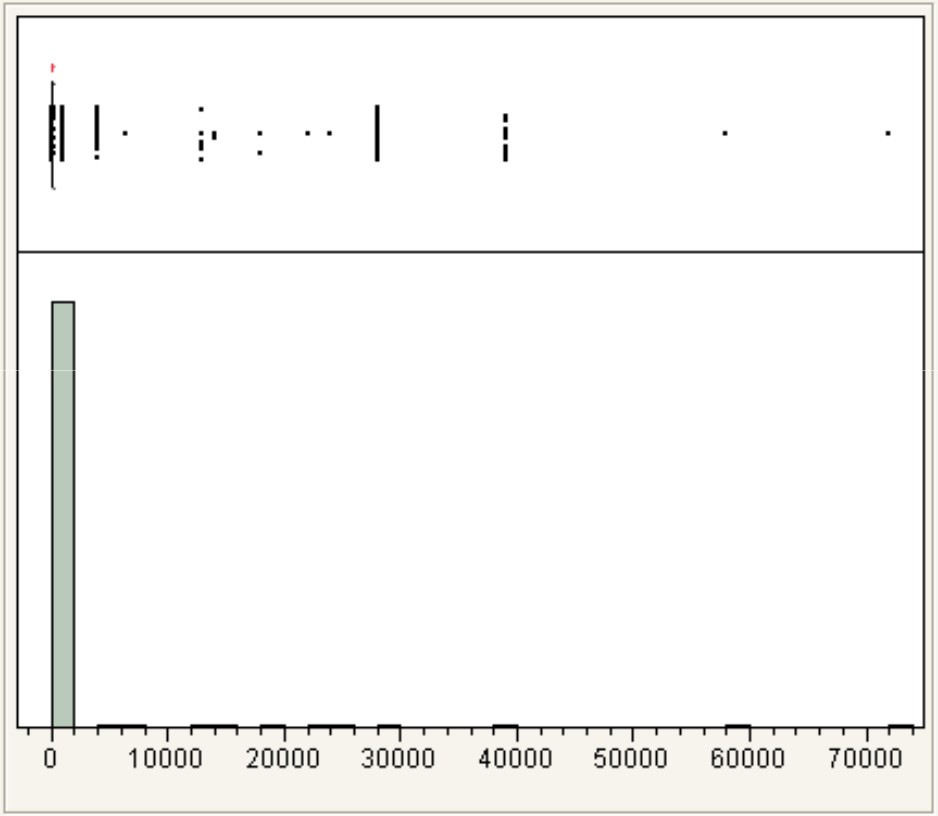
Mean	61.598019
Std Dev	16.666865
Std Err Mean	0.0624749
Upper 95% Mean	61.720469
Lower 95% Mean	61.475568
N	71170

T-Code

cup98lrn- Distribution of TCODE

Distributions

TCODE



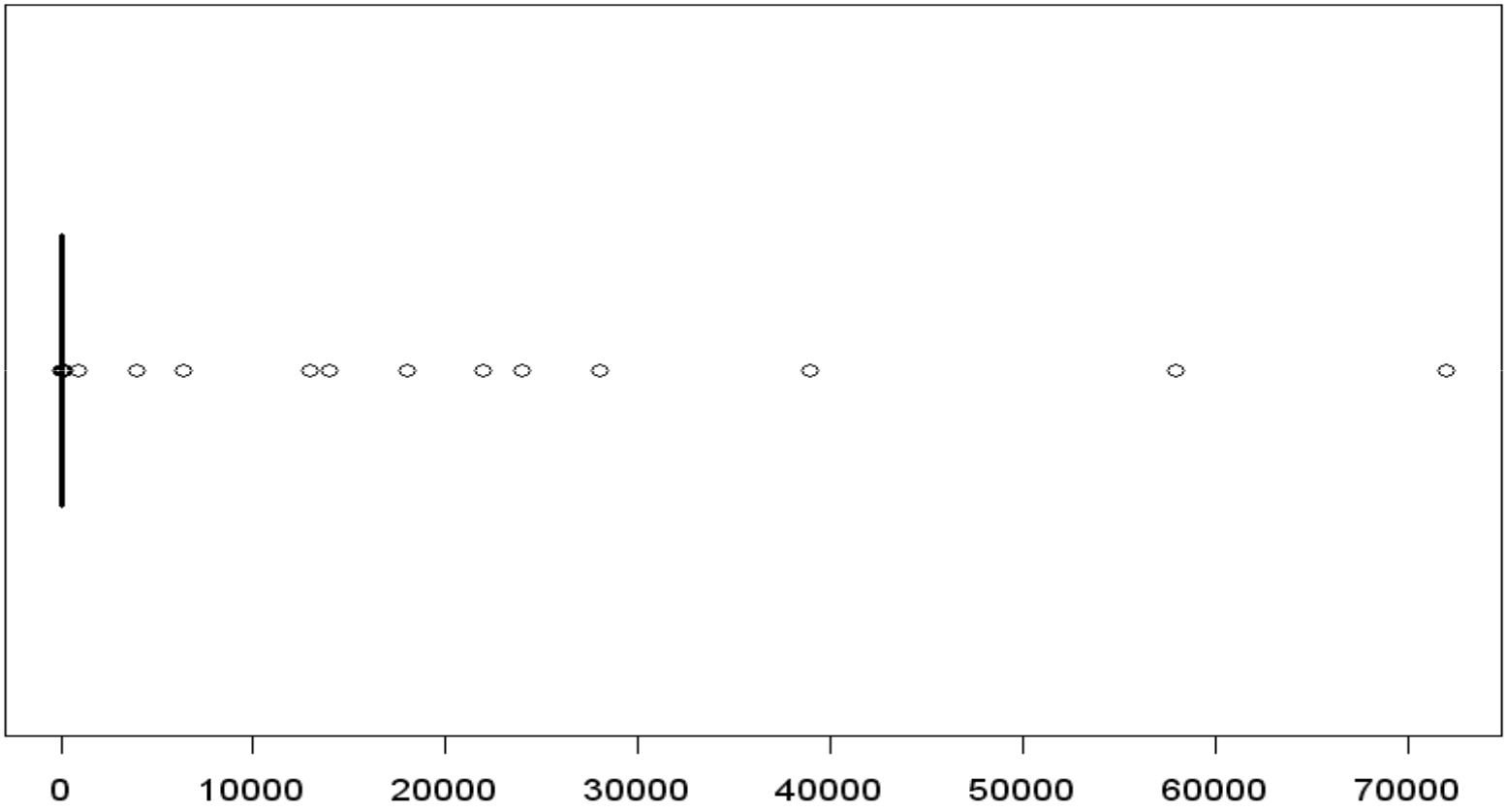
Quantiles

100.0%	maximum	72,002
99.5%		1,002
97.5%		28
90.0%		28
75.0%	quartile	2
50.0%	median	1
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

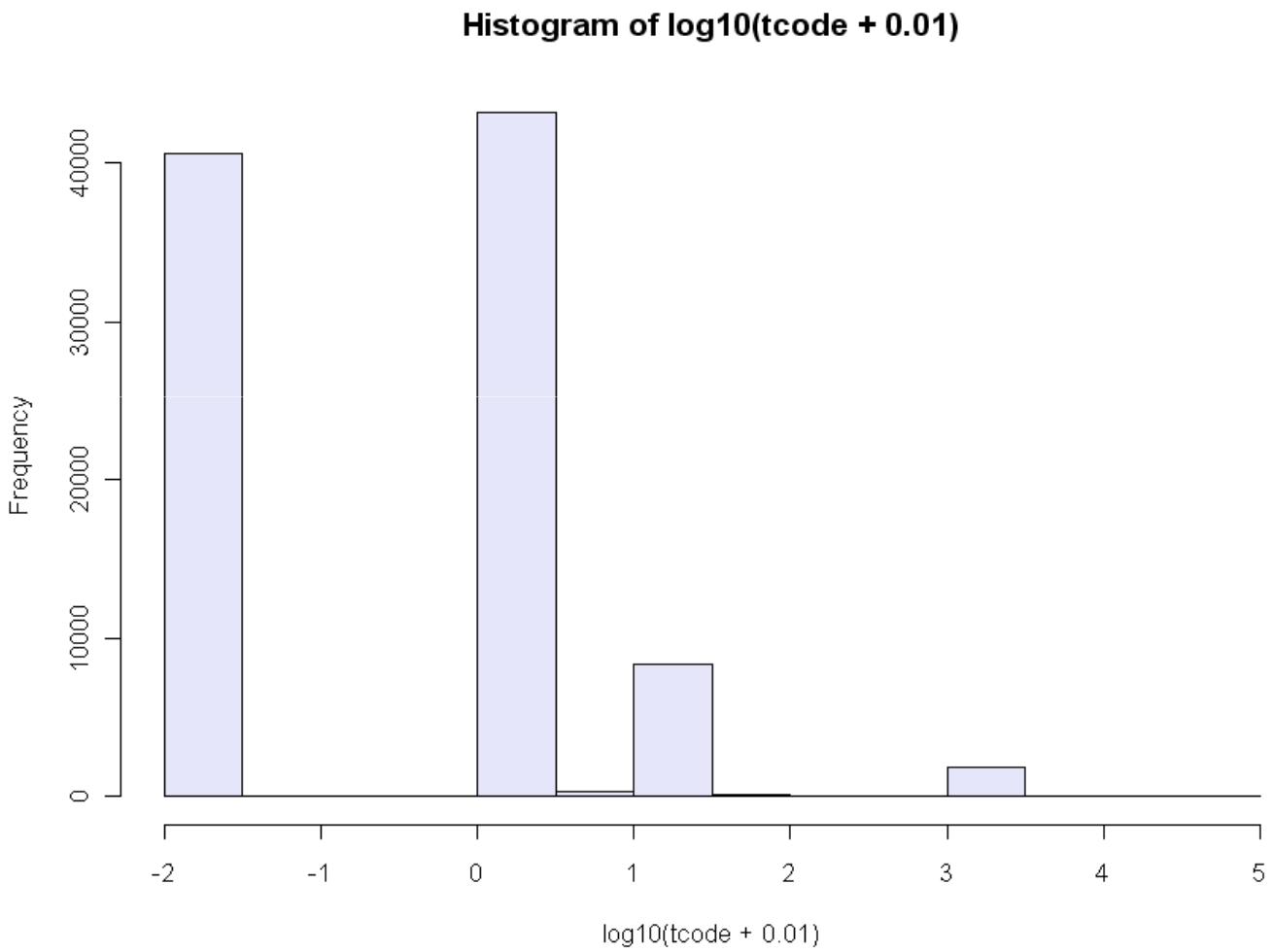
Moments

Mean	54.413105
Std Dev	957.50024
Std Err Mean	3.1122959
Upper 95% Mean	60.513171
Lower 95% Mean	48.313039
N	94,649

More T code



Transformation?



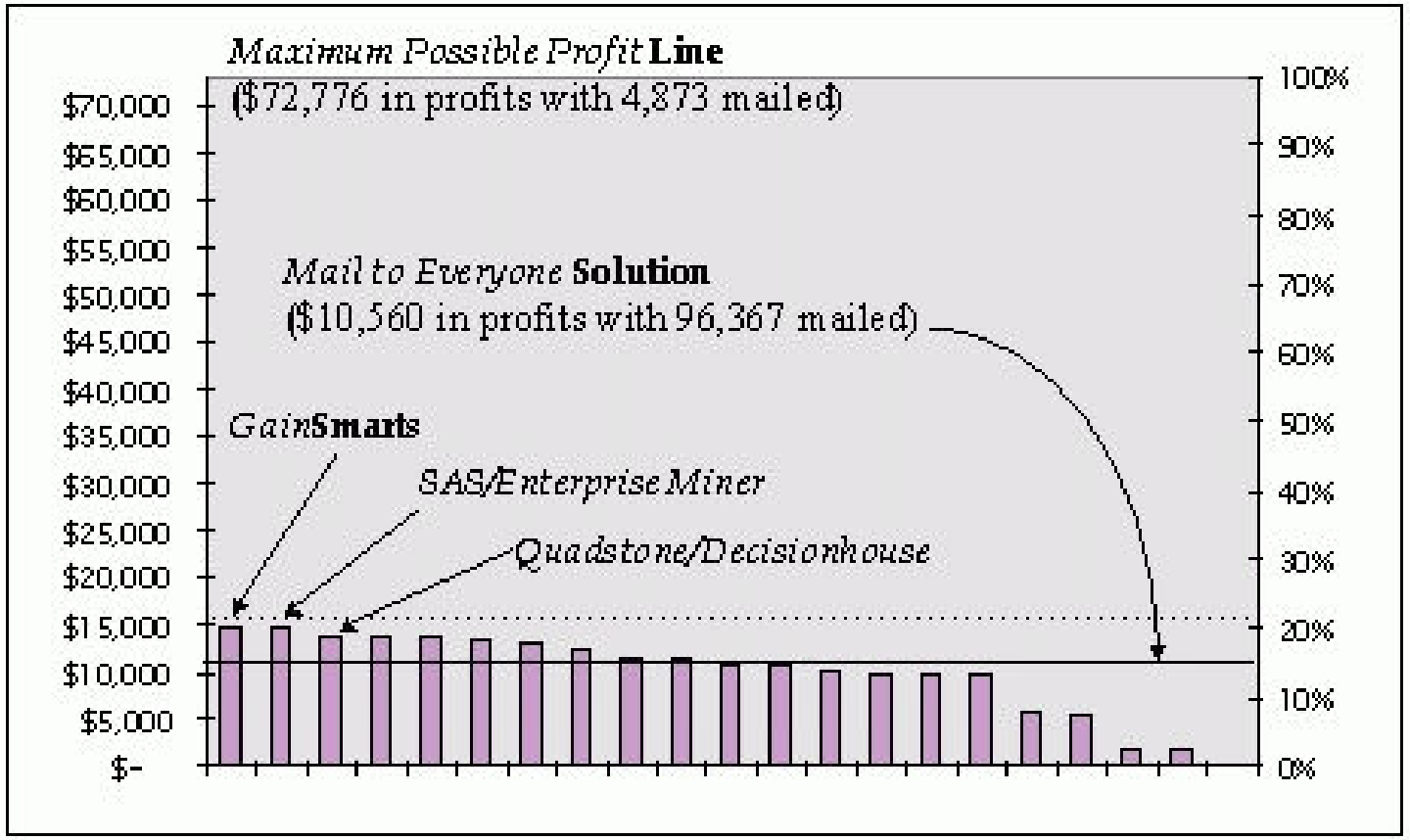
Categories?


0.0
0.1
0.2
0.3
0.4
0 1 2 3 4 6 7 9 12 13 14 17 18 21 22 23 24 25 27 28 30 38 40 42 44 45 47 50 61 72 76 93 94 96 116 136 228 1002 4004 13002 18002 24002 39002 72002

What does it mean?

T-Code	Title									
0	_	16	DEAN	48	CORPORAL	109	LIC.			
1	MR.	17	JUDGE	50	ELDER	111	SA.			
1001	MESSRS.	17002	JUDGE & MRS.	56	MAYOR	114	DA.			
1002	MR. & MRS.	18	MAJOR	59002	LIEUTENANT & MRS.	116	SR.			
2	MRS.	18002	MAJOR & MRS.	62	LORD	117	SRA.			
2002	MESDAMES	19	SENATOR	63	CARDINAL	118	SRTA.			
3	MISS	20	GOVERNOR	64	FRIEND	120	YOUR MAJESTY			
3003	MISSES	21002	SERGEANT & MRS.	65	FRIENDS	122	HIS HIGHNESS			
4	DR.	22002	COLNEL & MRS.	68	ARCHDEACON	123	HER HIGHNESS			
4002	DR. & MRS.	24	LIEUTENANT	69	CANON	124	COUNT			
4004	DOCTORS	26	MONSIGNOR	70	BISHOP	125	LADY			
5	MADAME	27	REVEREND	72002	REVEREND & MRS.	126	PRINCE			
6	SERGEANT	28	MS.	73	PASTOR	127	PRINCESS			
9	RABBI	28028	MSS.	75	ARCHBISHOP	128	CHIEF			
10	PROFESSOR	29	BISHOP	85	SPECIALIST	129	BARON			
10002	PROFESSOR & MRS.	31	AMBASSADOR	87	PRIVATE	130	SHEIK			
10010	PROFESSORS	31002	AMBASSADOR & MRS	89	SEAMAN	131	PRINCE AND PRINCESS			
11	ADMIRAL	33	CANTOR	90	AIRMAN	132	YOUR IMPERIAL MAJEST			
11002	ADMIRAL & MRS.	36	BROTHER	91	JUSTICE	135	M. ET MME.			
12	GENERAL	37	SIR	92	MR. JUSTICE	210	PROF.			
12002	GENERAL & MRS.	38	COMMODORE	100	M.					
13	COLONEL	40	FATHER	103	MLLE.					
13002	COLONEL & MRS.	42	SISTER	104	CHANCELLOR					
14	CAPTAIN	43	PRESIDENT	106	REPRESENTATIVE					
14002	CAPTAIN & MRS.	44	MASTER	107	SECRETARY					
15	COMMANDER	46	MOTHER	108	LT. GOVERNOR					
15002	COMMANDER & MRS.	47	CHAPLAIN							

KDD-CUP-98 Results (2 of 2)



KDD-CUP-98 Results (1 of 2)

Participants	Sum of Actual Profits	Number Mailed	Average Profits
Student #1 $15,024	$ 14,712.24	56,330	0.26
Student #2 $14,695	$ 14,662.43	55,838	0.26
Student #3 $14,345	$ 13,954.47	57,836	0.24
# 4	$ 13,824.77	55,650	0.25
# 5	$ 13,794.24	51,906	0.27
# 6	$ 13,598.05	55,830	0.24
# 7	$ 13,040.46	60,901	0.21
# 8	$ 12,298.23	48,304	0.25
# 9	$ 11,422.77	56,144	0.20
# 10	$ 11,276.46	90,976	0.12
# 11	$ 10,719.88	62,432	0.17
# 12	$ 10,706.34	65,286	0.16
# 13	$ 10,112.08	64,044	0.16
# 14	$ 10,048.72	76,994	0.13
# 15	$ 9,740.72	54,195	0.18
# 16	$ 9,463.77	79,294	0.12
# 17	$ 5,682.91	51,477	0.11
# 18	$ 5,483.67	30,539	0.18
# 19	$ 1,924.69	50,475	0.04
# 20	$ 1,706.17	42,270	0.04
# 21	$ (53.68)	1,551	-0.03

Ismail Parsa KDD-CUP-98 8/98 epsilon

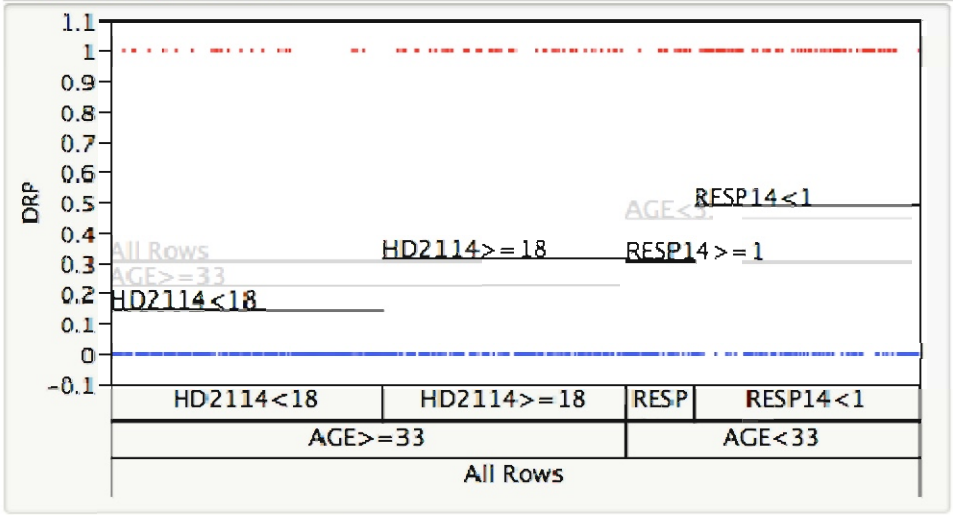
Ignoring What's Not There

Depression Clinical Trial Study

- Designed to study antidepressant efficacy
  - Measured via Hamilton Rating Scale
- Side effects
  - Sexual dysfunction
  - Misc safety and tolerability issues
- 428 patients
- Two antidepressants + placebo

Depression Study: Partition of DRP

Partition for DRP



Split Prune

RSquare	RMSE	N	Number of Splits	Imputes	AICc
0.084	0.4401157	428	3	12	522.219

All Rows

Count	428	LogWorth	Difference
Mean	0.3037383	5.3158193	0.22172
Std Dev	0.4604092		

AGE>=33

Count	273	LogWorth	Difference
Mean	0.2234432	2.5166817	0.16424
Std Dev	0.417318		

AGE<33

Count	155	LogWorth	Difference
Mean	0.4451613	1.2610131	0.18184
Std Dev	0.4985946		

HD2114<18

Count	144
Mean	0.1458333
Std Dev	0.354171

▶ Candidates

HD2114>=18

Count	129
Mean	0.3100775
Std Dev	0.4643283

▶ Candidates

RESP14>=1

Count	36
Mean	0.3055556
Std Dev	0.4671766

▶ Candidates

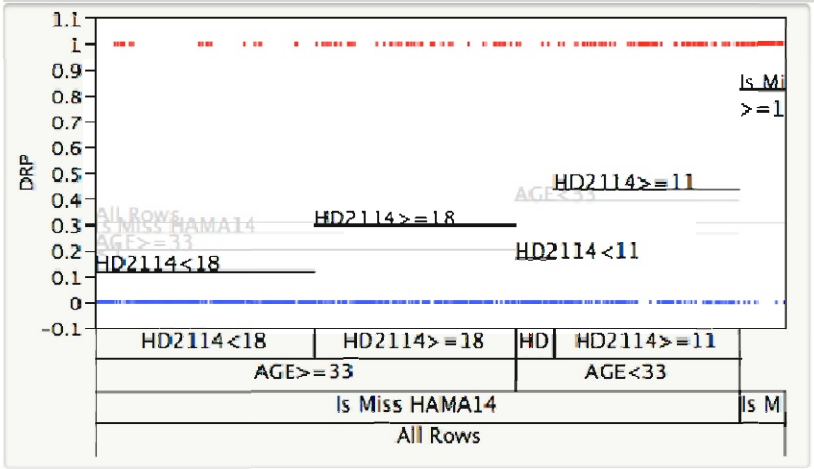
RESP14<1

Count	119
Mean	0.487395
Std Dev	0.5019546

▶ Candidates

Depression Study: Partition of DRP

Partition for DRP



Split Prune

RSquare	RMSE	N	Number of Splits	AICc
0.162	0.4210218	428	4	486.31

All Rows

Count	428
Mean	0.3037383
Std Dev	0.4604092

LogWorth 10.199513 Difference 0.55393

Is Miss HAMA14 <1

Count	400
Mean	0.2675
Std Dev	0.4432097

LogWorth 3.2981702 Difference 0.18542

Is Miss HAMA14 >=1

Count	28
Mean	0.8214286
Std Dev	0.390021

Candidates

AGE>=33

Count	261
Mean	0.2030651
Std Dev	0.4030535

LogWorth 2.5166817 Difference 0.17835

AGE<33

Count	139
Mean	0.3884892
Std Dev	0.4891695

LogWorth 0.8526589 Difference 0.26812

HD2114<18

Count	136
Mean	0.1176471
Std Dev	0.3233808

Candidates

HD2114>=18

Count	125
Mean	0.296
Std Dev	0.458328

Candidates

HD2114<11

Count	24
Mean	0.1666667
Std Dev	0.3806935

Candidates

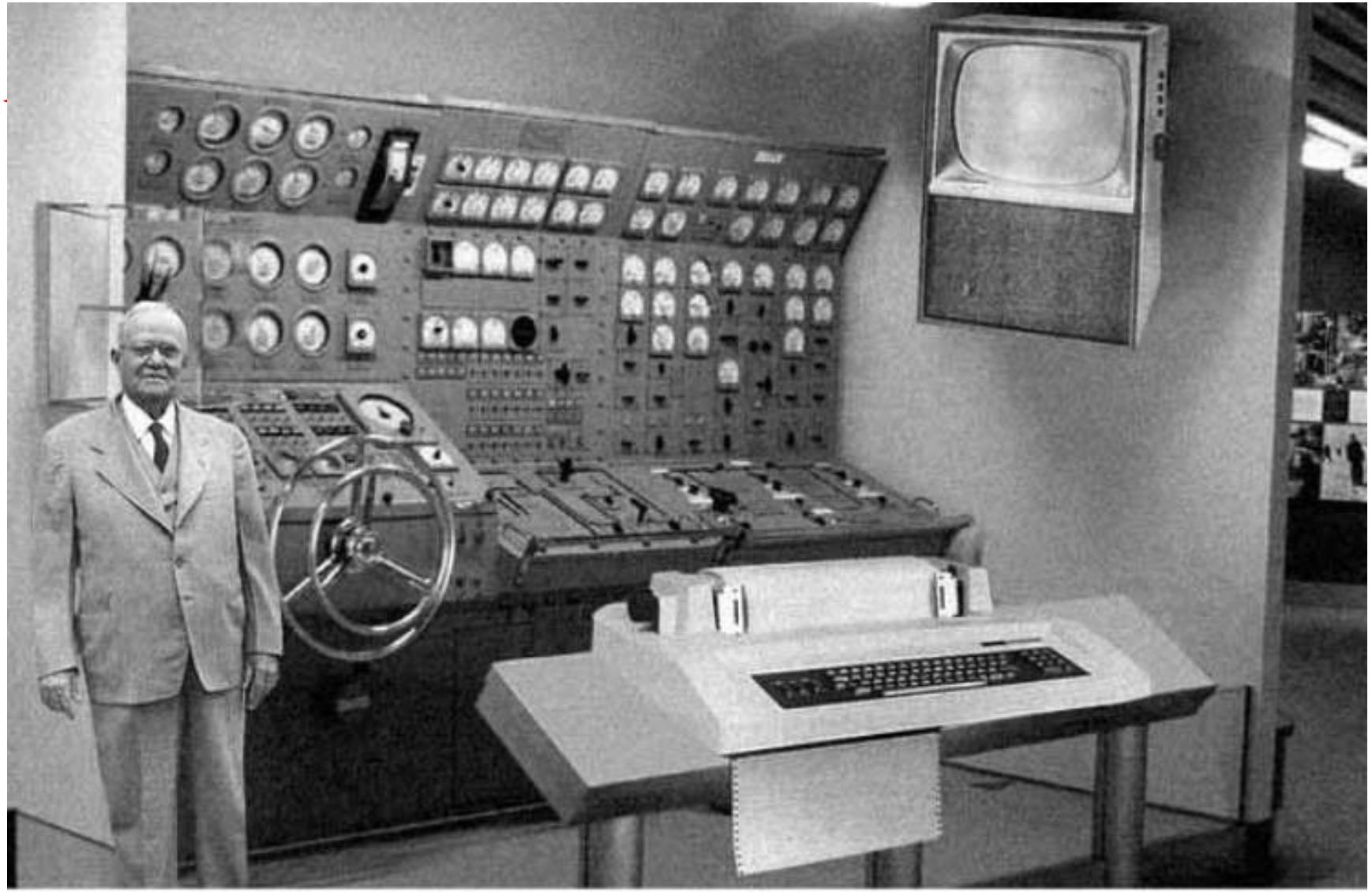
HD2114>=11

Count	115
Mean	0.4347826
Std Dev	0.4978979

Candidates

Falling in Love With Your Models

Prediction is Hard



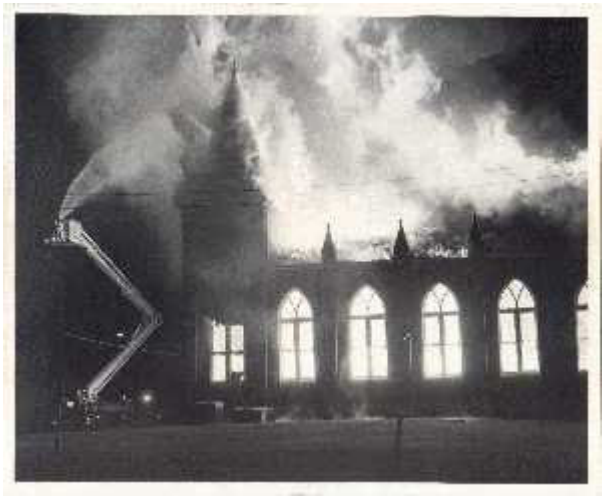
Scientists from the RAND Corporation have created this model to illustrate how a "home computer" could look like in the year 2004. However the needed technology will not be economically feasible for the average home. Also the scientists readily admit that the computer will require not yet invented technology to actually work, but 50 years from now scientific progress is expected to solve these problems. With teletype interface and the Fortran language, the computer will be easy to use.

Car Insurance

- 42800 mature policies on fleets of commercial car insurance
  - 65 Potential Predictors
  - Can we find a pattern for the unprofitable policies?

ITI Cavtat

But It Worked for Cars!



- Liability for churches
  - Some Predictors
    - Net Premium Value
    - Property Value
    - Coastal (yes/no)
    - Inner100 (a.k.a., highly-urban) (yes/no)
    - High property value Neighborhood (yes/no)
    - Indicator Class
      - 1 (Church/House of worship)
      - 2 (Sexual Misconduct – Church)
      - 3 (Add'l Sex. Misc. Covg Purchased)
      - 4 (Not-for-profit daycare centers)
      - 5 (Dwellings – One family (Lessor's risk))
      - 6 (Bldg or Premises – Office – Not for profit)
      - 7 (Corporal Punishment – each faculty member)
      - 8 (Vacant land- not for profit)
      - 9 (Private, not for profit, elementary, Kindergarten and Jr. High Schools)
      - 10 (Stores – no food or drink – not for profit)
      - 11 (Bldg or Premises – Maintained by insured (lessor's risk) – not for profit)
      - 12 (Sexual misconduct – diocese)

ITI Cavtat

Predicting Malignancy

- Breast cancer data from mammograms
  - Error rates by trained radiologists are near 25% for both false positives and false negatives
- Newer equipment is prohibitively expensive for the developing world
- Early detection of breast cancer is crucial
- Cumulative type I error over a decade is near 100% leading to needless biopsies

The Data

- 1618 mammograms showing clustered microcalcifications
  - Biostatistics Dept Institut Curie
- Variables
  - Response: Malignant or not
  - Predictors: Age, Tissue Type (light/dense) Size (mm), Number of microcalc, Number of suspicious clusters, Shape of microcalc (1-5), Polyshape?(y/n), Shape of cluster (1,2,3), Retro (cluster near nipple?), Deep? (y/n)

Combining Models

- Bagging (Bootstrap Aggregation)
  - Bootstrap a data set repeatedly
  - Take many versions of same model (e.g. tree)
    - Random Forest Variation
  - Form a committee of models
  - Take majority rule of predictions
- Boosting
  - Create repeated samples of weighted data
  - Weights based on misclassification
  - Combine by majority rule, or linear combination of predictions

Bootstrap Forest Wins

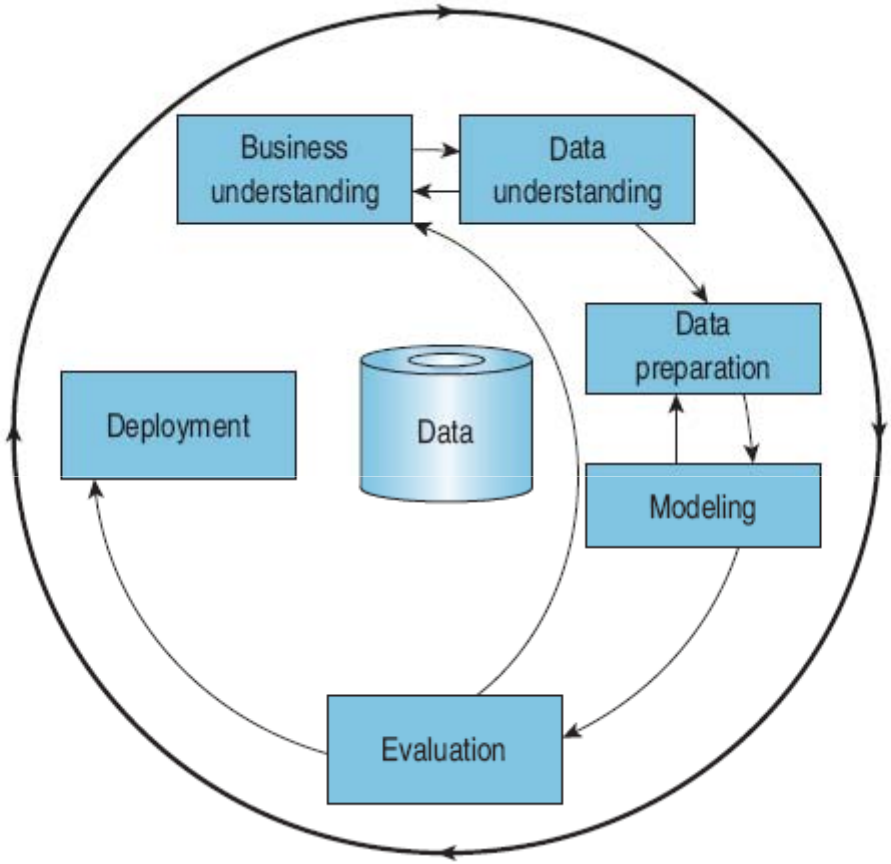
	False Positives	False Negatives
Simple Tree	32.20%	33.70%
Neural Network	25.50%	31.70%
Boosted Trees	24.90%	32.50%
Boostrap Forest	**19.30%**	**28.80%**
Radiologists	22.40%	35.80%

What About Interpretability?

- Column Contributions
  - Bootstrap Forest
  - Even better than “Shaking the Tree”
- Use as a Benchmark
  - Explore interpretable models that have same out of sample predictive power as ensemble methods

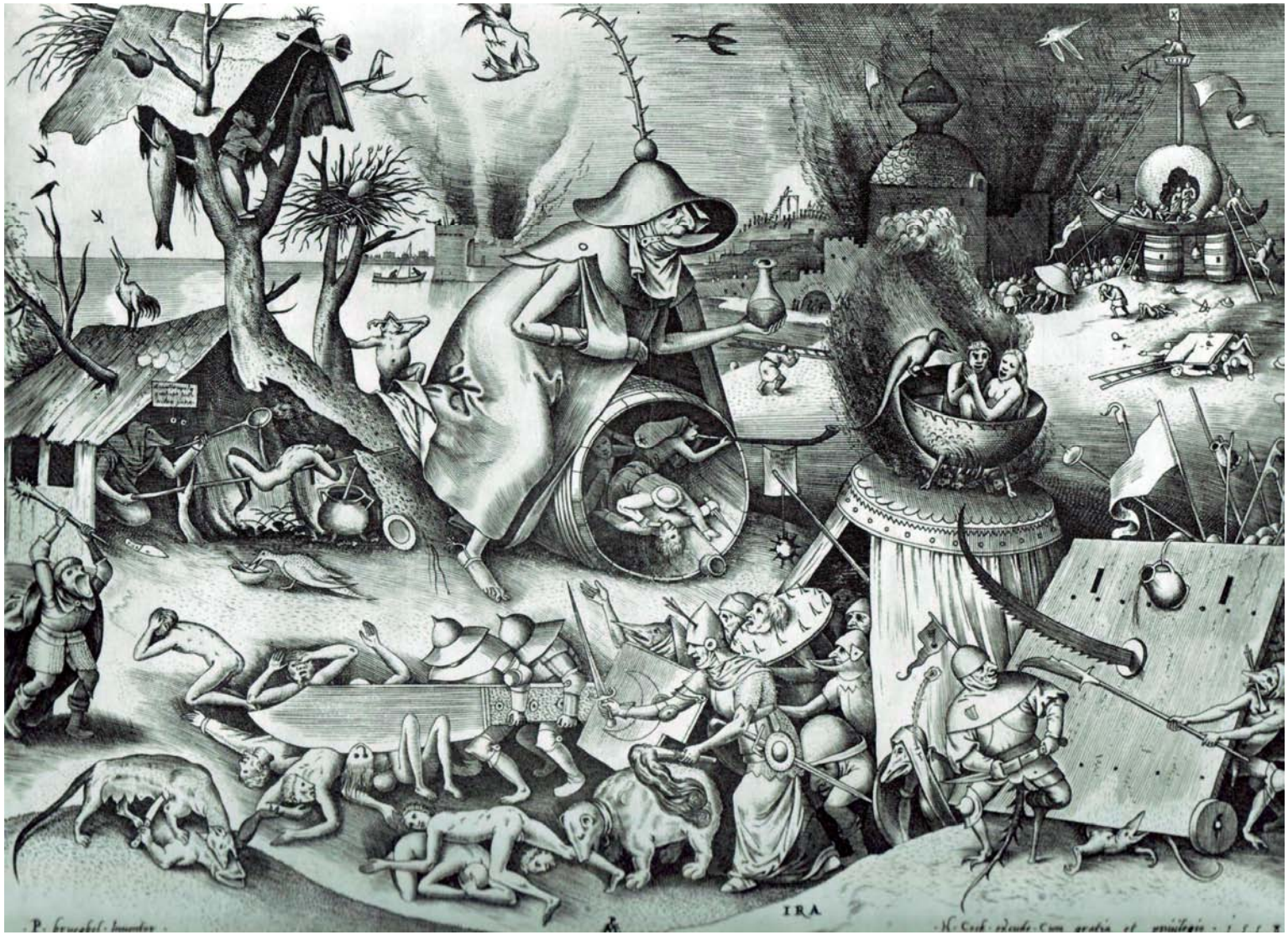
Going it Alone

The Process



CRoss-Industry Standard Process for Data Mining (CRISP –DM)

Using Bad Data

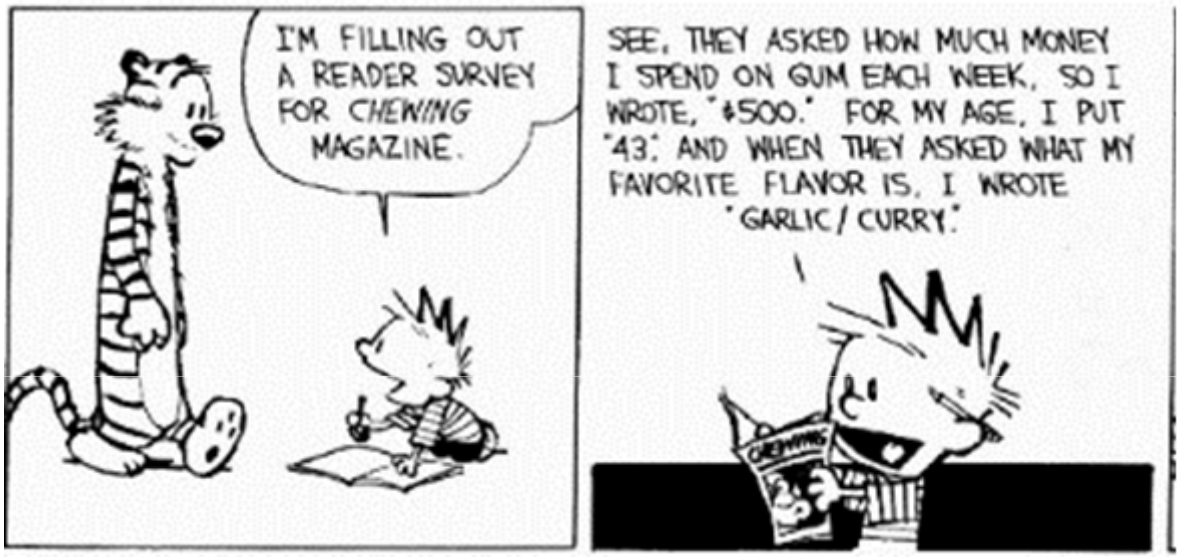


Where Do They Come From?

- The *Times* of London reports 30kg bat
  - The mysterious moving decimal
- 40% of doctors born on Veteran's Day 1911
  - Data Entry
- Mars Lander lost -- $125 M
  - Confusion of acceleration units (metric – newtons/sec vs. English pound/sec) Everywhere
- Ovarian Cancer cure published in *Lancet*
  - Differences due to lab practice, not treatment

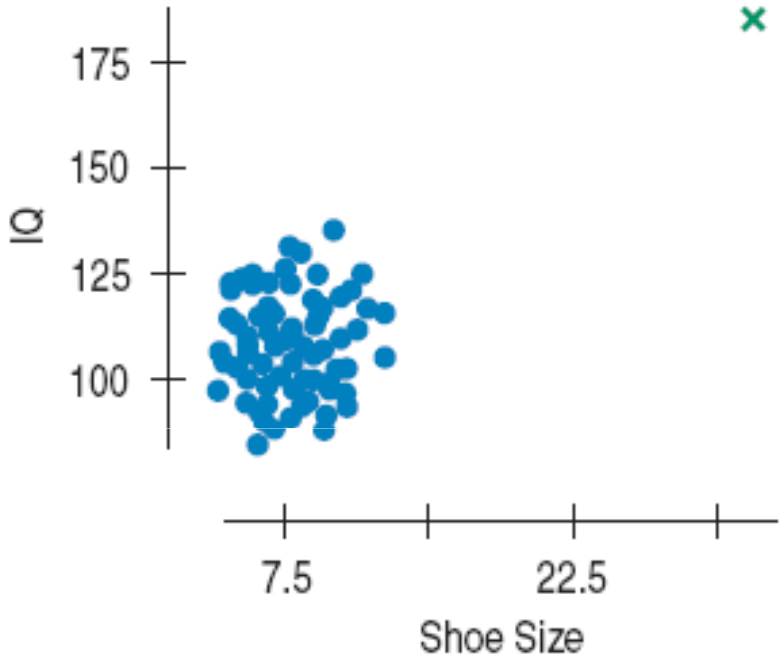
Where Do They Really Come From?

Where Do They Really Come From?



Where Do They Really Come From?

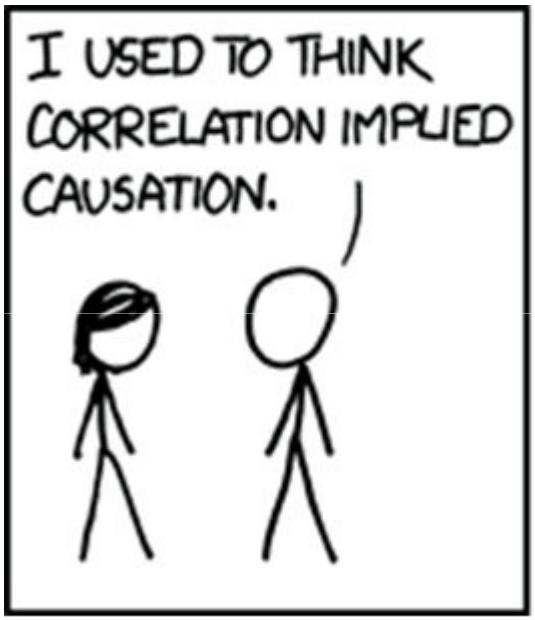
What Can They Do?



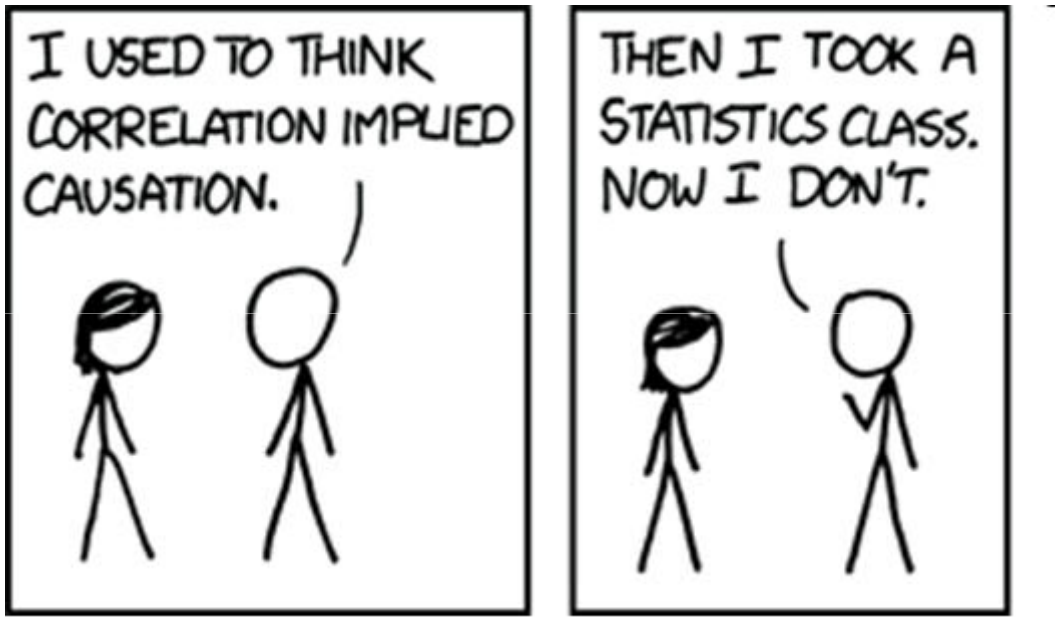
- Study by Large Credit Card Issuer
- Entire effect was due to one customer who charged $3M

Confusing Correlation and Causation

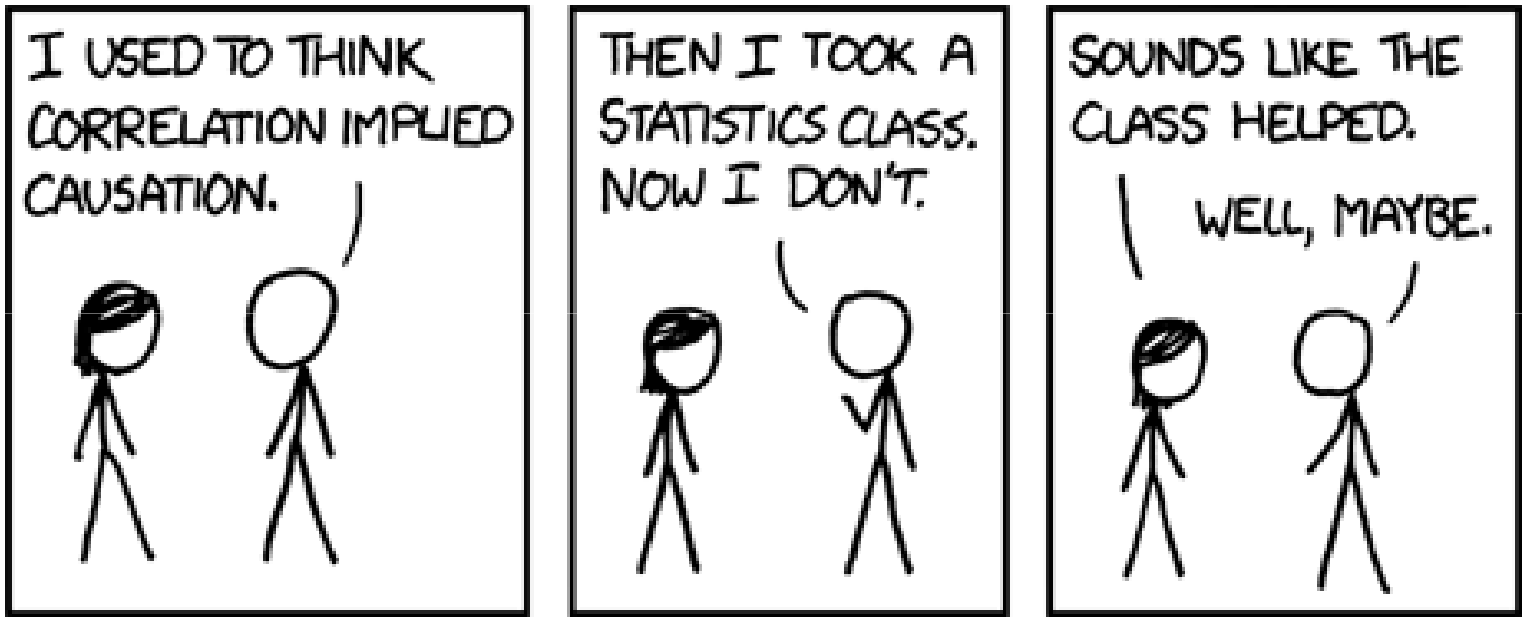
Did it Matter?



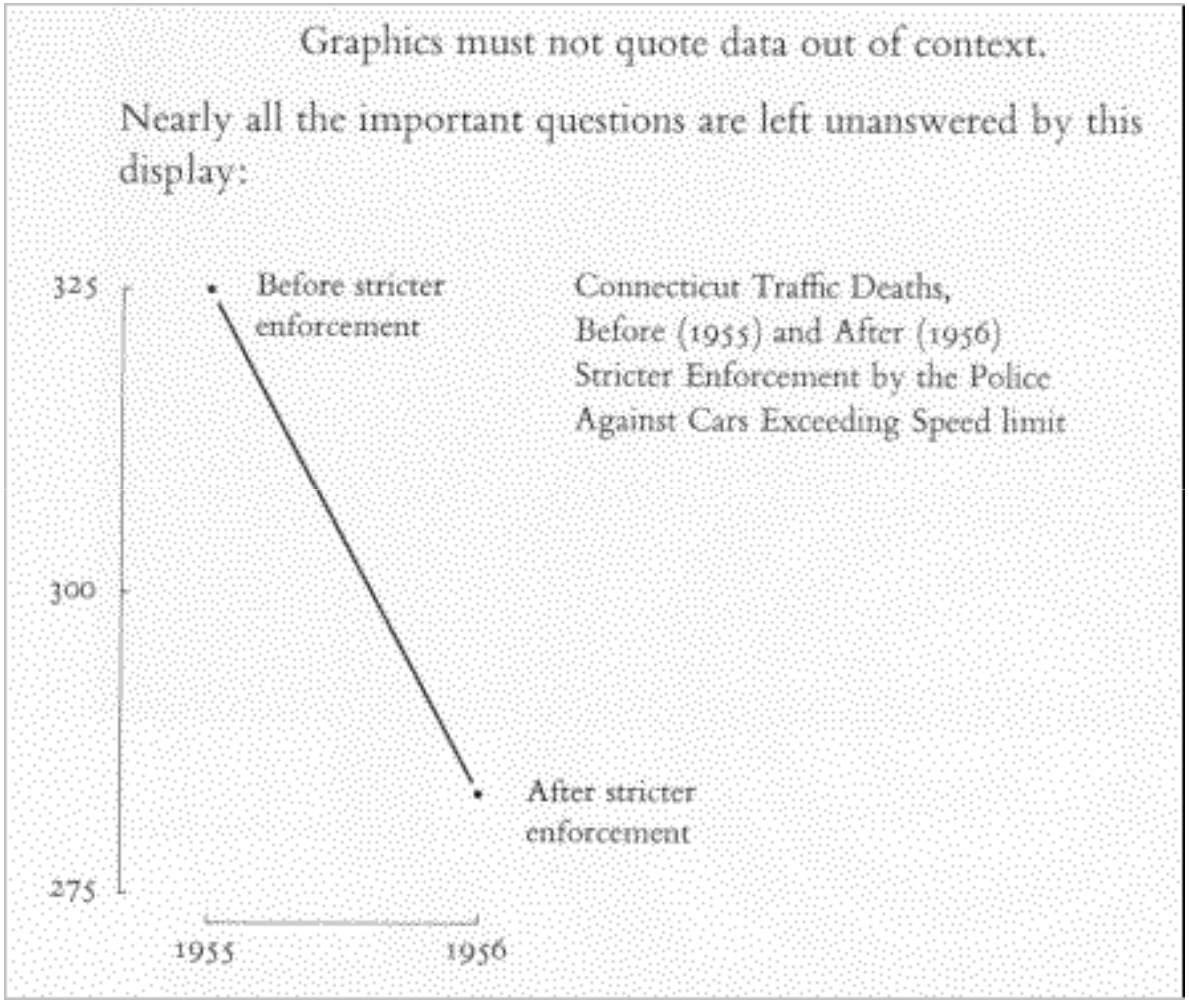
Did it Matter?



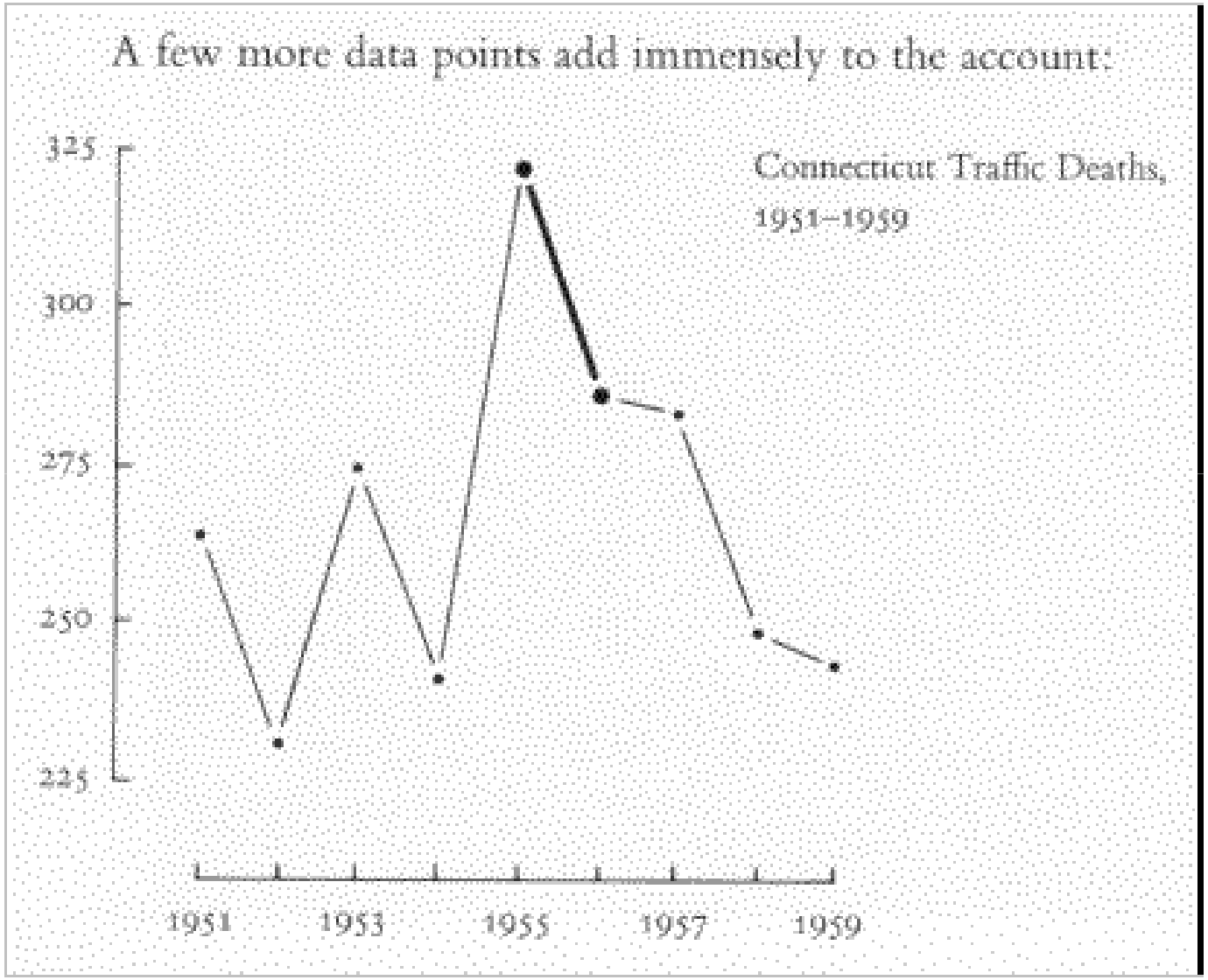
Did it Matter?



The Managers' (and all of our) Folly



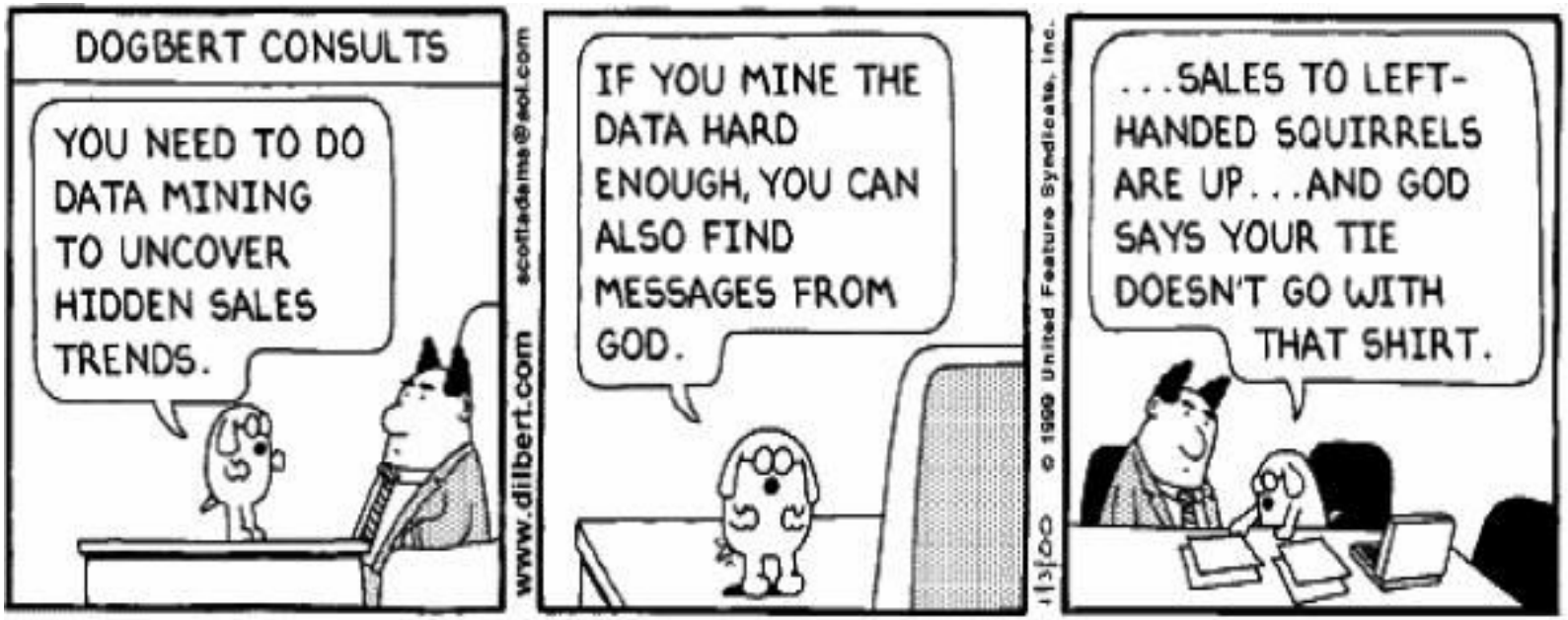
Other Factors?



Associations

Associations

Associations



Less Far Fetched?

- Did you know?
  - Shorter men are at greater risk for heart attacks?
- Hunch
  - 55 million answers to questions from 1 million users
- Did you know?
  - People who swat flies like *USA Today*?
  - People who believe in alien abductions prefer Pepsi to Coke
  - People who eat fresh fruit every day are more likely to buy a pricey digital camera
  - People who cut sandwiches diagonally prefer Ray-Ban sunglasses
- Social Network Marketing Produces More Loyal Customers
  - Chicken or egg?

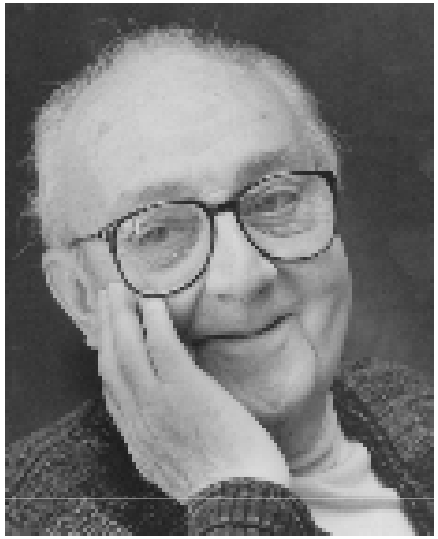
The End of Science

ITI Cavtat

Poor George

George Box

> "All models are wrong, but some are useful"



Peter Norvig

> "All models are wrong, and increasingly you can succeed without them."

Really?

Chris Anderson

> "With enough data, the numbers speak for themselves"

The Seven Virtues

- Define the Problem
- Prepare the Data
  - use Domain Knowledge
- Be Open to New Methods and Models
- Be Aware of Missing Data
- Work in Teams
- Ensure Data Quality
- Use Models, not just Associations