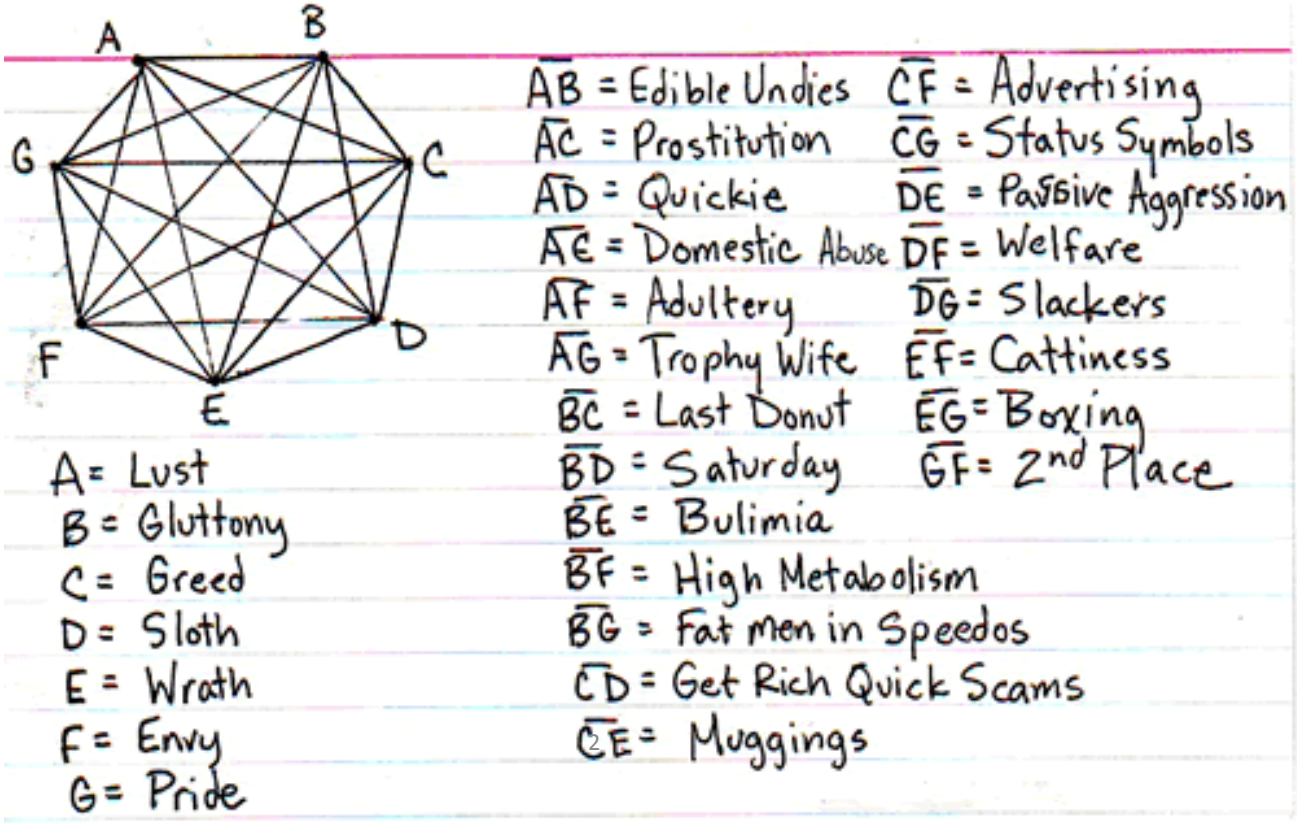
Seven Deadly Sins of Big Data

Richard De Veaux
Williams College
FTC
October 5, 2017

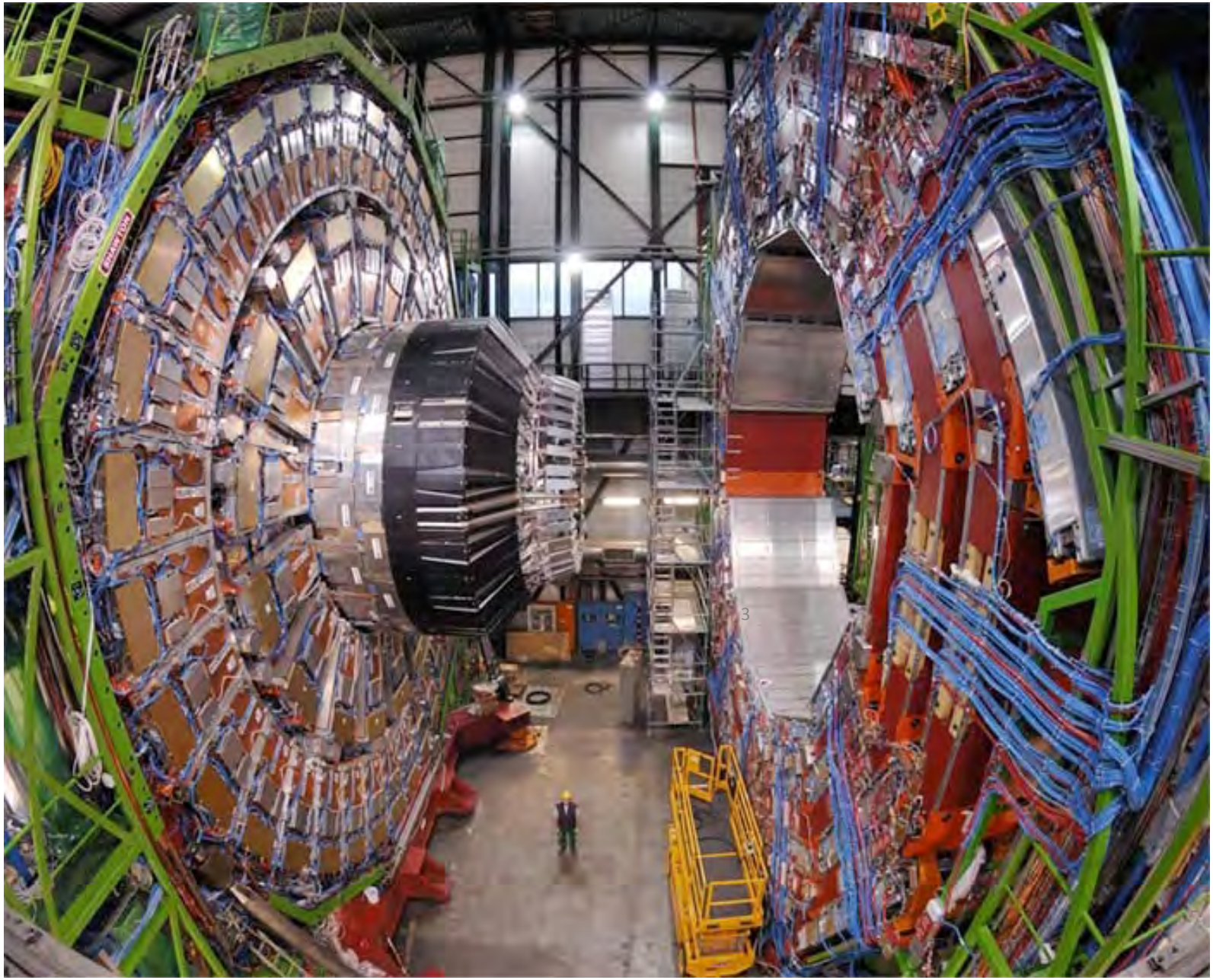
The Original Seven Sins

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

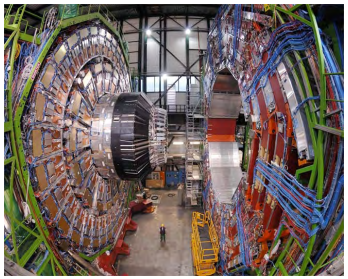


Thanks to Ewan's Corner,
Source Unknown

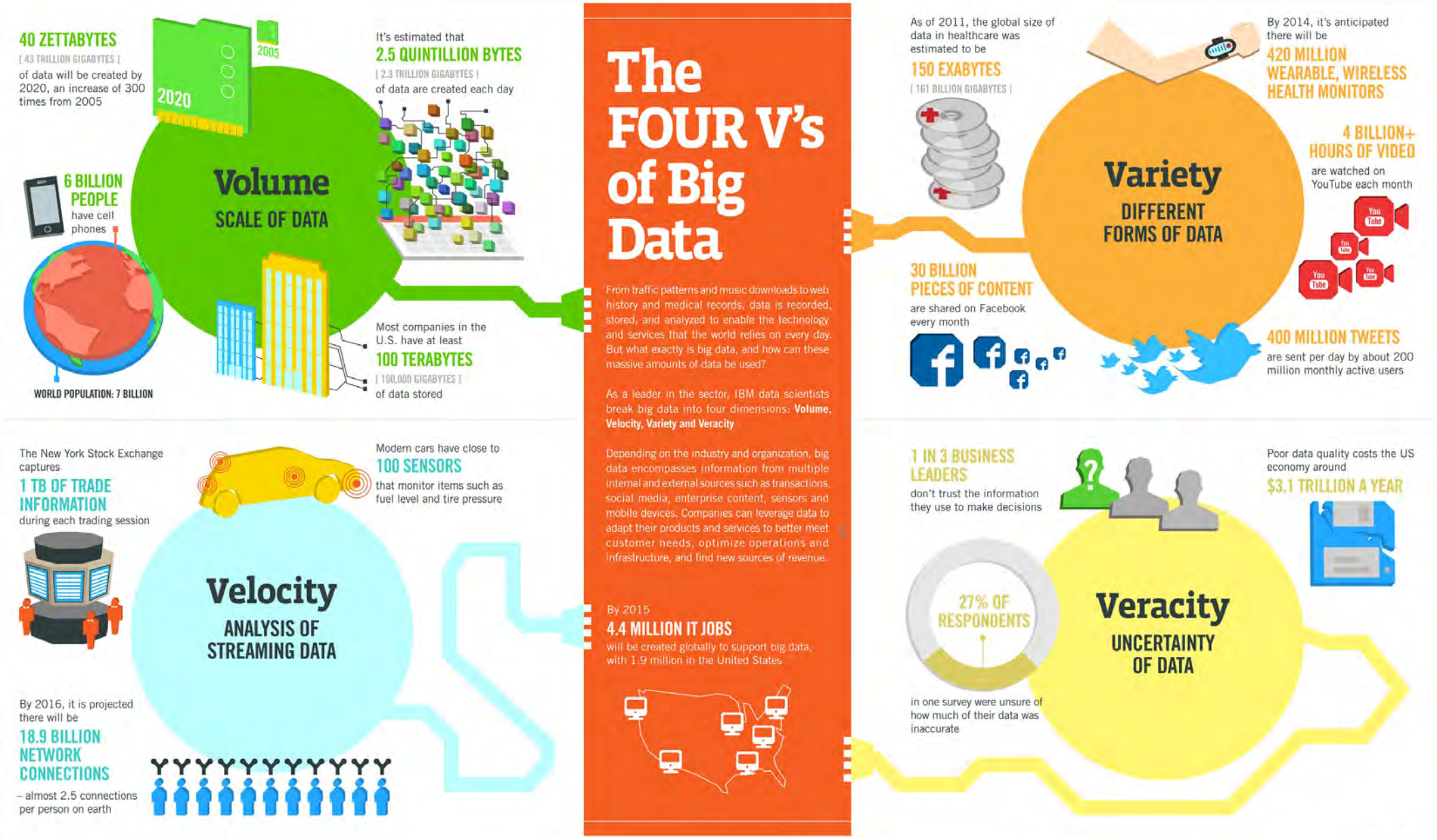
What is Big Data?



How Big Is Big?



What makes Big data Big?



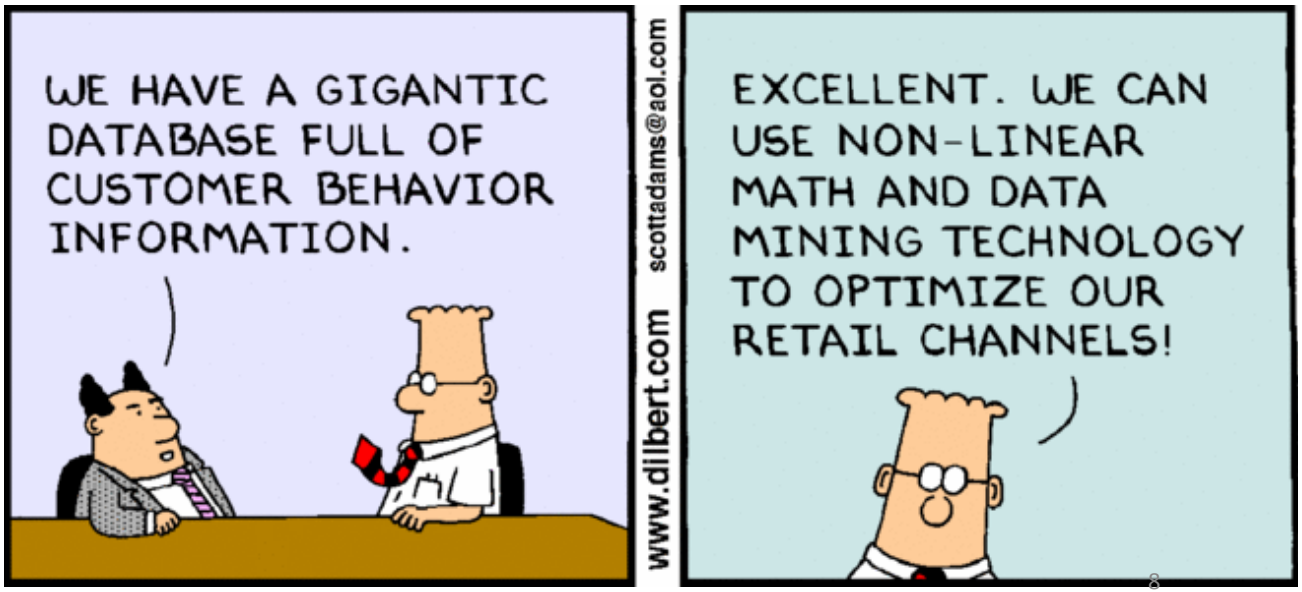
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Origins of Big Data

What is it really?

What is it really?



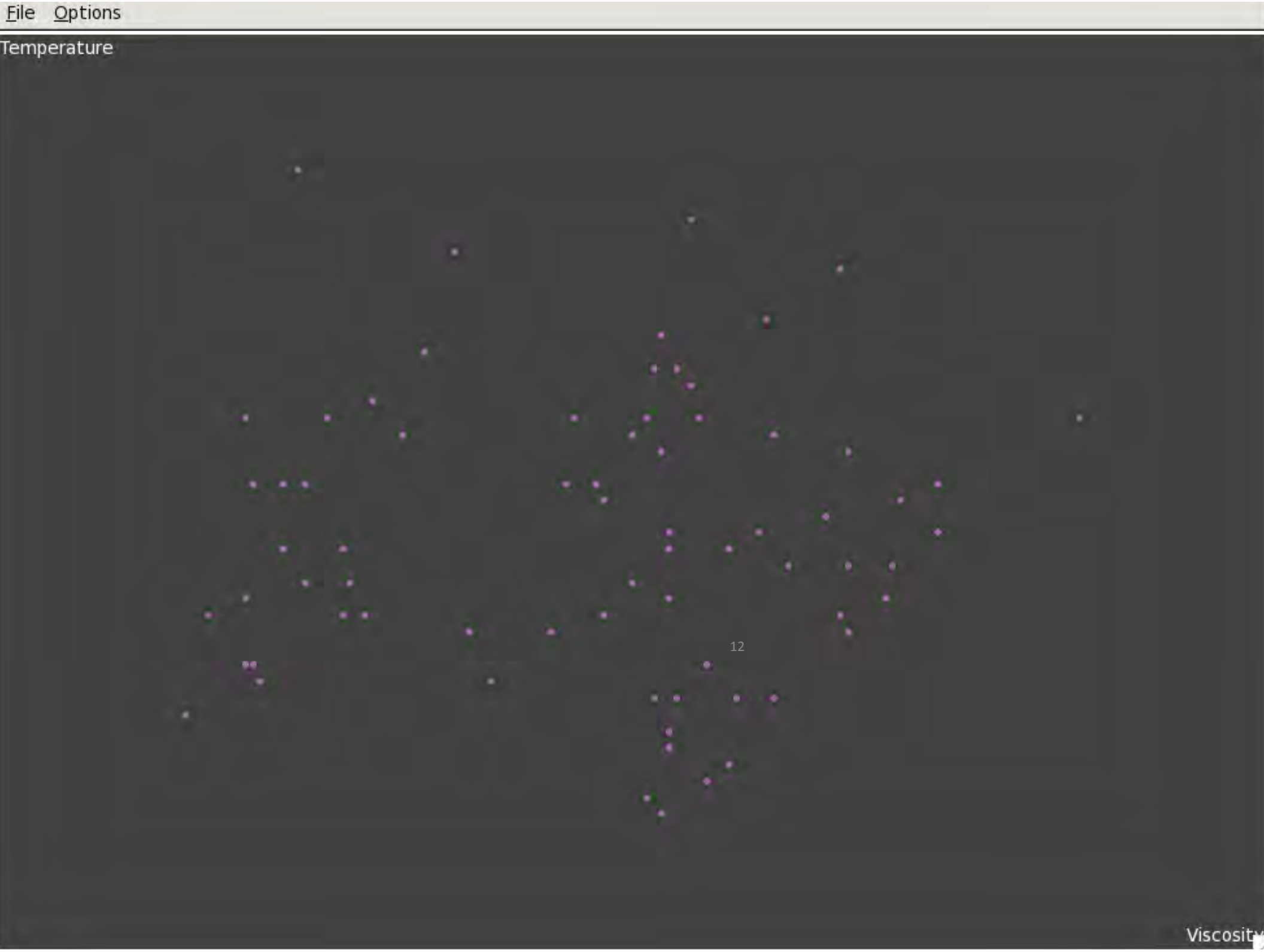
What is it really?

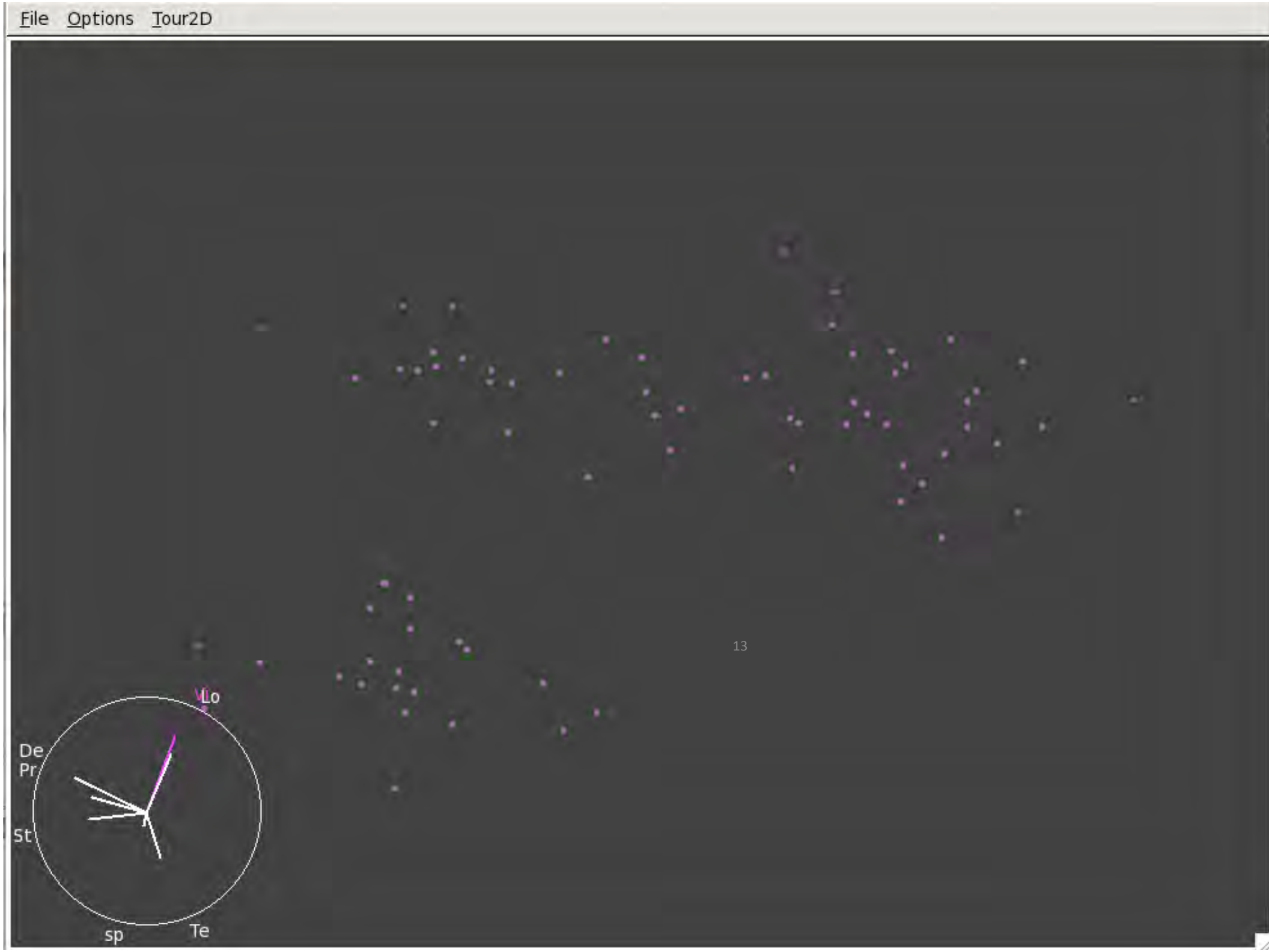
So, What are the Seven Deadly Sins of Big Data?

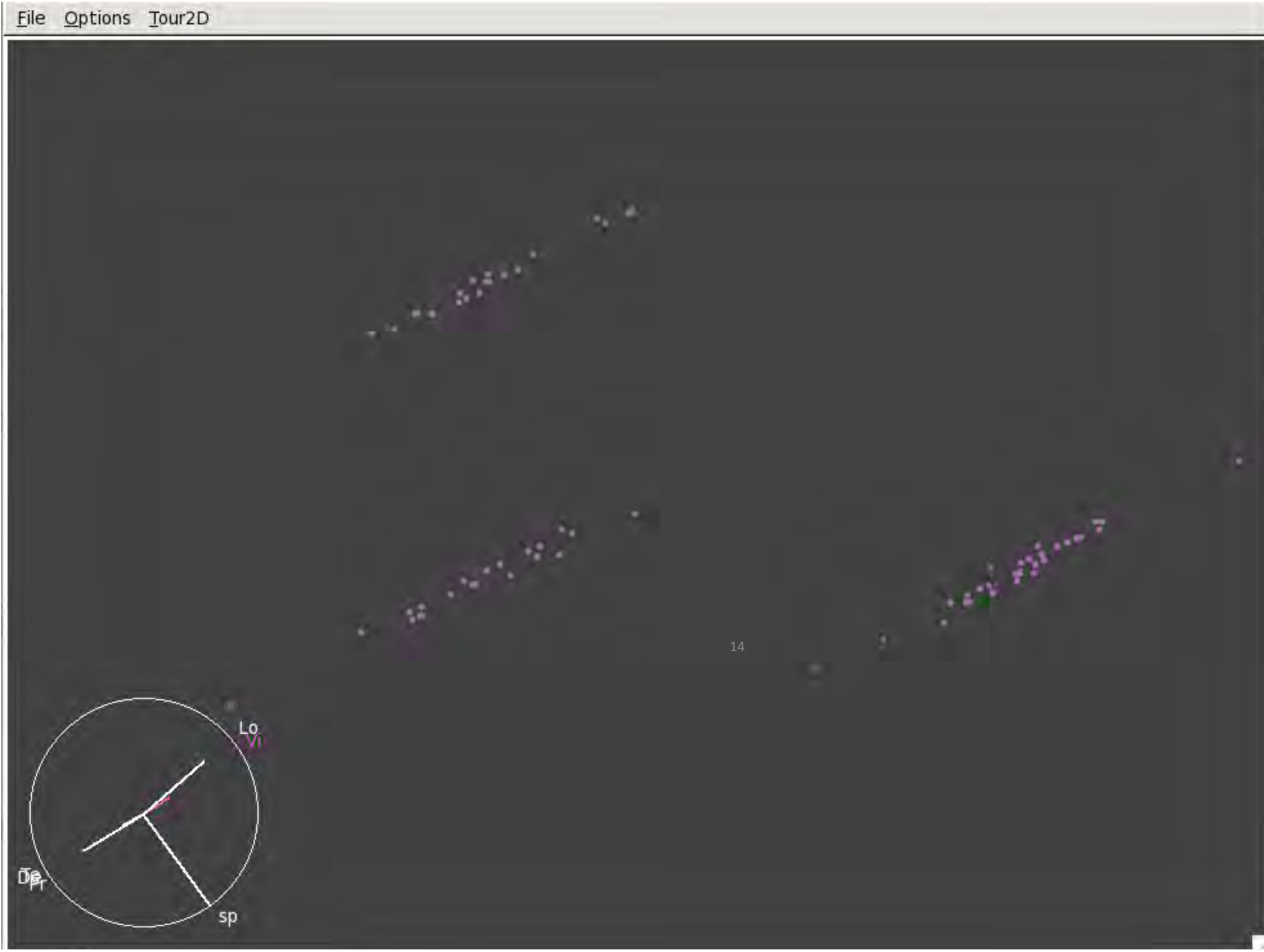
1. Failing to Define the Problem

Production Analysis

- Statistician studies properties for 74000 recent samples from production
  - Viscosity
  - Max temperature to failure
  - Max pressure to failure
  - Density
  - Load at failure
  - Stress at failure


File Options
Temperature
12
Viscosity


File
Options
Tour2D
13
Lo
De
Pr
St
sp
Te


File
Options
Tour2D
14
Lo
sp

Three Groups!!!

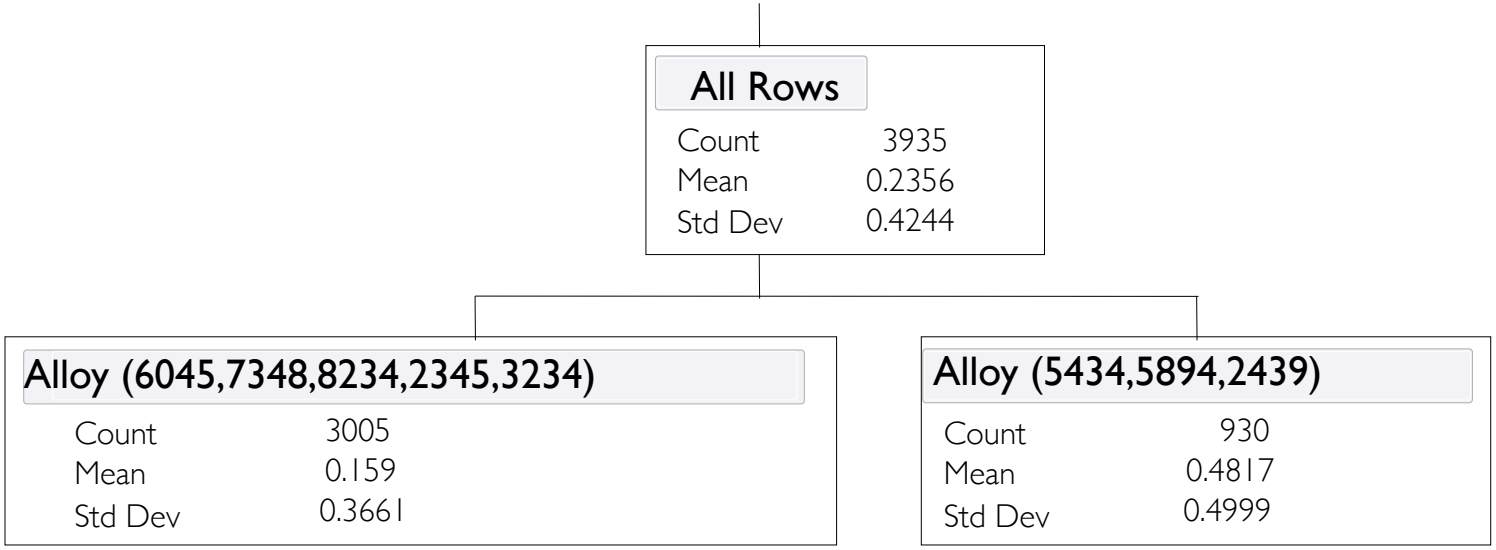
"Do you know we make 3 products?"

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

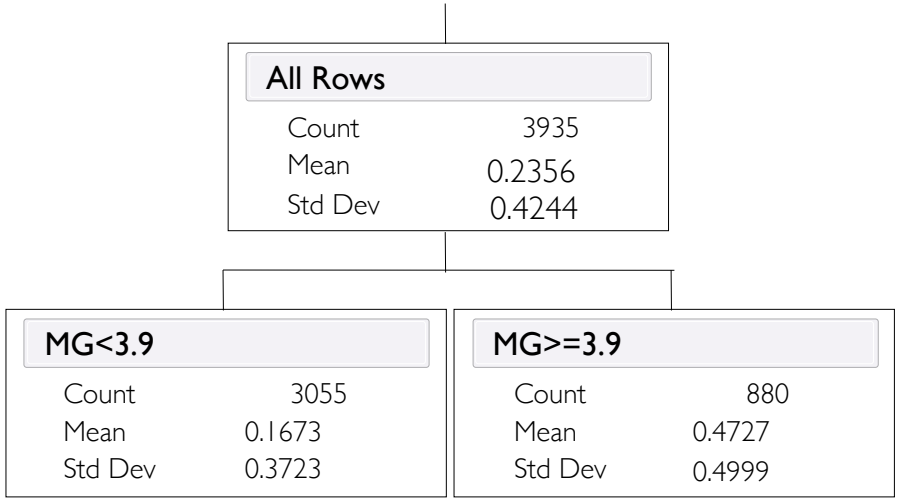
ALCOA

Ingots – First Tree



We know that!! – some alloys are hard to make. That's why we gave you the data in the first place.

Second Tree



What do you think is in those alloys?

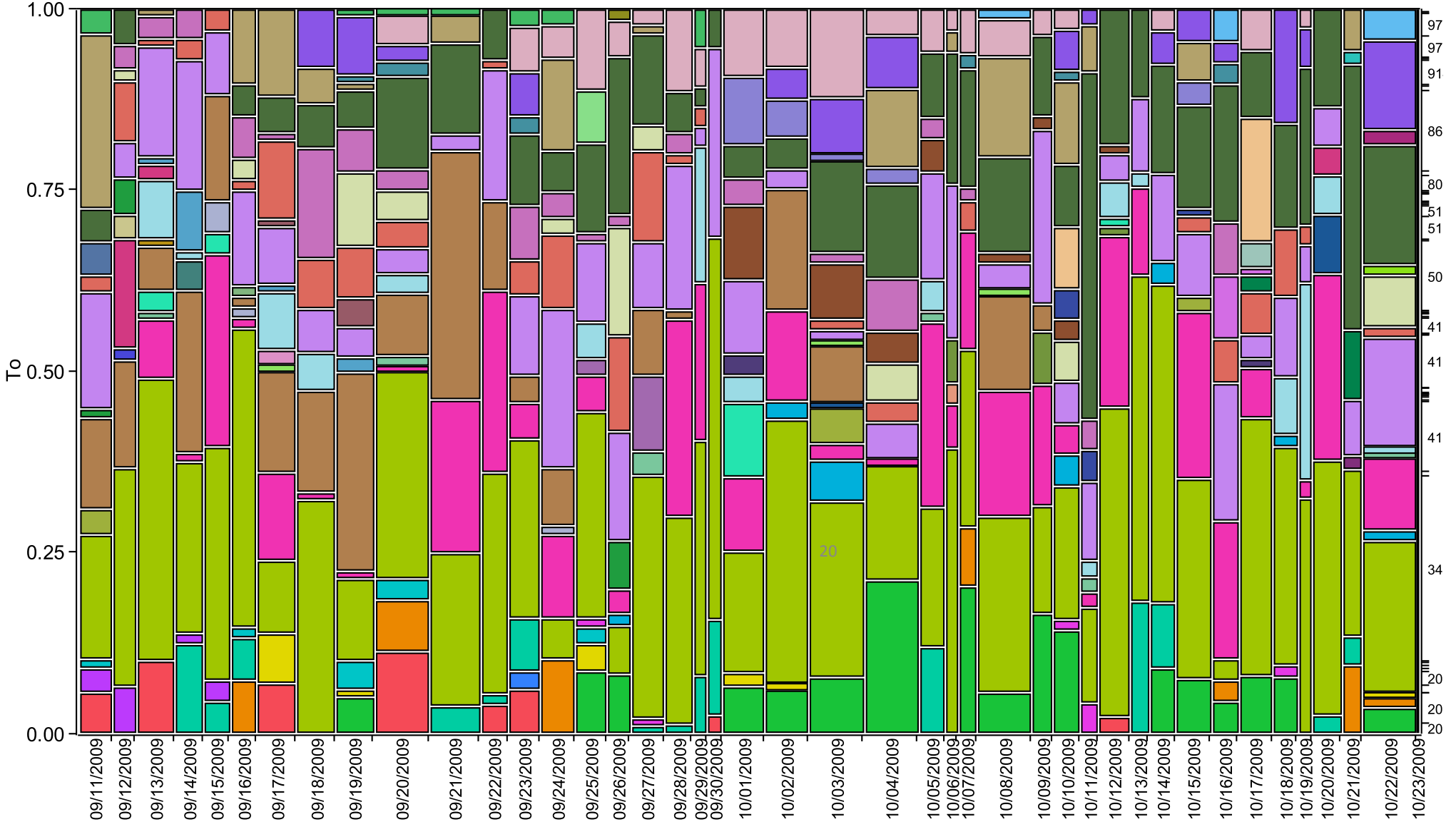
One More Time

- Looks like Chrome (!?)
- Really?
- Did that "solve" the problem?
  - No, but... experimental design
  - Enabled us to *focus* on the important variables

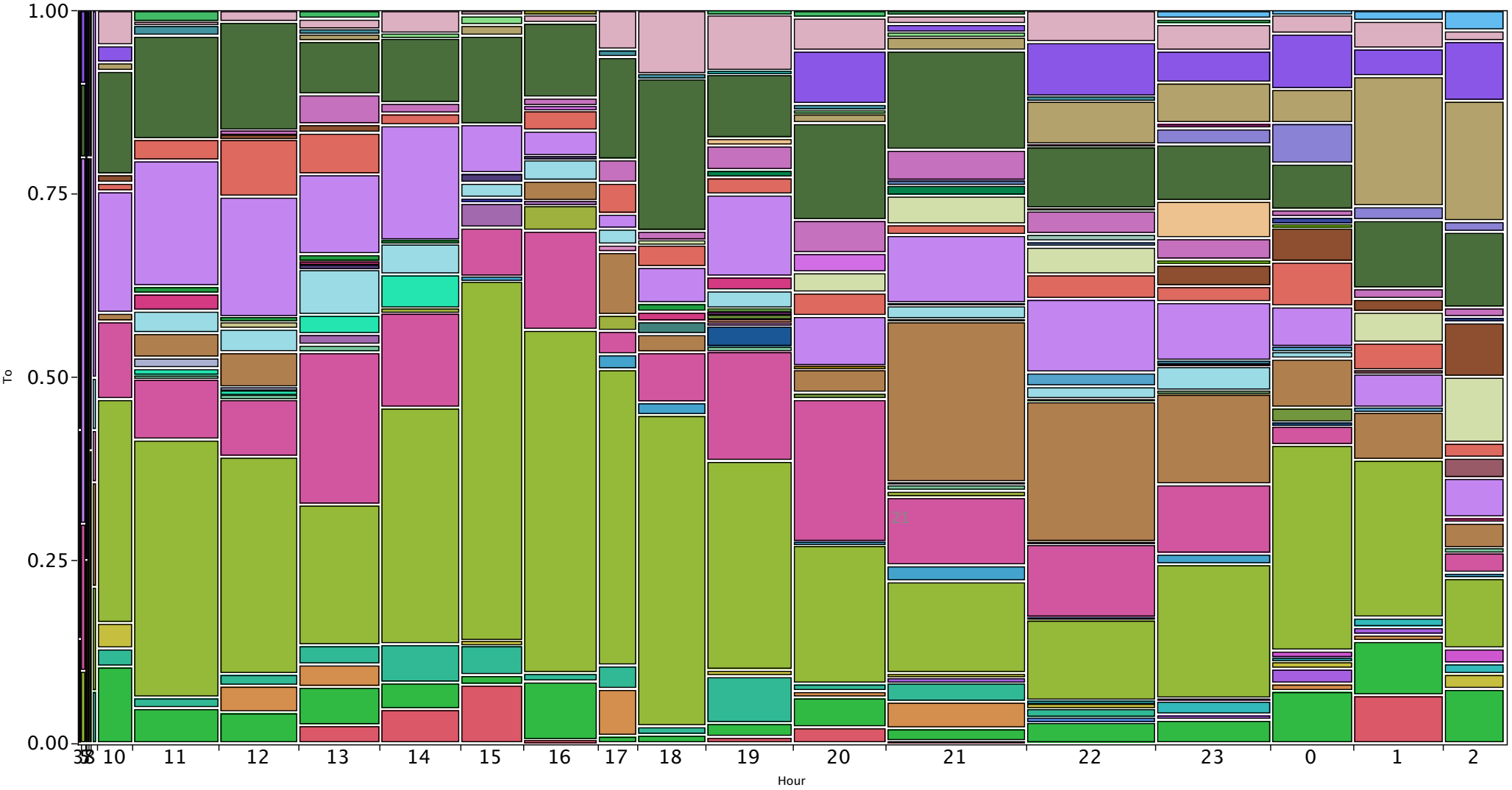
Oh, that's funny!

-Issac Asimov

Mosaic Plot


1.00
0.75
0.50
0.25
0.00
To
09/11/2009
09/12/2009
09/13/2009
09/14/2009
09/15/2009
09/16/2009
09/17/2009
09/18/2009
09/19/2009
09/20/2009
09/21/2009
09/22/2009
09/23/2009
09/24/2009
09/25/2009
09/26/2009
09/27/2009
09/28/2009
09/29/2009
09/30/2009
10/01/2009
10/02/2009
10/03/2009
10/04/2009
10/05/2009
10/06/2009
10/07/2009
10/08/2009
10/09/2009
10/10/2009
10/11/2009
10/12/2009
10/13/2009
10/14/2009
10/15/2009
10/16/2009
10/17/2009
10/18/2009
10/19/2009
10/20/2009
10/21/2009
10/22/2009
10/23/2009
20
97
97
91
86
80
51
51
50
41
41
41
34
20
20
20

By Hour…


1.00
0.75
0.50
0.25
0.00
To
10
11
12
13
14
15
16
17
18
19
20
21
22
23
0
1
2
Hour

2. Underestimating Data Preparation

Data Preparation

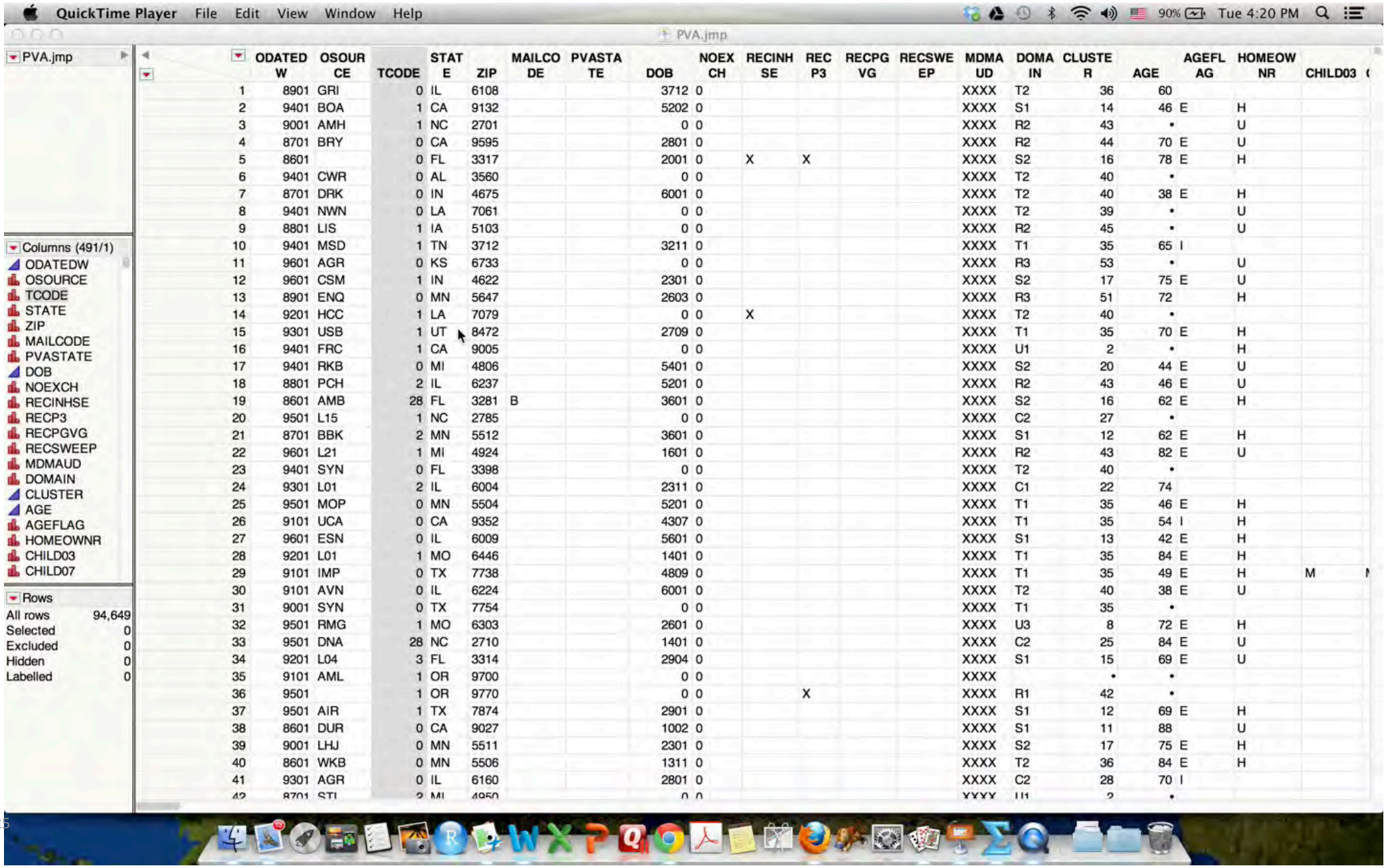
60% to 95% of the time is spent preparing the data

87.1% of all statistics are made up

Why is This Hard?

- 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)
  - 479 (!) demographic variables
    - Past giving, income, age etc etc etc
  - Recent campaign (only for training set)
    - Did they give? (Target B)
- Who should get the current mailing?
    - What's a cost effective strategy
    - How much did they give (Target D)

What's Hard Exactly?

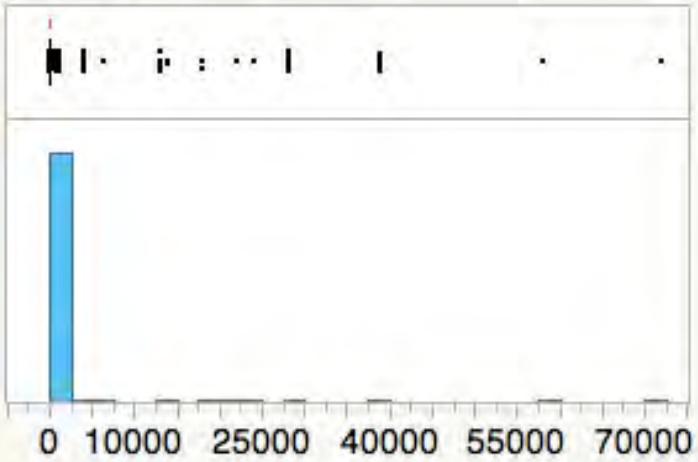


	ODATEDW	OSOURCE	TCODE	STATE	ZIP	MAILCODE	PVASTATE	DOB	NOEXCH	RECINHSE	RECP3	RECPGVG	RECSWEEP	MDMAUD	DOMAIN	CLUSTER	AGE	AGEFLAG	HOMEOWNR	CHILD03
1	8901	GRI	0	IL	6108			3712	0					XXXX	T2	36	60			
2	9401	BOA	1	CA	9132			5202	0					XXXX	S1	14	46	E	H	
3	9001	AMH	1	NC	2701			0	0					XXXX	R2	43	•		U	
4	8701	BRY	0	CA	9595			2801	0					XXXX	R2	44	70	E	U	
5	8601		0	FL	3317			2001	0	X	X			XXXX	S2	16	78	E	H	
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	
8	9401	NWN	0	LA	7061			0	0					XXXX	T2	39	•		U	
9	8801	LIS	1	IA	5103			0	0					XXXX	R2	45	•		U	
10	9401	MSD	1	TN	3712			3211	0					XXXX	T1	35	65	I		
11	9601	AGR	0	KS	6733			0	0					XXXX	R3	53	•		U	
12	9601	CSM	1	IN	4622			2301	0					XXXX	S2	17	75	E	U	
13	8901	ENQ	0	MN	5647			2603	0					XXXX	R3	51	72		H	
14	9201	HCC	1	LA	7079			0	0	X				XXXX	T2	40	•			
15	9301	USB	1	UT	8472			2709	0					XXXX	T1	35	70	E	H	
16	9401	FRC	1	CA	9005			0	0					XXXX	U1	2	•		H	
17	9401	RKB	0	MI	4806			5401	0					XXXX	S2	20	44	E	U	
18	8801	PCH	2	IL	6237			5201	0					XXXX	R2	43	46	E	U	
19	8601	AMB	28	FL	3281	B		3601	0					XXXX	S2	16	62	E	H	
20	9501	L15	1	NC	2785			0	0					XXXX	C2	27	•			
21	8701	BBK	2	MN	5512			3601	0					XXXX	S1	12	62	E	H	
22	9601	L21	1	MI	4924			1601	0					XXXX	R2	43	82	E	U	
23	9401	SYN	0	FL	3398			0	0					XXXX	T2	40	•			
24	9301	L01	2	IL	6004			2311	0					XXXX	C1	22	74			
25	9501	MOP	0	MN	5504			5201	0					XXXX	T1	35	46	E	H	
26	9101	UCA	0	CA	9352			4307	0					XXXX	T1	35	54	I	H	
27	9601	ESN	0	IL	6009			5601	0					XXXX	S1	13	42	E	H	
28	9201	L01	1	MO	6446			1401	0					XXXX	T1	35	84	E	H	
29	9101	IMP	0	TX	7738			4809	0					XXXX	T1	35	49	E	H	M
30	9101	AVN	0	IL	6224			6001	0					XXXX	T2	40	38	E	U	
31	9001	SYN	0	TX	7754			0	0					XXXX	T1	35	•			
32	9501	RMG	1	MO	6303			2601	0					XXXX	U3	8	72	E	H	
33	9501	DNA	28	NC	2710			1401	0					XXXX	C2	25	84	E	U	
34	9201	L04	3	FL	3314			2904	0					XXXX	S1	15	69	E	U	
35	9101	AML	1	OR	9700			0	0					XXXX		•	•			
36	9501		1	OR	9770			0	0		X			XXXX	R1	42	•			
37	9501	AIR	1	TX	7874			2901	0					XXXX	S1	12	69	E	H	
38	8601	DUR	0	CA	9027			1002	0					XXXX	S1	11	88		U	
39	9001	LHJ	0	MN	5511			2301	0					XXXX	S2	17	75	E	H	
40	8601	WKB	0	MN	5506			1311	0					XXXX	T2	36	84	E	H	
41	9301	AGR	0	IL	6160			2801	0					XXXX	C2	28	70	I		

T-Code

Distributions

TCODE



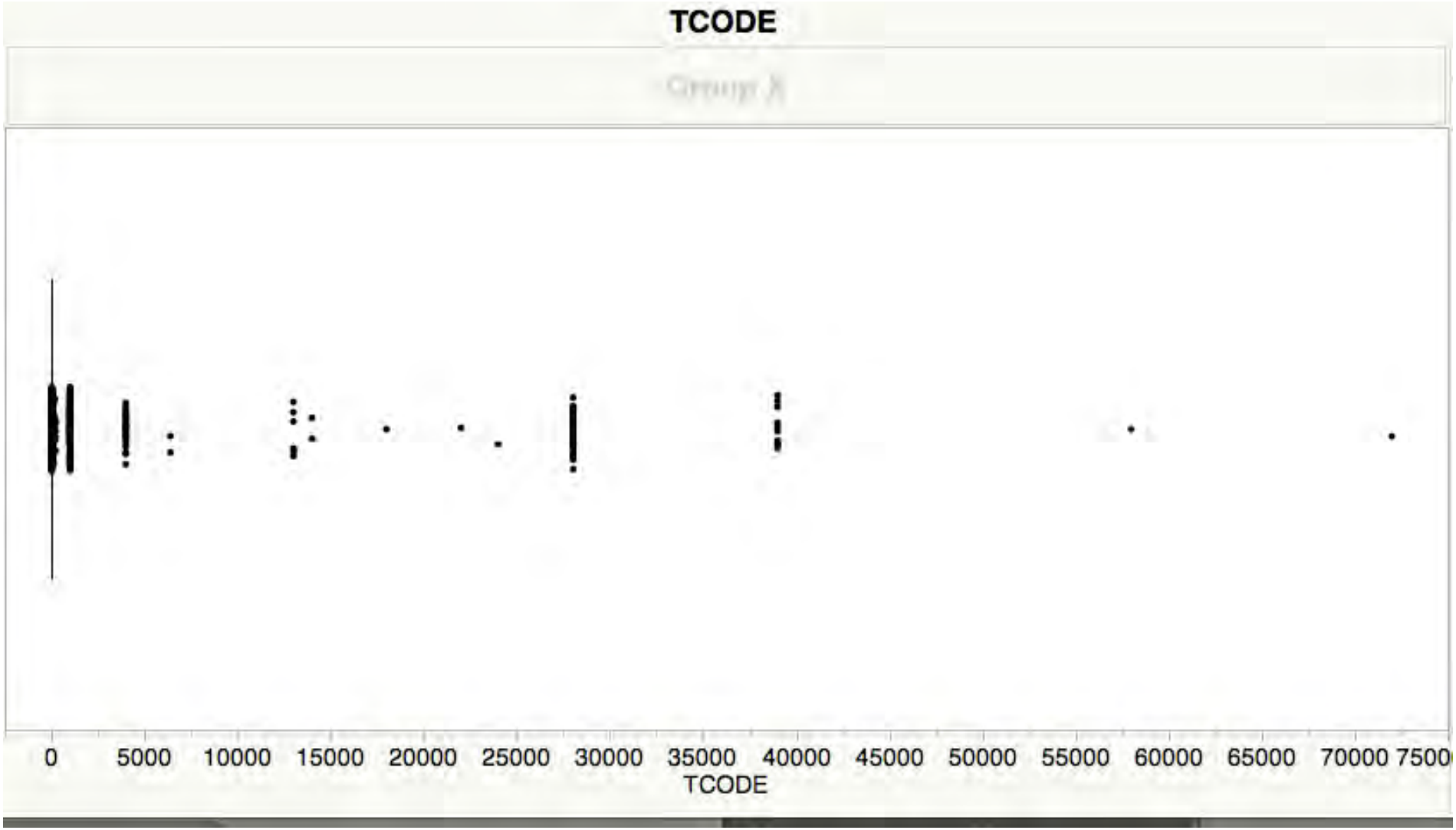
Quantiles

100.0%	maximum	72002
99.5%		1002
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

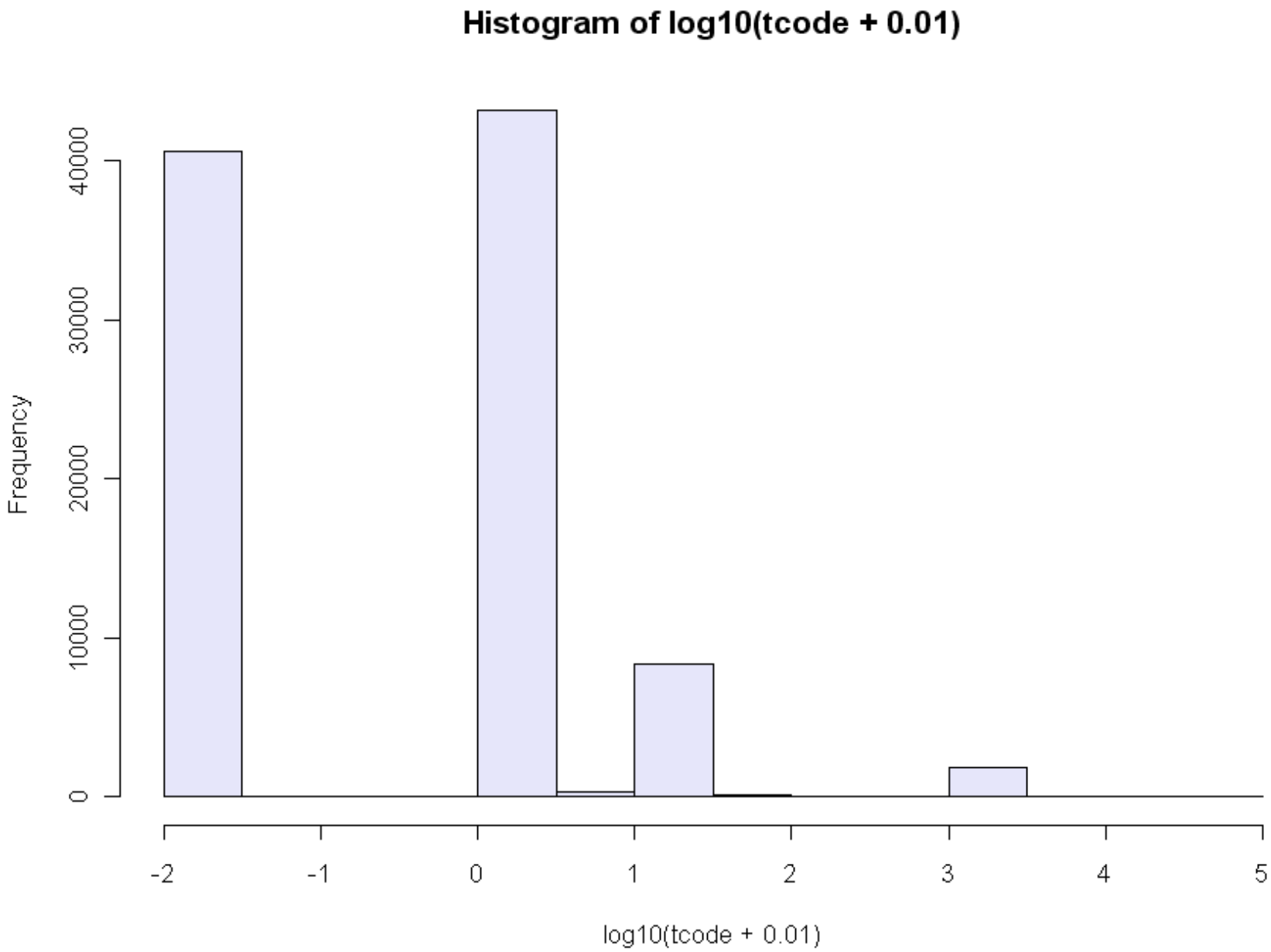
Summary Statistics

Mean	54.413105
Std Dev	957.50024
Std Err Mean	3.1122959
Upper 95% Mean	60.513171
Lower 95% Mean	48.313039
N	94649

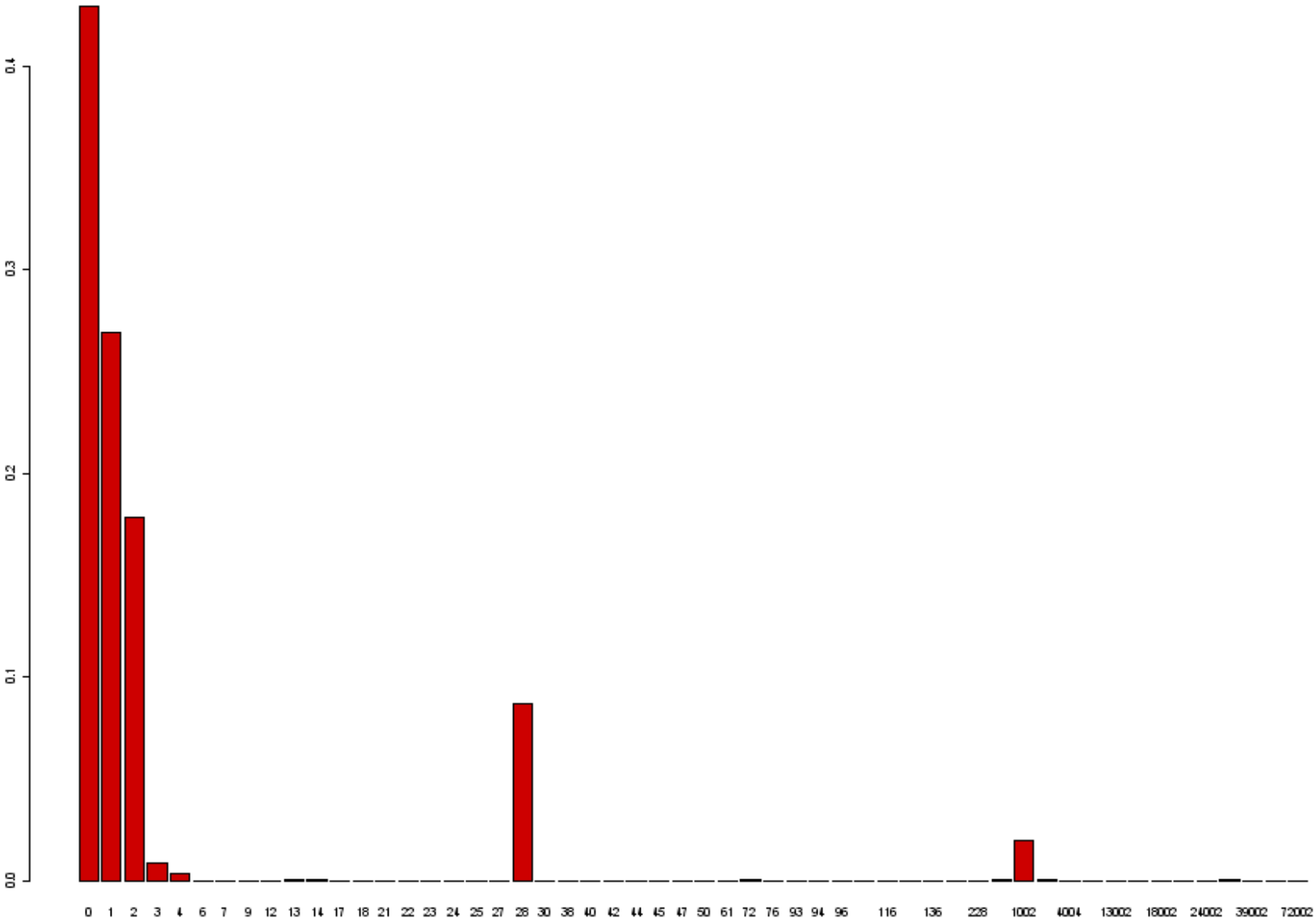
Zoom of Boxplot



Transformation?



Maybe Categories?



What is it, Anyway?

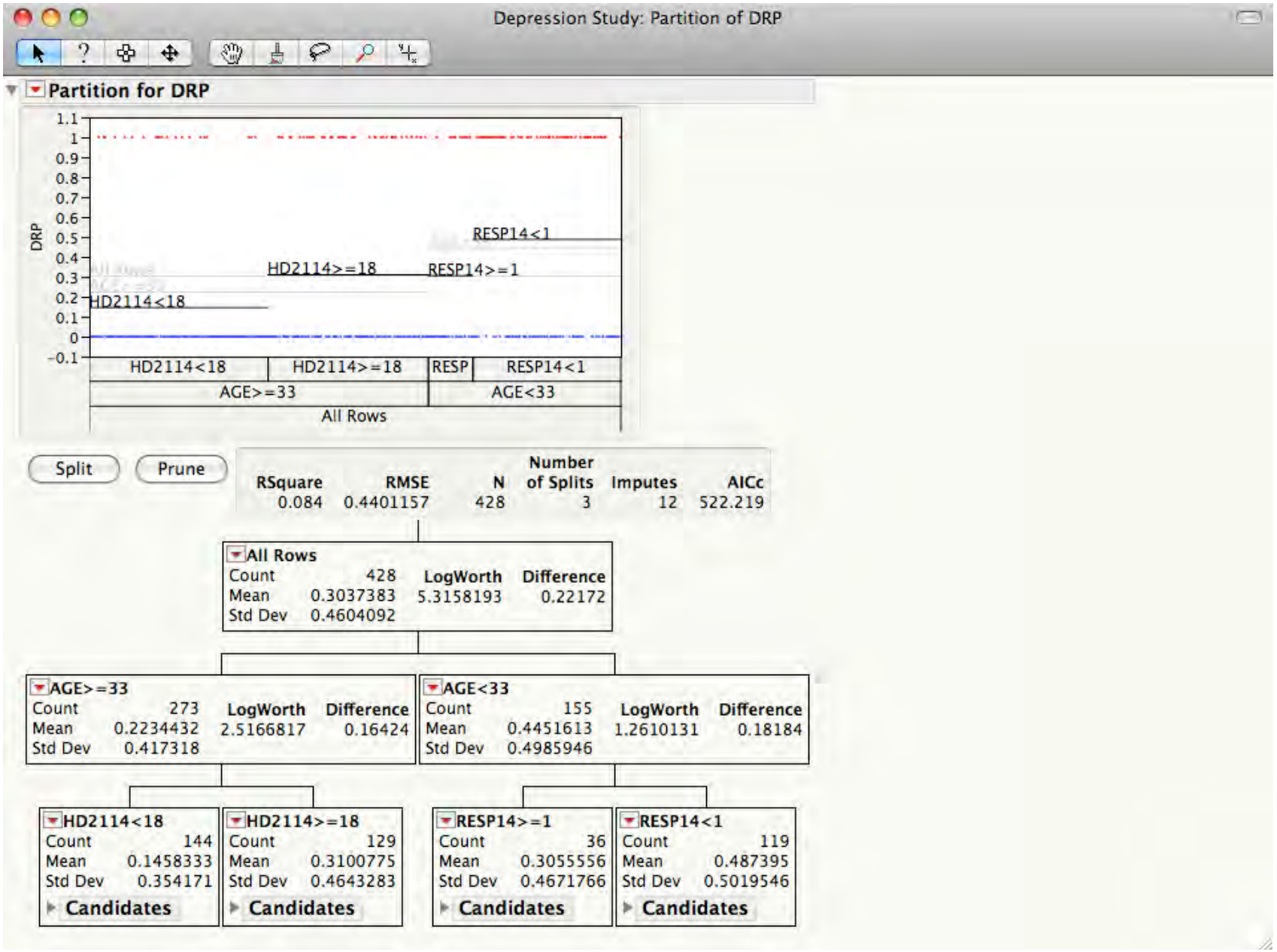
T-Code	Title
0	_
1	MR.
1001	MESSRS.
1002	MR. & MRS.
2	MRS.
2002	MESDAMES
3	MISS
3003	MISSES
4	DR.
4002	DR. & MRS.
4004	DOCTORS
5	MADAME
6	SERGEANT
9	RABBI
10	PROFESSOR
10002	PROFESSOR & MRS.
10010	PROFESSORS
11	ADMIRAL
11002	ADMIRAL & MRS.
12	GENERAL
12002	GENERAL & MRS.
13	COLONEL
13002	COLONEL & MRS.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

3. Ignoring What's Not There

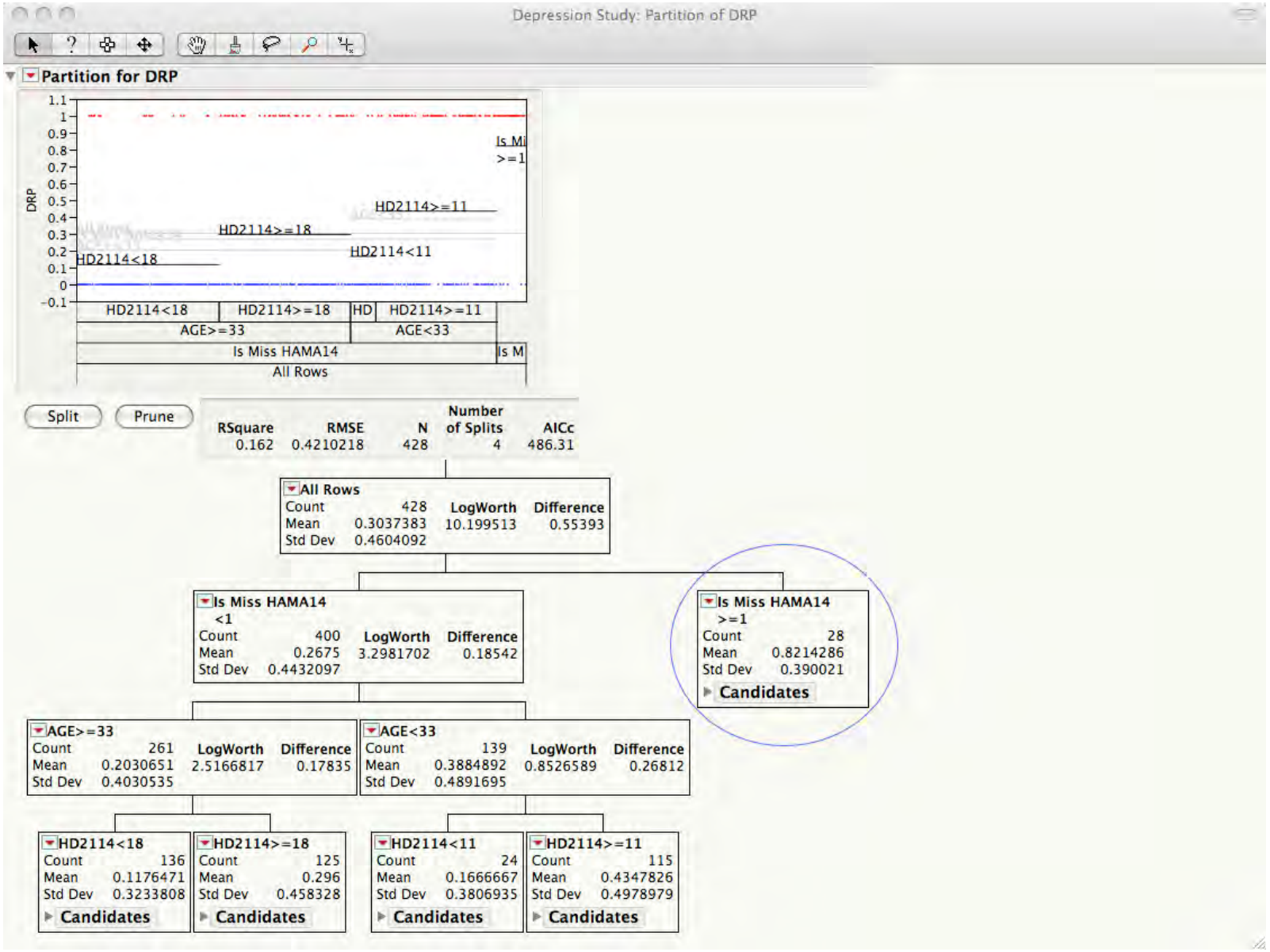
Depression Clinical Trial Study

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

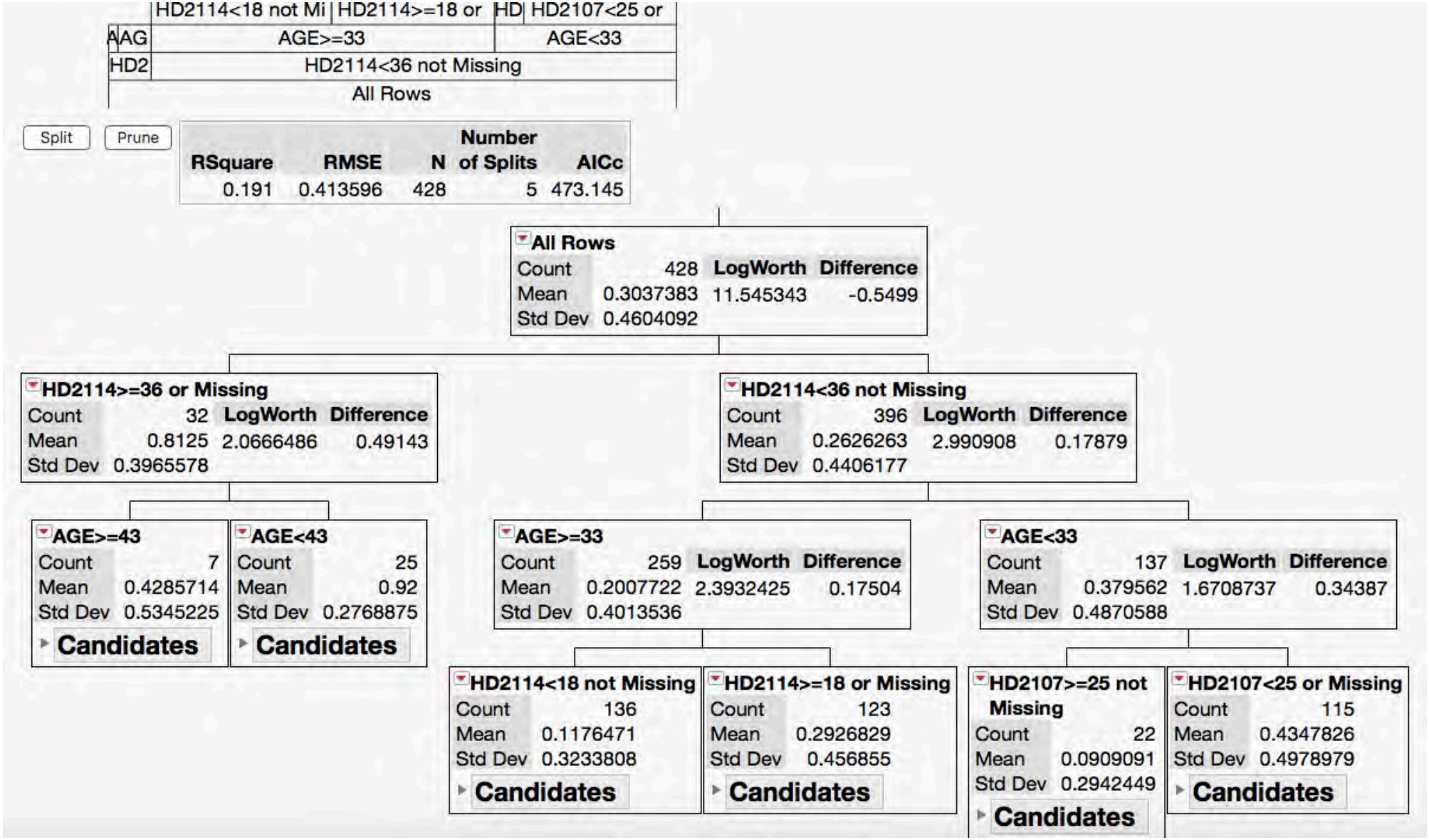
The Usual Suspects



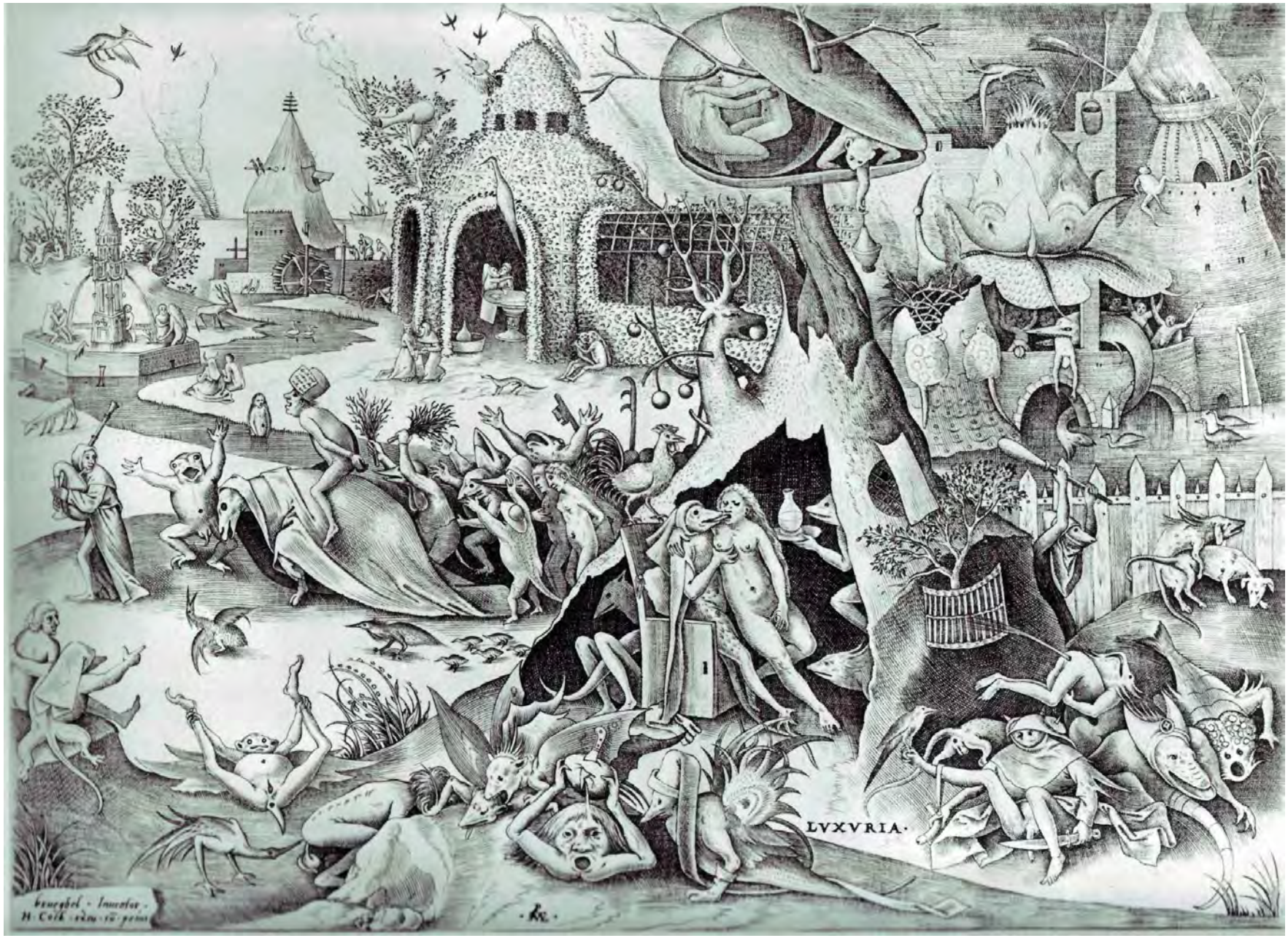
What's Missing?



Now Automatic



4. Falling in Love With Your Models



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
- Trees did even worse than radiologists

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

Random Forest Wins

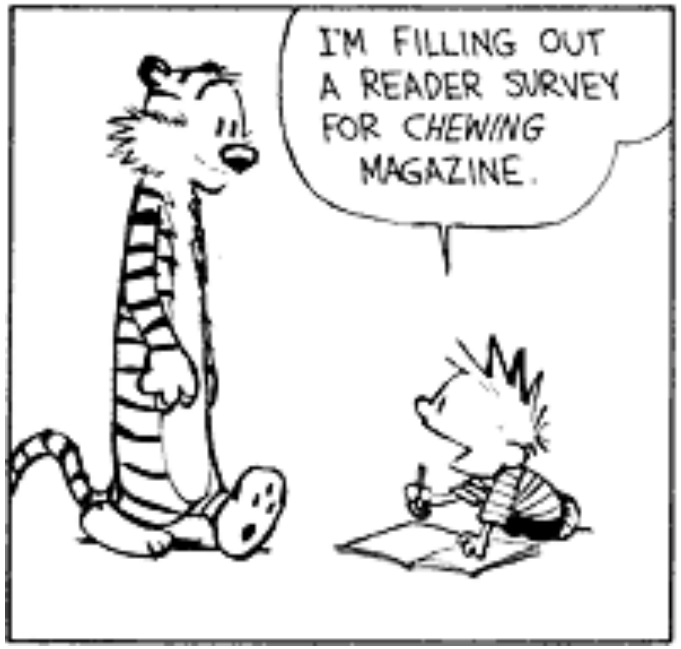
	False Positives	False Negatives
Simple Tree	32.20%	33.70%
Neural Network	25.50%	31.70%
Boosted Trees	24.90%	32.50%
Boostrap Forest	**19.30%**	**28.80%**
Radiologists	22.40%	35.80%

5. Using Bad Data

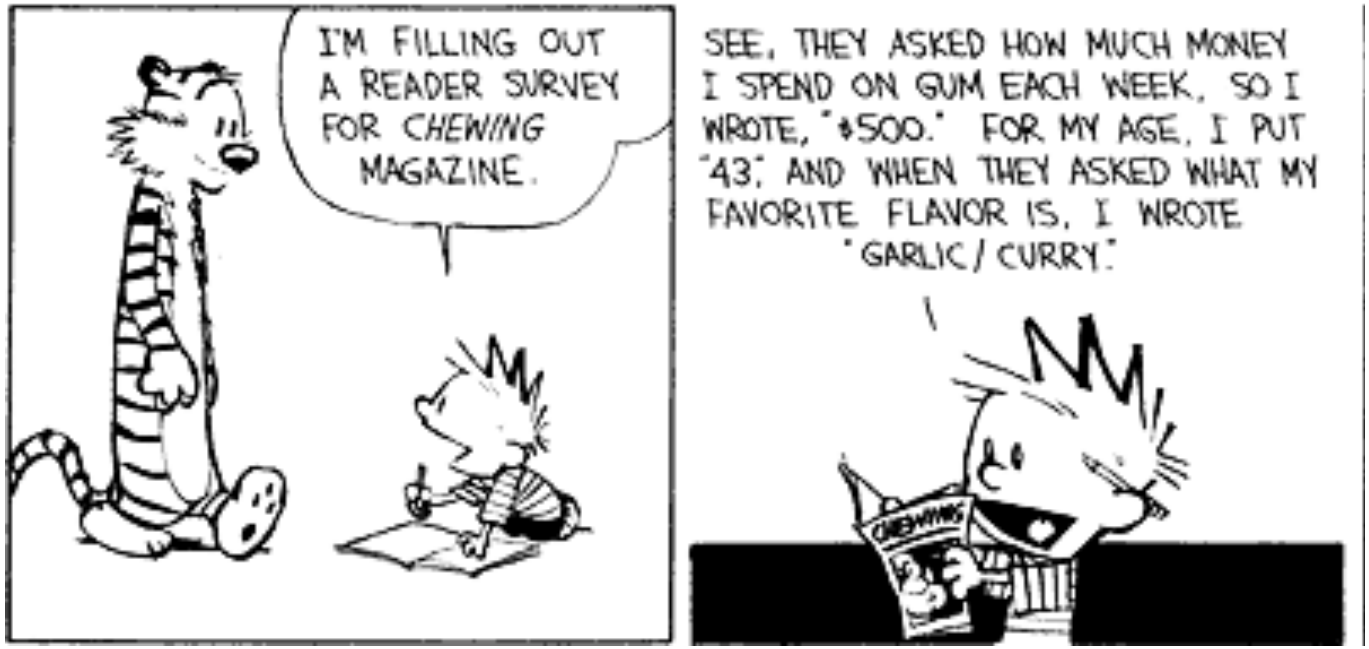
Where Do They Come From?

- ✦The *Times* of London reports 63.2kg 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

Data Quality

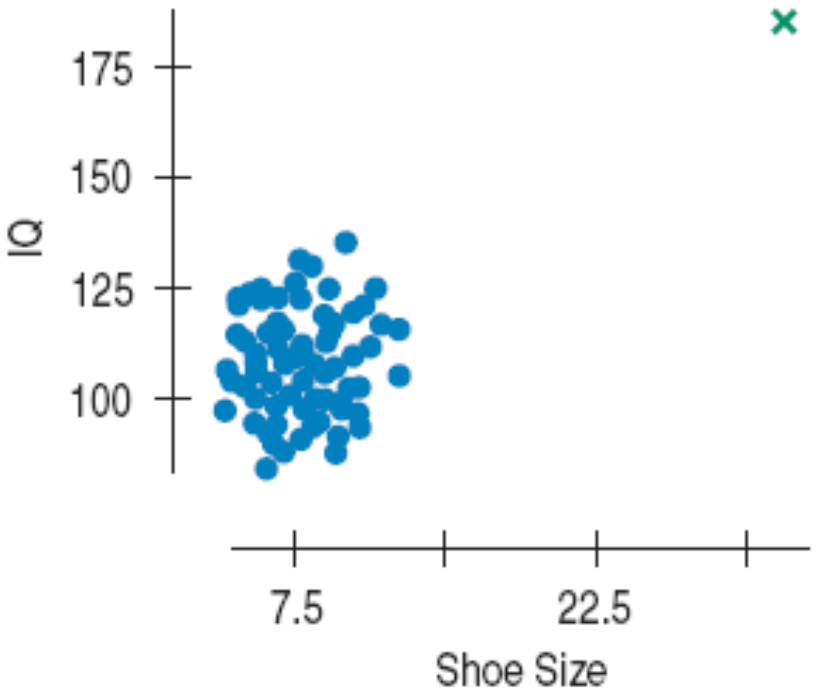


Data Quality



Data Quality

What Can They Do?



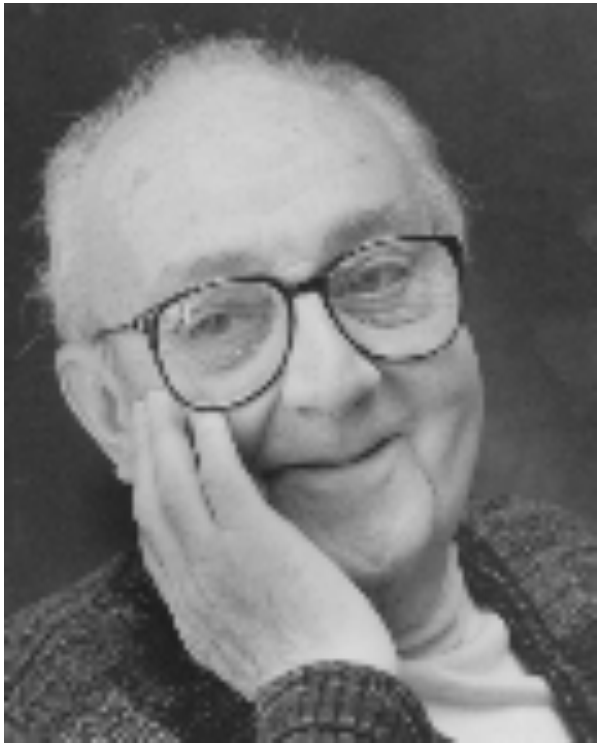
- Study by Large Credit Card Issuer
- Entire effect was due to one customer who charged $3M

6. Confusing Correlation and Causation

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 End of Science

Did you know?

- Shorter men have greater risk of heart attack

Did you know?

- Who is more likely to have a heart attack in 10 yrs?

Did you know?

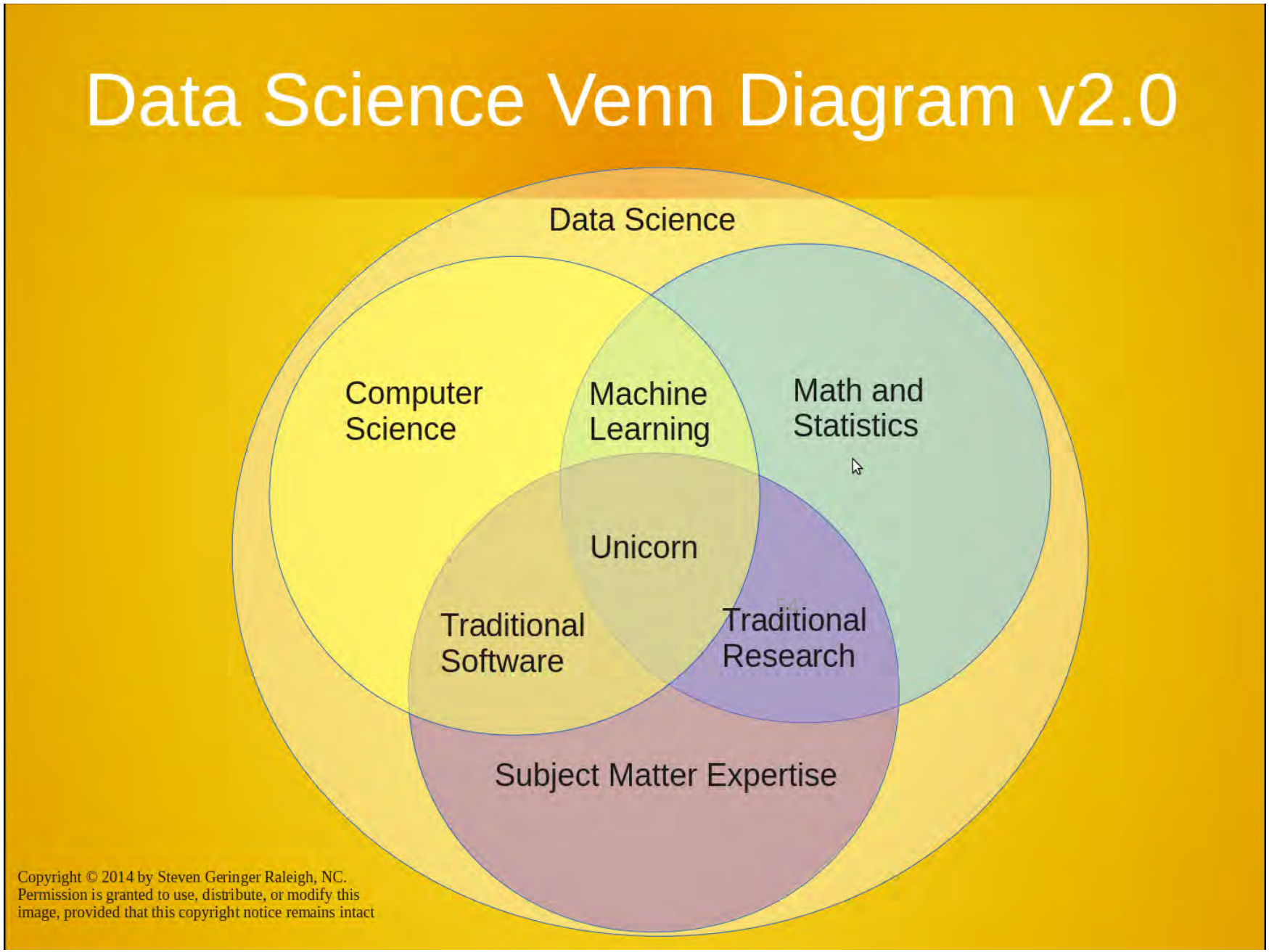
- High popsicle sales are strongly correlated with shark

Did you know?

- People who believe in alien abductions prefer Pepsi to Coke

7. Not Taking Your Anti-Hubristines

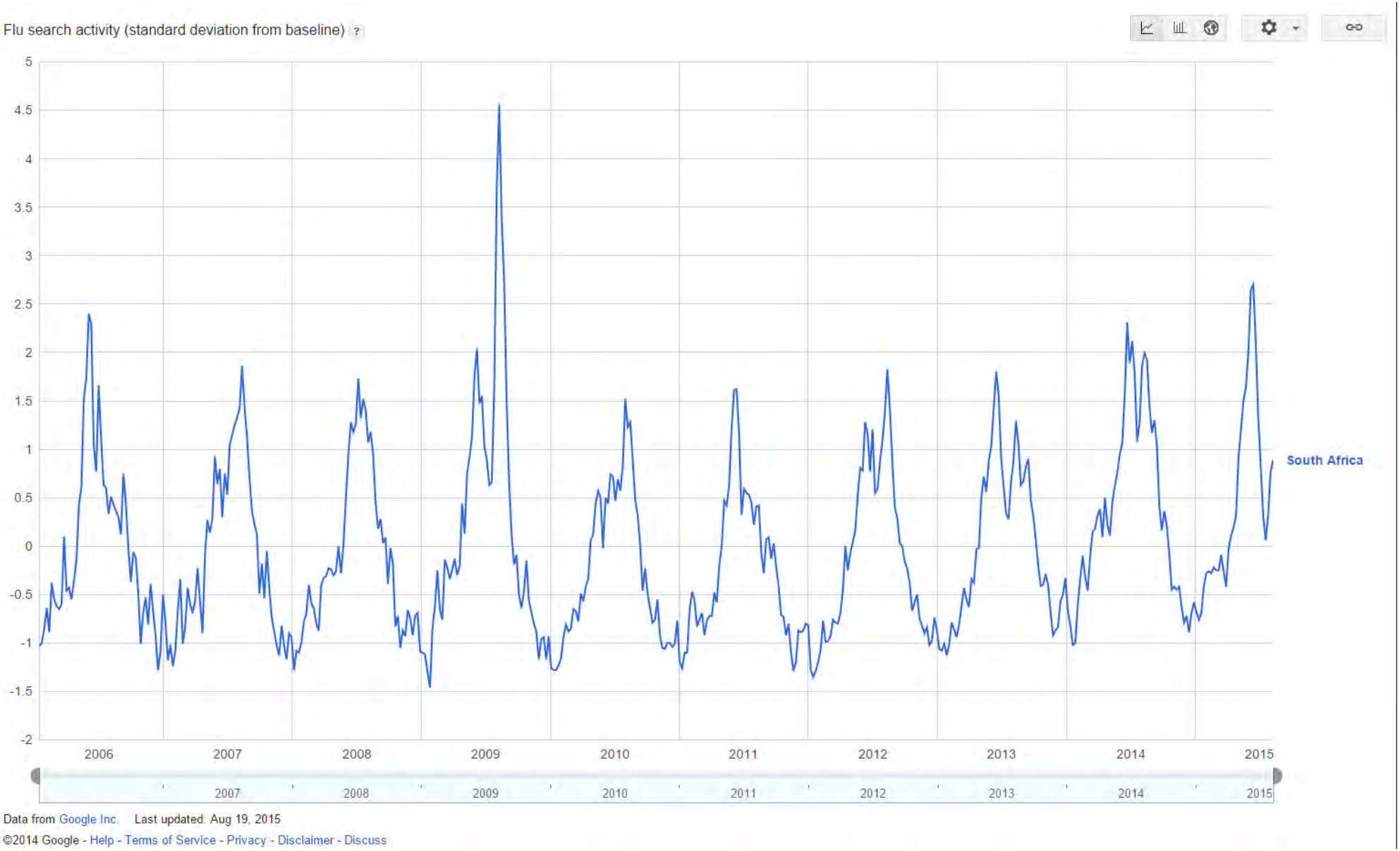
Data Science



Or?


Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil

Google Predicts Flu Before CDC!!



Or… Did They?

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

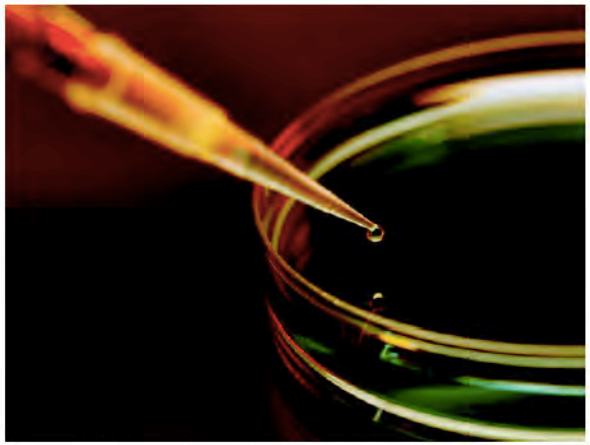
David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

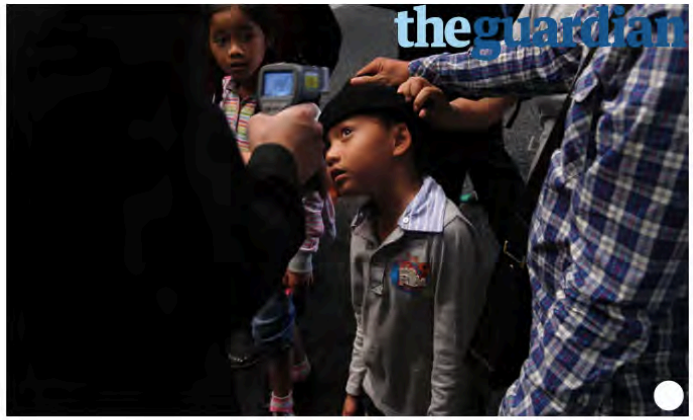
In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS





Google Flu Trends is no longer good at predicting flu, scientists find

The Seven Virtues

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

Thank you!!