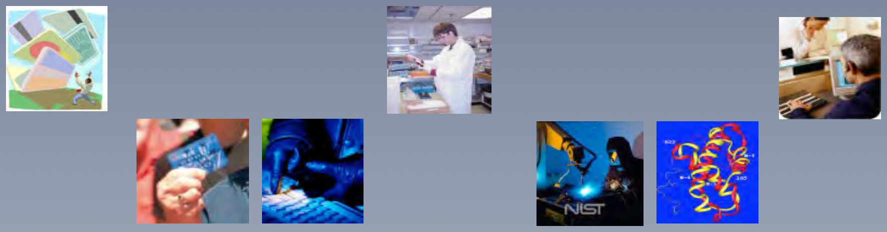


SUCCESSFUL DATA MINING IN PRACTICE

ASA SHORT COURSE

Richard D. De Veaux
Williams College
JSM Baltimore 2017
deveaux@williams.edu

WILLIAMS COLLEGE

WHY DATA MINING?

Data = $$

WHAT IS DATA MINING?

WHAT IS DATA MINING?

Data Mining

Where else would you go?

DATA MINING DEFINITIONS

data mining
noun *Digital Technology*.
the process of collecting, searching through, and analyzing a large amount of data in a database, as to discover patterns or relationships: *the use of data mining to detect fraud.*

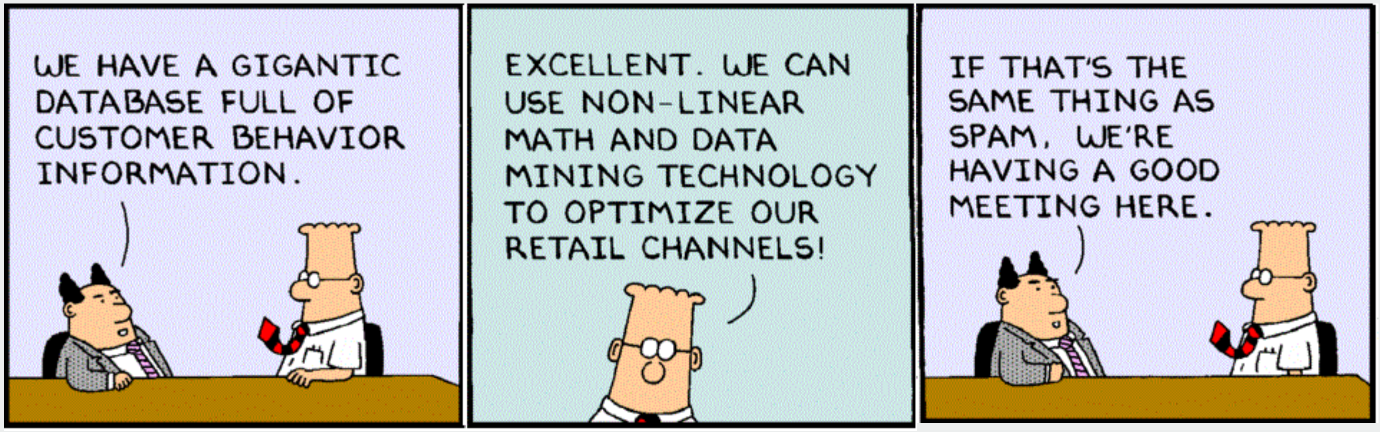
Origin:
1985-90

data mining
is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems

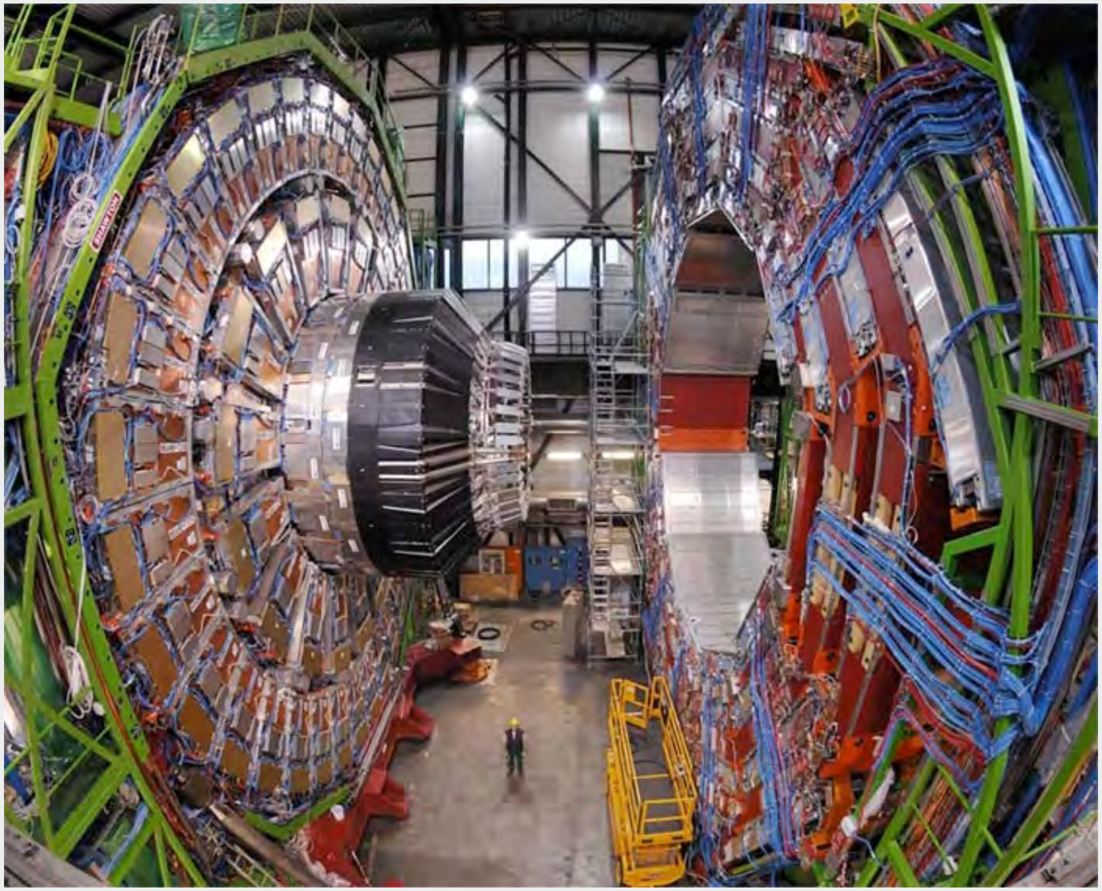
data mining
is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both

data mining
the practice of searching through large amounts of computerized data to find useful patterns or trends. A major goal of data mining is to discover previously unknown relationships among the data

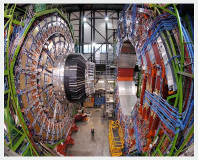
WHAT IS IT REALLY?

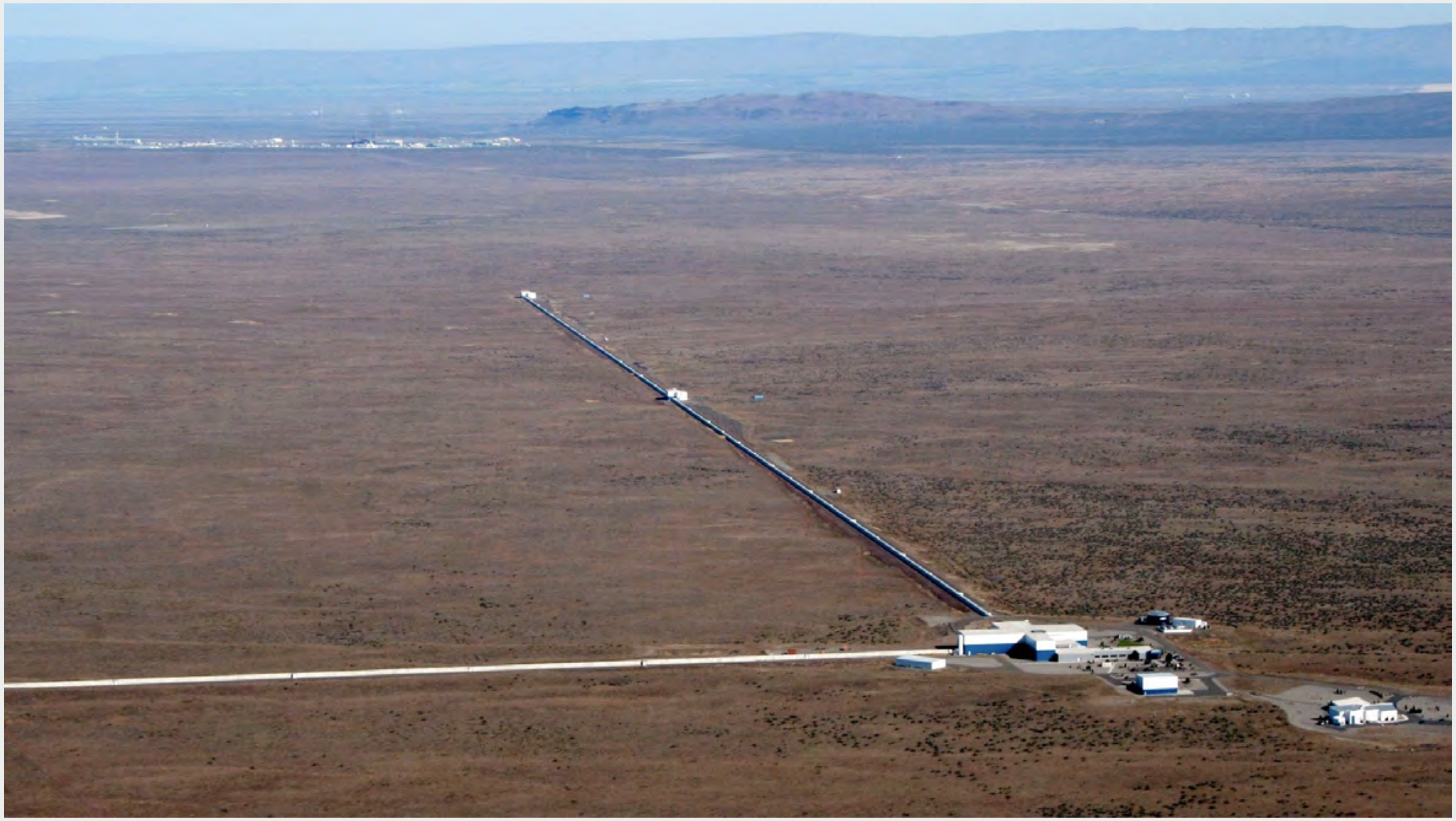


BIG DATA?

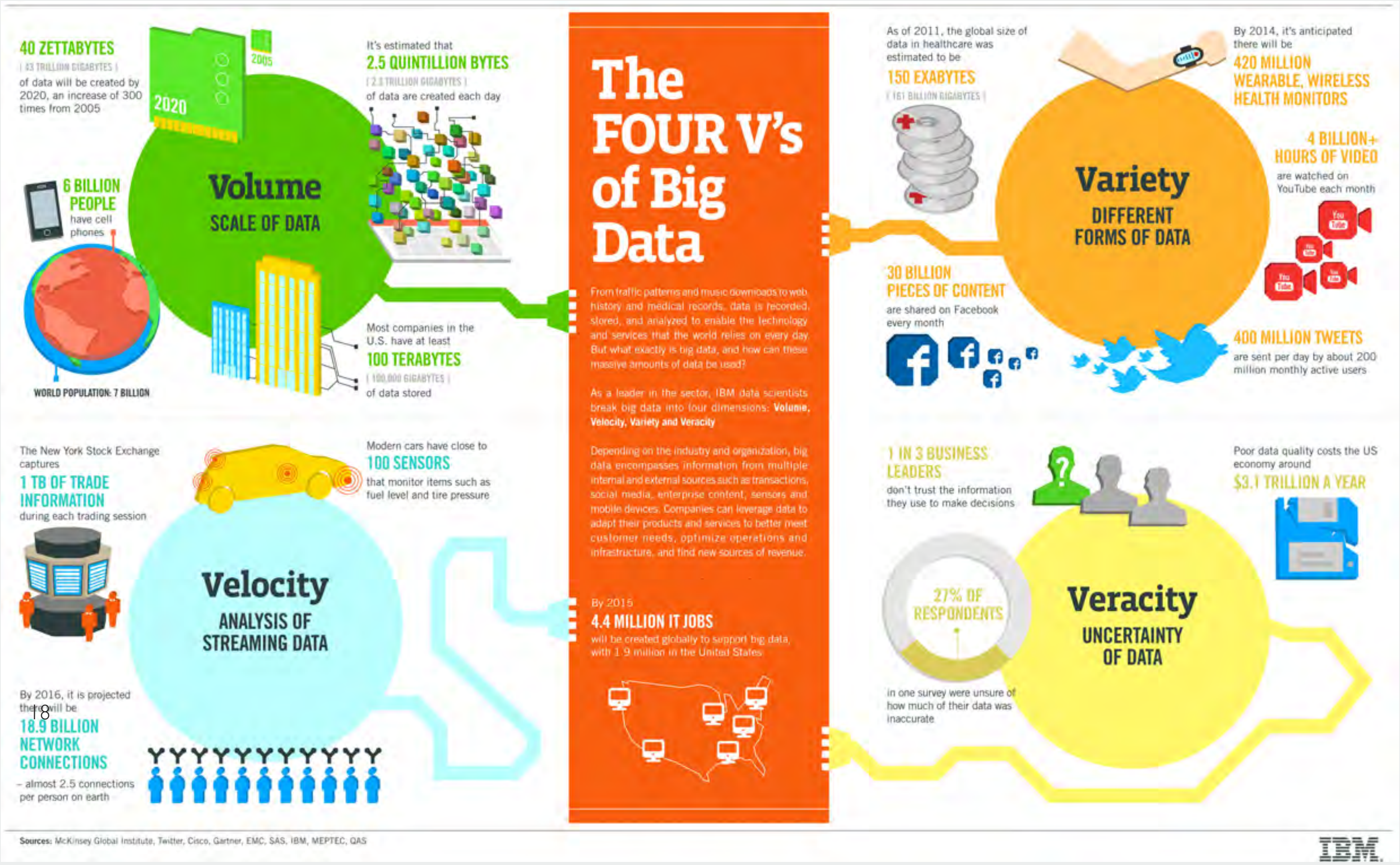


HOW BIG IS BIG?





WHAT IS BIG DATA ALL ABOUT?



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

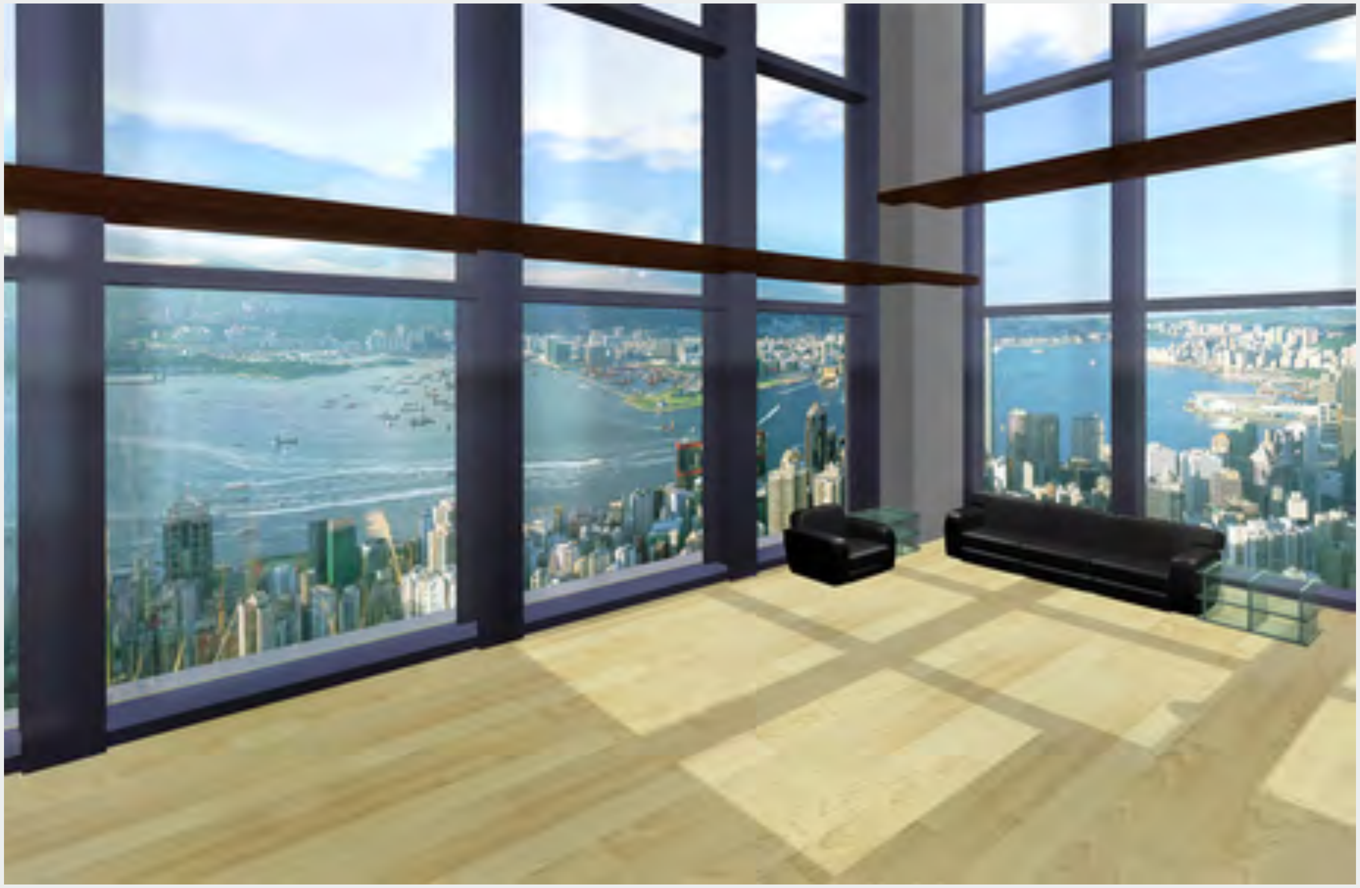
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

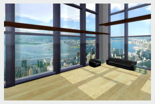
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

WHERE IS DATA MINING USED?

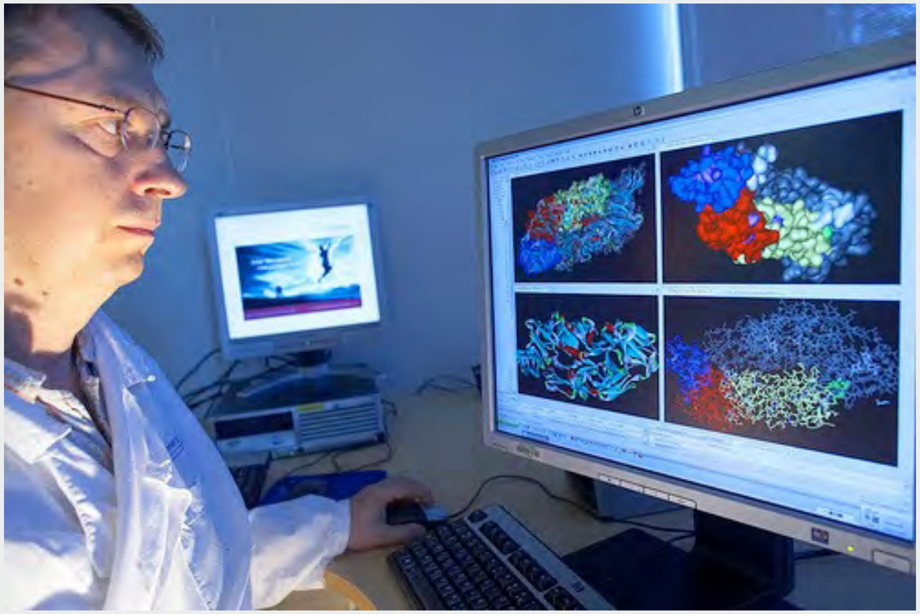
WHERE IS DATA MINING USED?



WHERE IS DATA MINING USED?



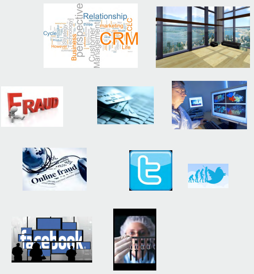
WHERE IS DATA MINING USED?



WHERE IS DATA MINING USED?

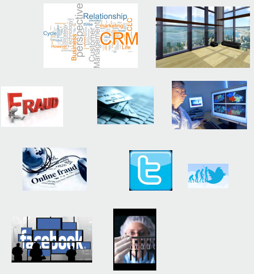
WHERE IS DATA MINING USED?

WHERE IS DATA MINING USED?



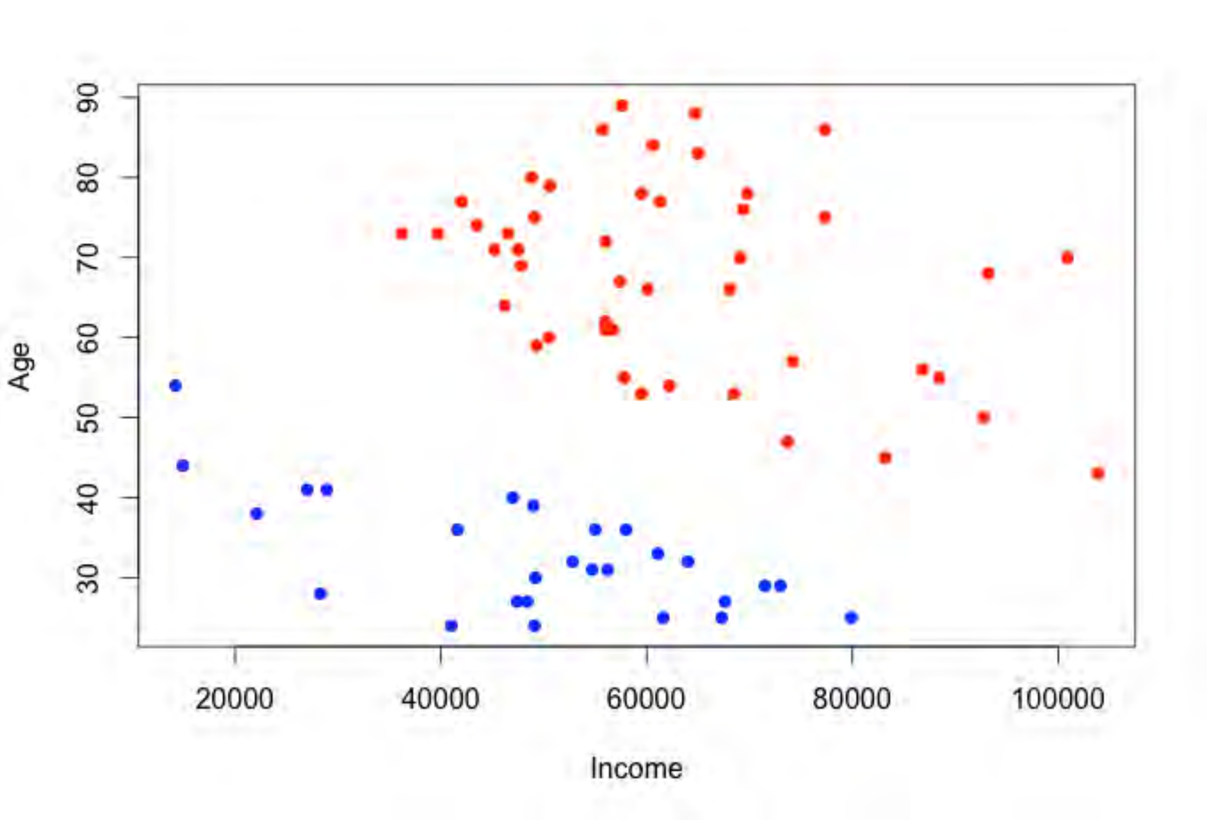
Don’t mix any 0’s in with the 1’s

WHERE IS DATA MINING USED?

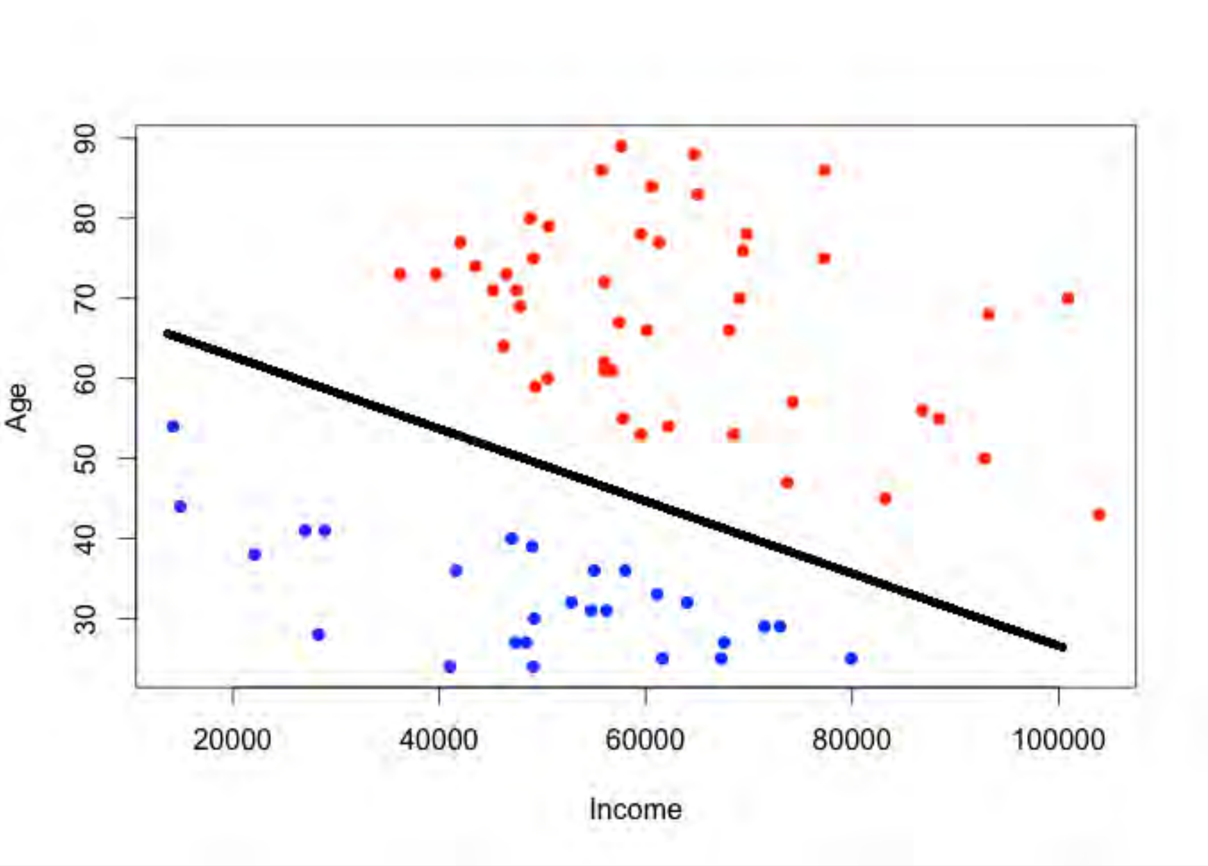


Don’t mix any 0’s in with the 1’s
and don’t miss *any* 1’s !

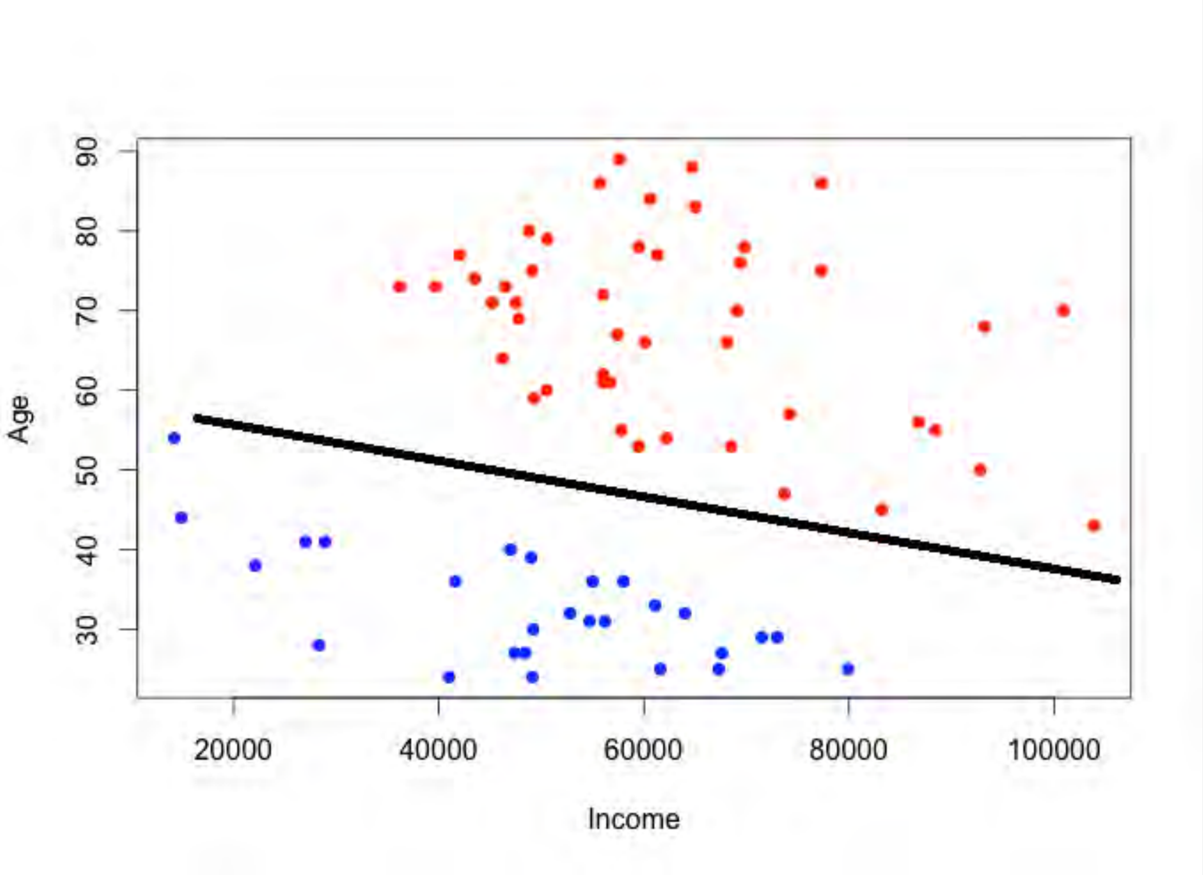
CONCEPTUAL CHALLENGES



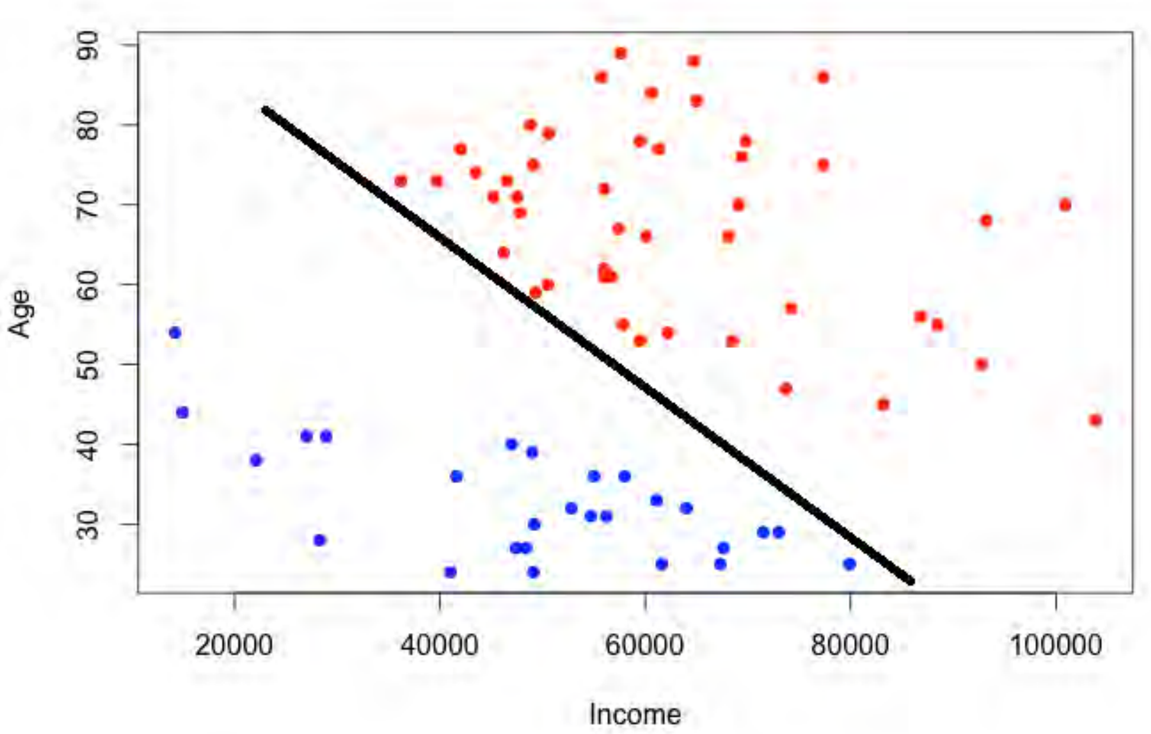
CONCEPTUAL CHALLENGES



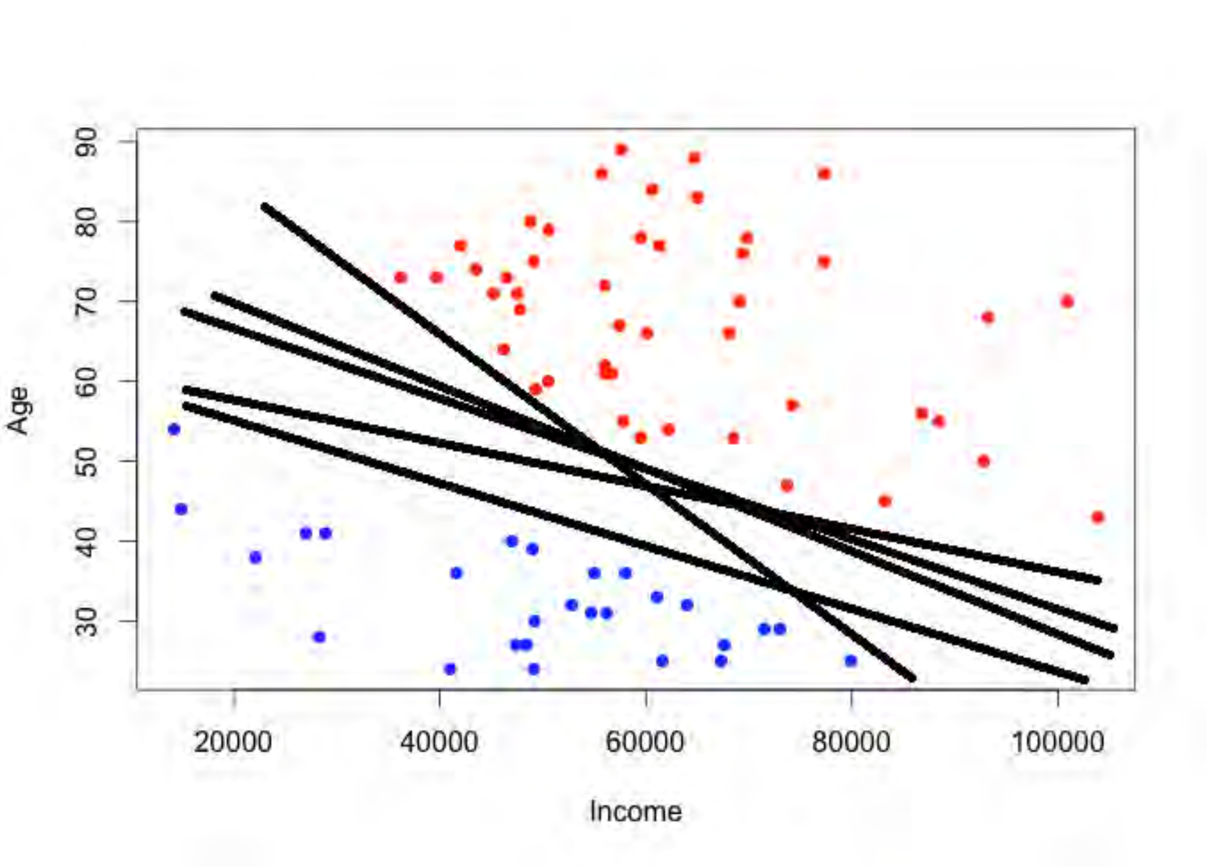
CONCEPTUAL CHALLENGES



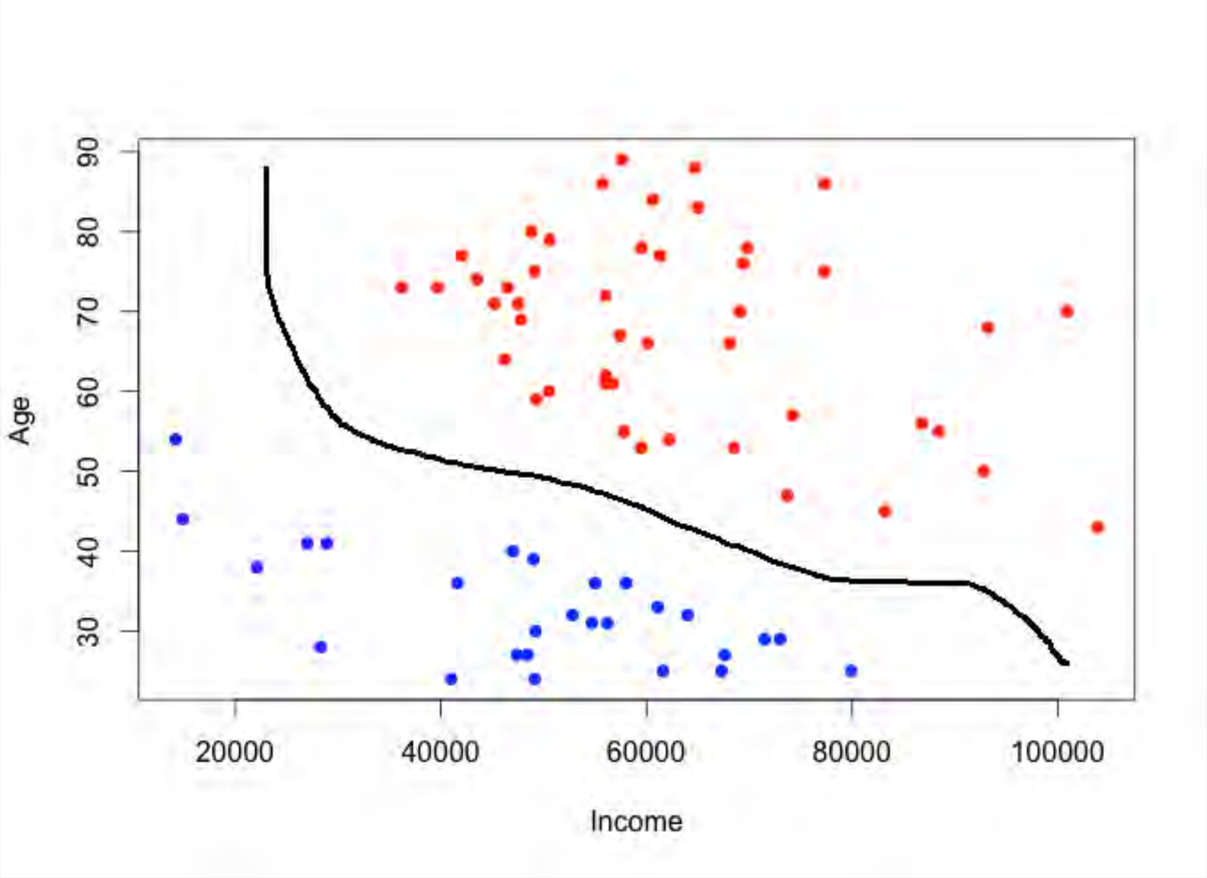
CONCEPTUAL CHALLENGES



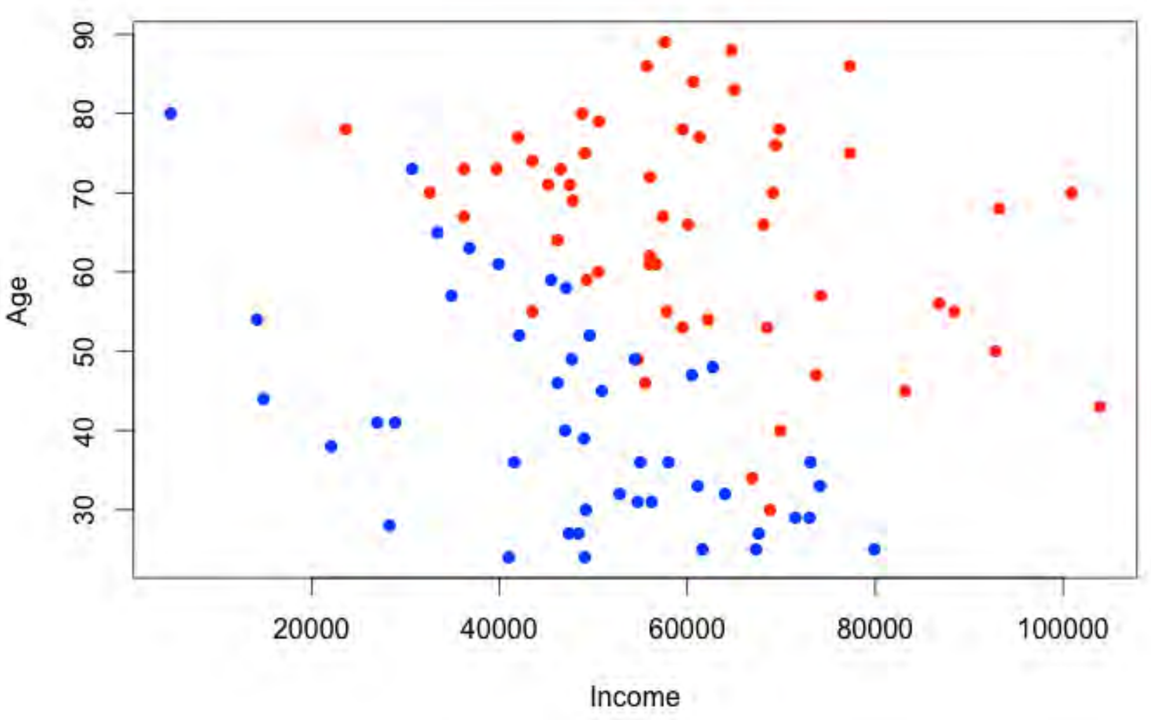
CONCEPTUAL CHALLENGES



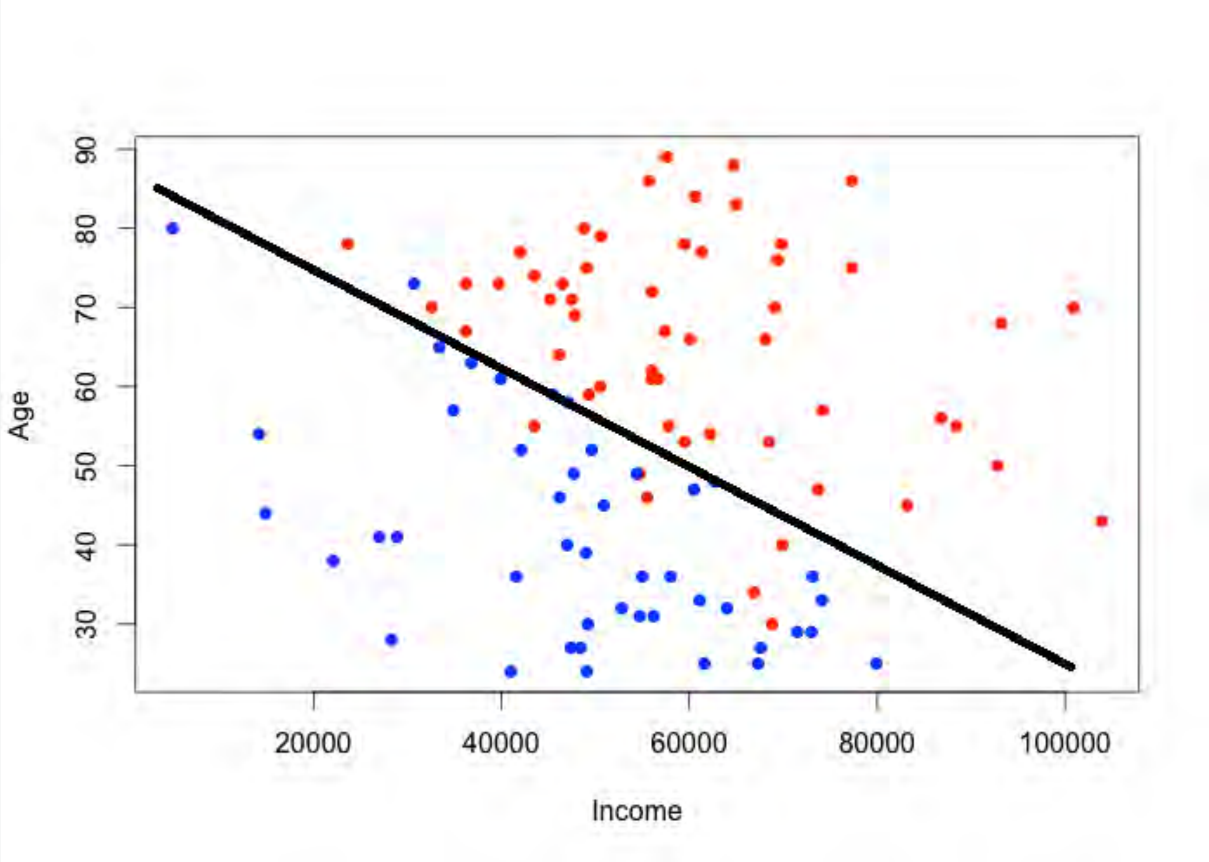
CONCEPTUAL CHALLENGES



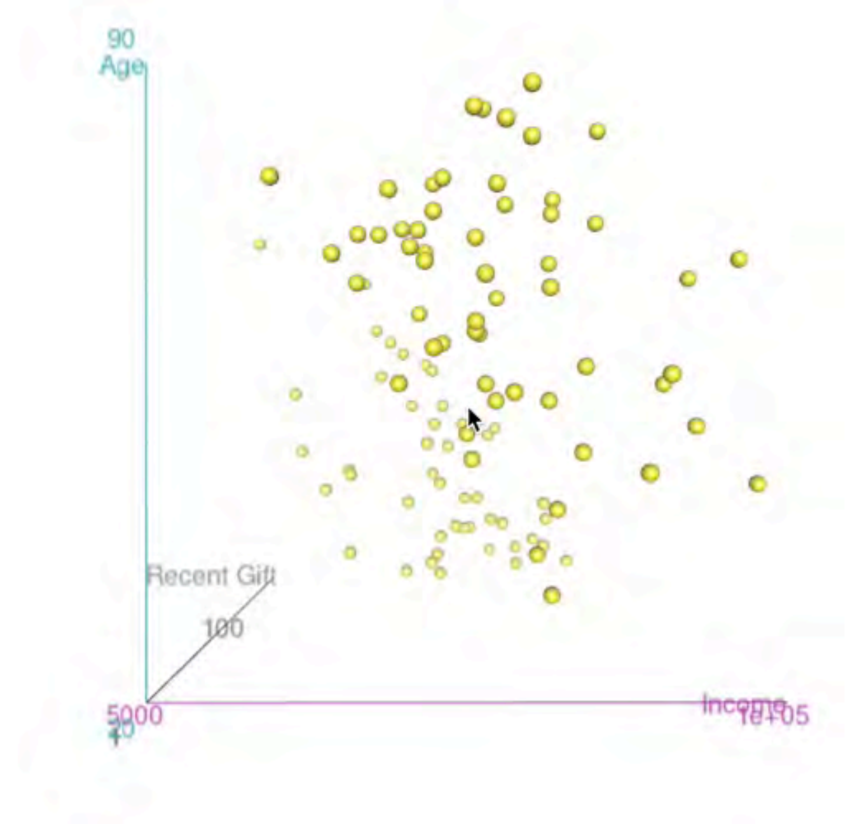
CONCEPTUAL CHALLENGES



CONCEPTUAL CHALLENGES



ANOTHER VARIABLE?



MODELS – A PARTIAL LIST

- **Traditional Statistical Methods and Models**
  - Linear regression, logistic regression, splines, smoothers etc
  - Not originally part of the Data Mining suite
- **Visualization**
- **Generalized Regression**
  - Change the fitting criterion
- **Classification and Regression Trees**
- **Dimension Reduction**
  - Principal Components
  - Projection Pursuit

MODELS – A PARTIAL LIST

- **Neural Networks**
- **K-nearest Neighbor Methods**
- **Naïve Bayes**
- **Combining Models**
  - Bagging (Random Forests)
  - Boosting
- **Clustering (unsupervised learning)**

George Box

All models are wrong
but...some are useful

MODELS

All models are wrong

but...some are useful

George Box

Statisticians, like artists, have the bad habit of falling in love with their models

MODELS

All models are wrong

but...some are useful

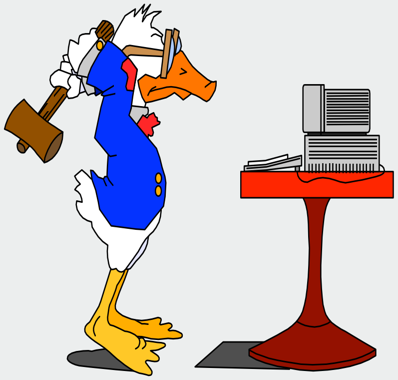
George Box

Statisticians, like artists, have the bad habit of falling in love with their models

WHAT'S DIFFERENT?

- **Massive Amounts of Data**
  - Number of rows (cases)
  - Number of columns (variables)
- **Low Signal to Noise**
  - Many irrelevant variables — but which ones?
  - Subtle relationships
  - Variation and data quality
- **UPS**
  - 16TB — Mostly tracking
- **Facebook**
  - 200-300TB photos/month
- **Google**
  - 1 PB every 72 minutes (!?!)

USERS ARE DIFFERENT



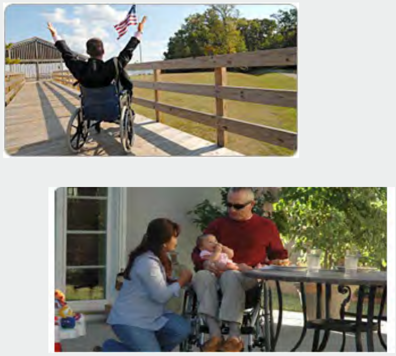
- Users
- Domain experts, not statisticians
- Have too much data
- Want automatic methods
- Want useful information without spending all their time doing statistical analysis

MYTHS

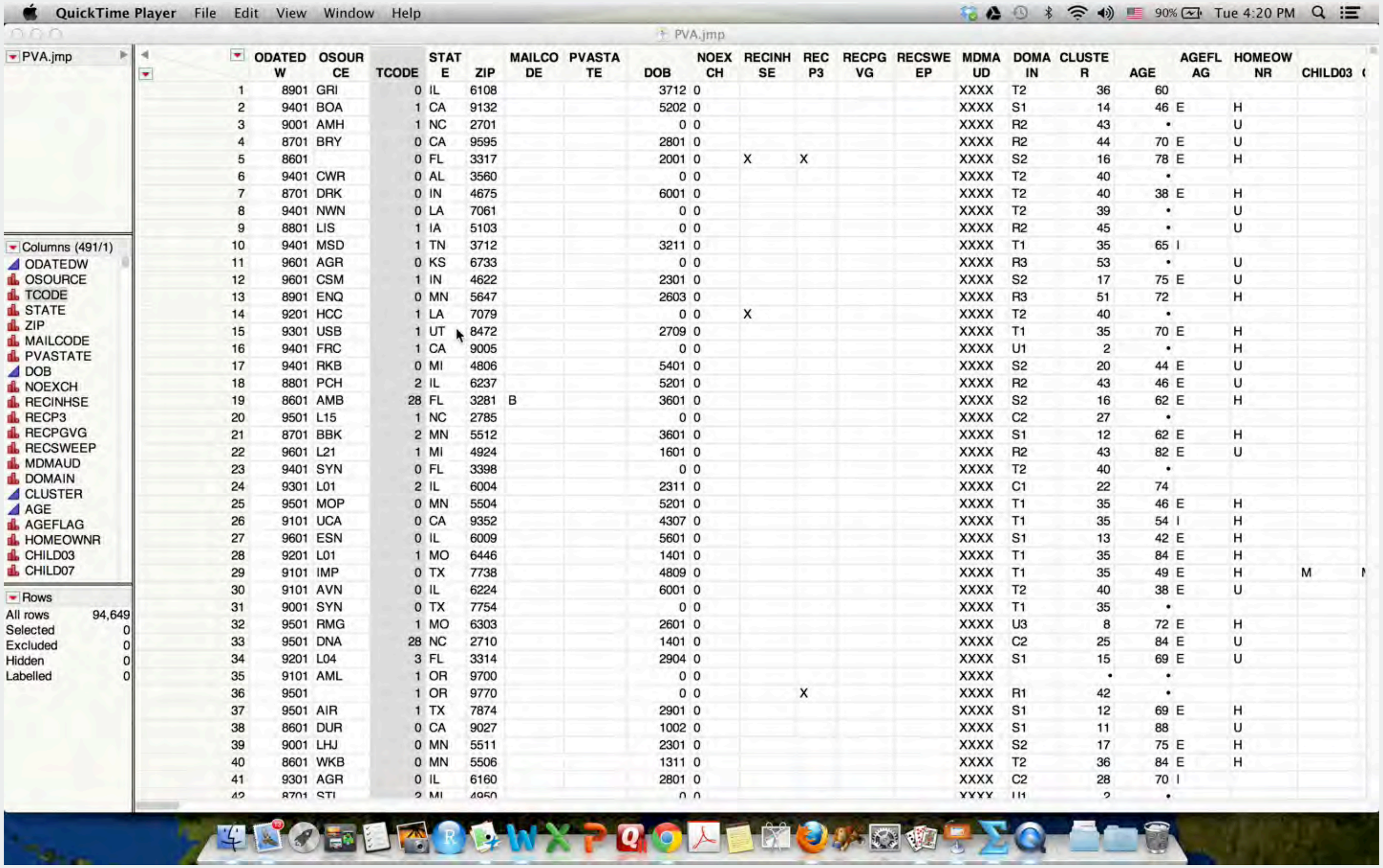
- Find answers to unasked questions
- Eliminate the need to understand your business
- Continuously monitor your data base for interesting patterns
- Eliminate the need to collect good data
- Eliminate the need to have good data analysis skills

WHAT'S 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?



PVA.jmp

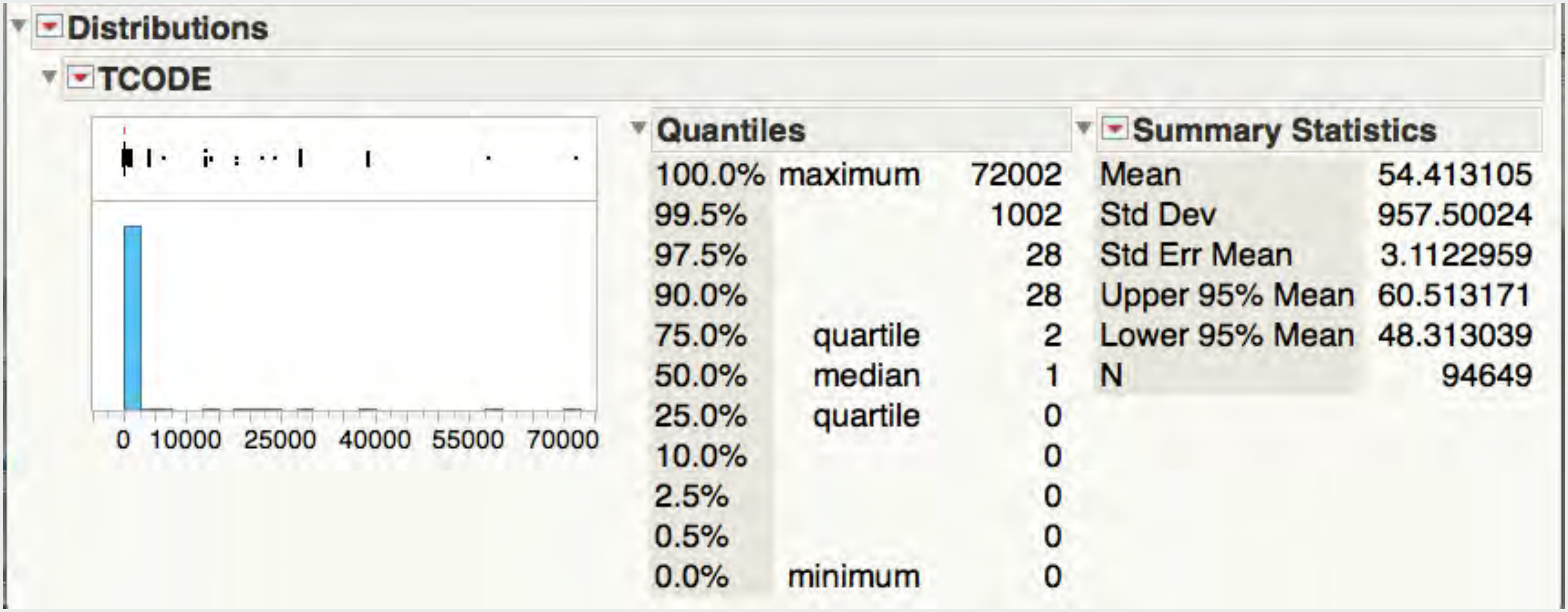
Columns (491/1)
- ODATEDW
- OSOURCE
- TCODE
- STATE
- ZIP
- MAILCODE
- PVASTATE
- DOB
- NOEXCH
- RECINHSE
- RECP3
- RECPGVG
- RECSWEEP
- MDMAUD
- DOMAIN
- CLUSTER
- AGE
- AGEFLAG
- HOMEOWNR
- CHILD03
- CHILD07

Rows

All rows	94,649
Selected	0
Excluded	0
Hidden	0
Labelled	0

	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

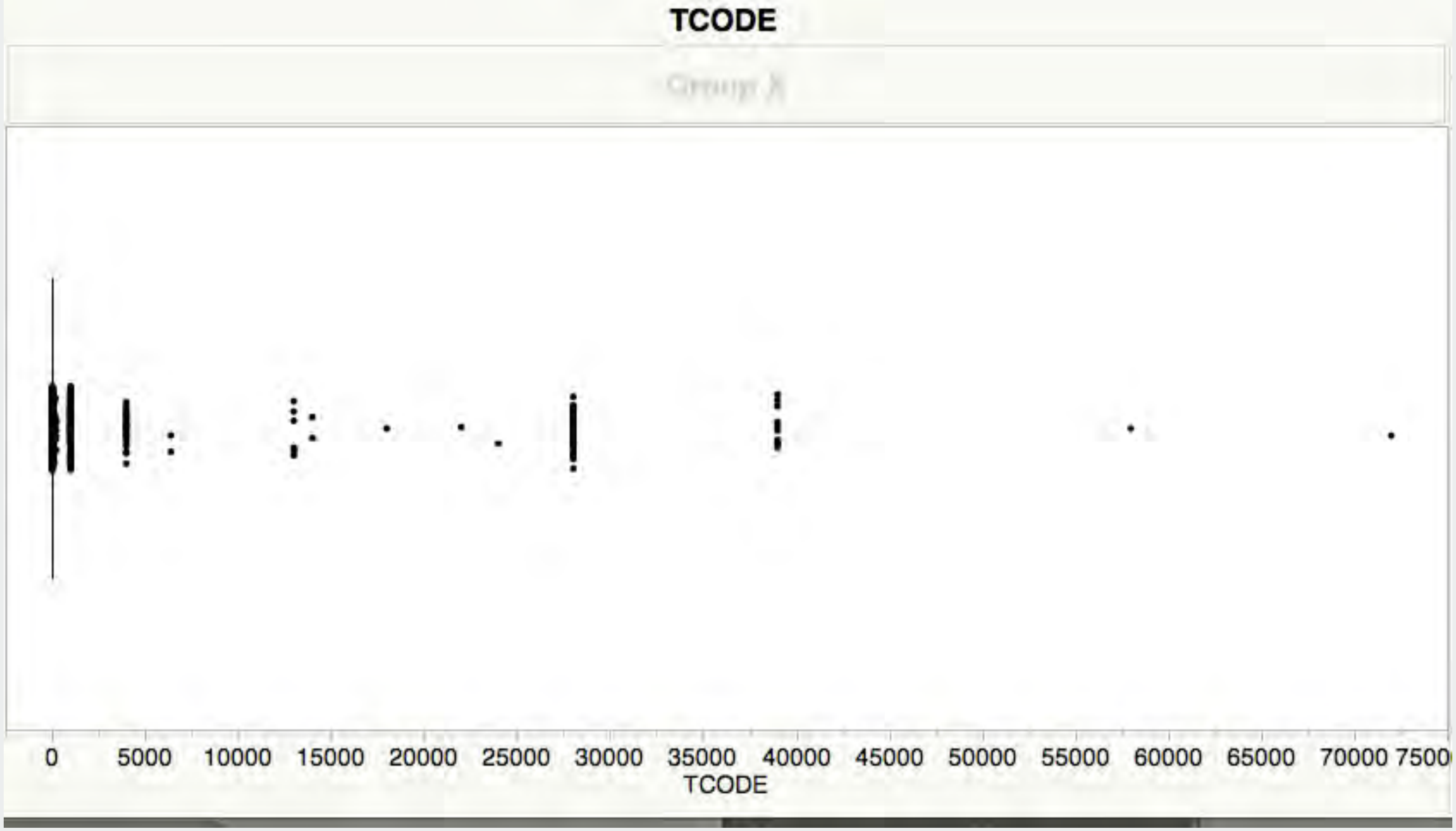
Quantiles

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

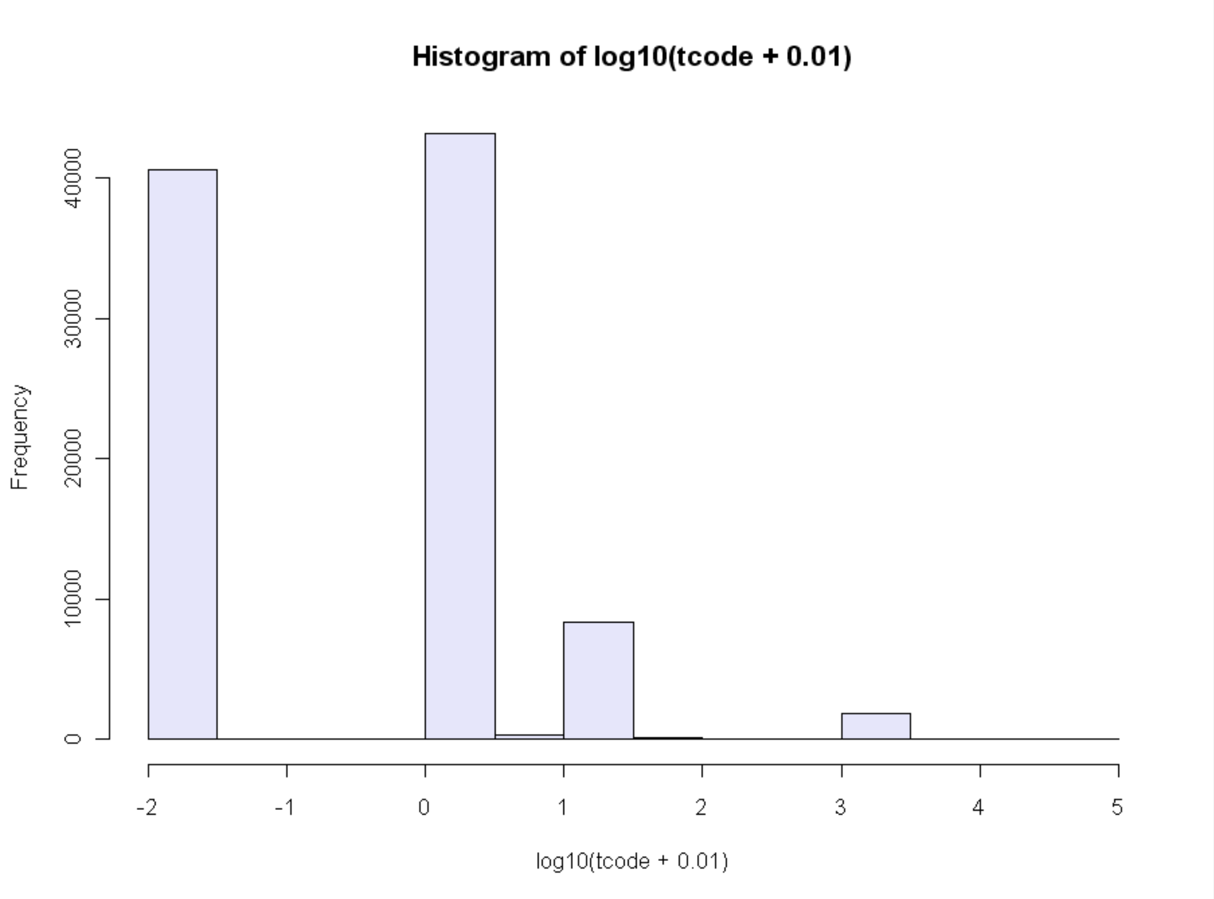
Summary Statistics

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

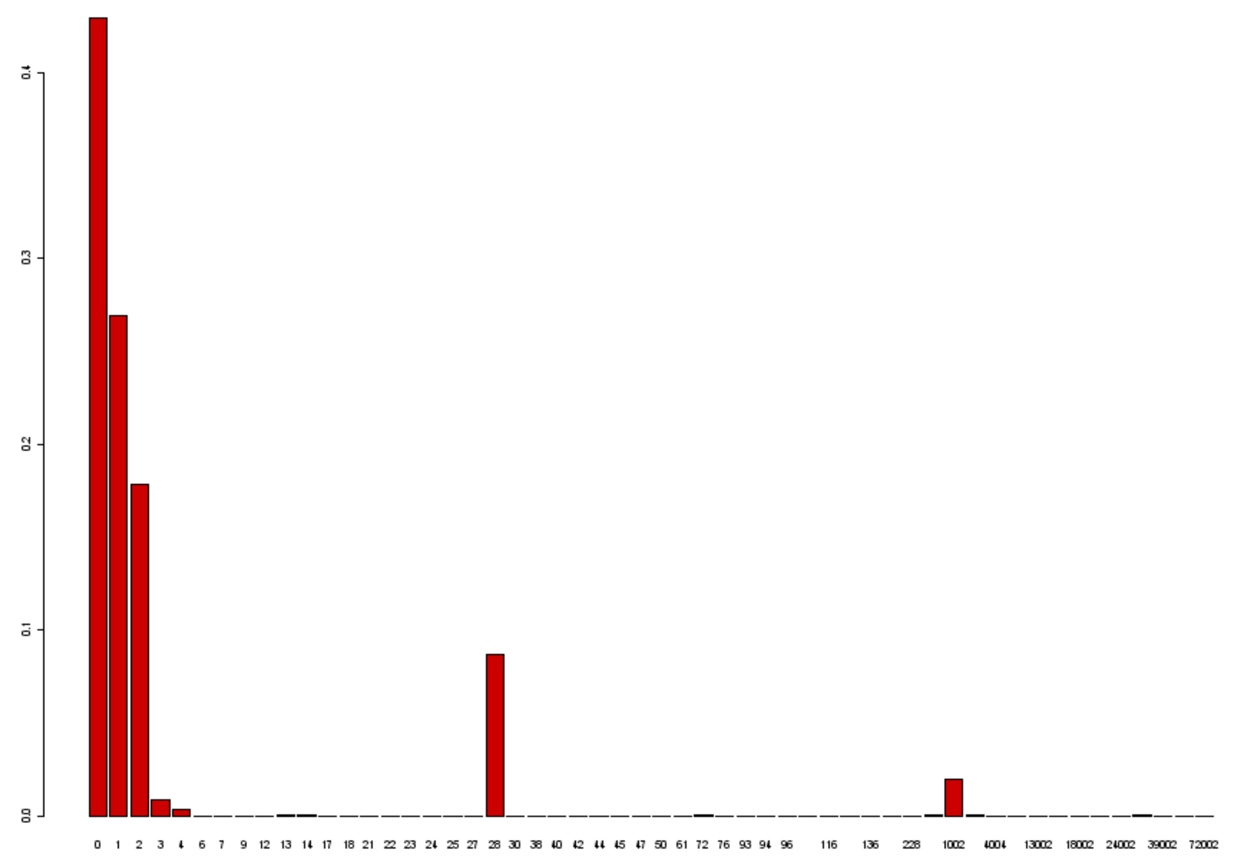
ZOOM OF BOXPLOT



TRANSFORMATION?



MAYBE CATEGORIES?



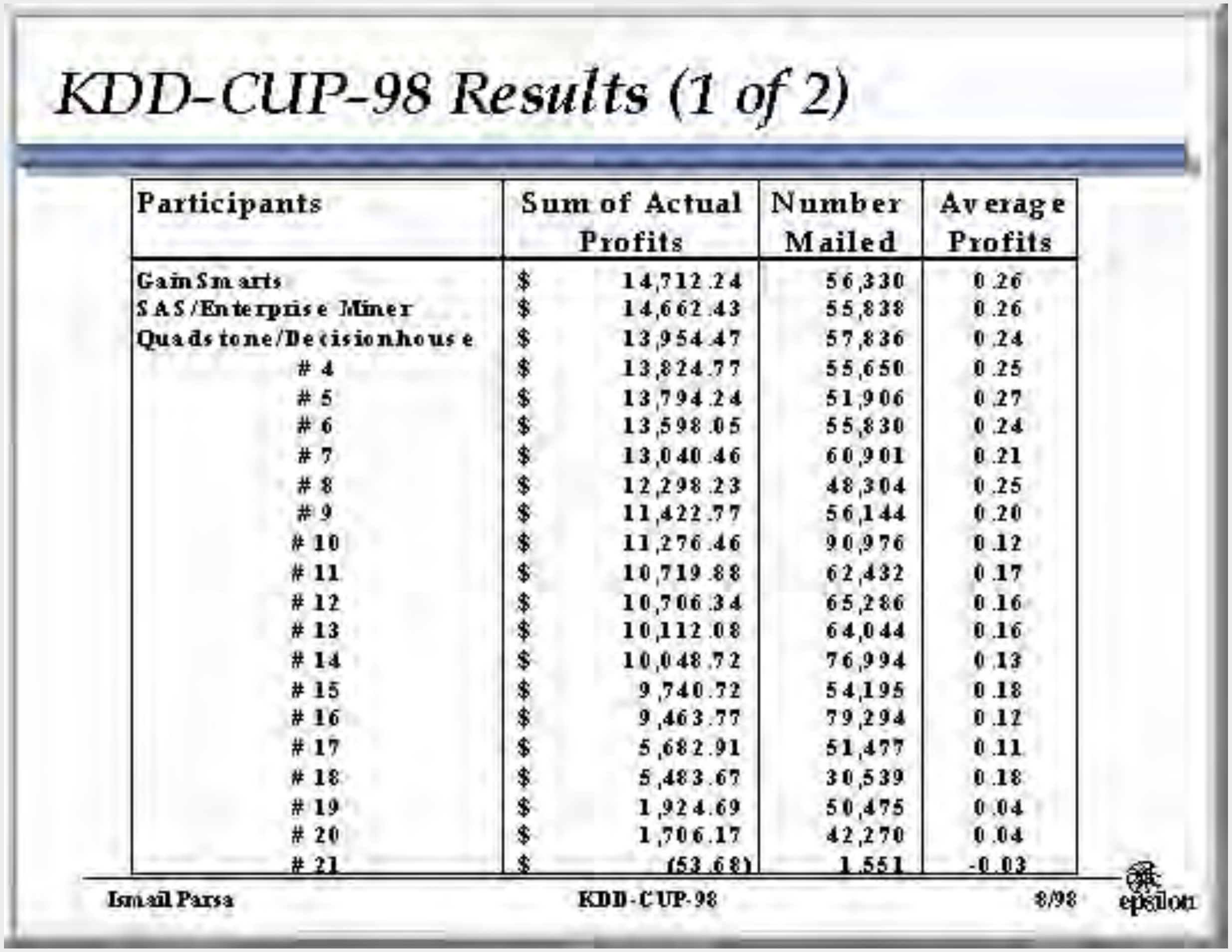
WHAT IS IT, ANYWAY?

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.

RESULTS FOR PVA DATA SET

- If entire list (100,000 donors) is mailed, net donation is $10,500
- Using predictive analytics, this was increased 41.37%

KDD RESULTS

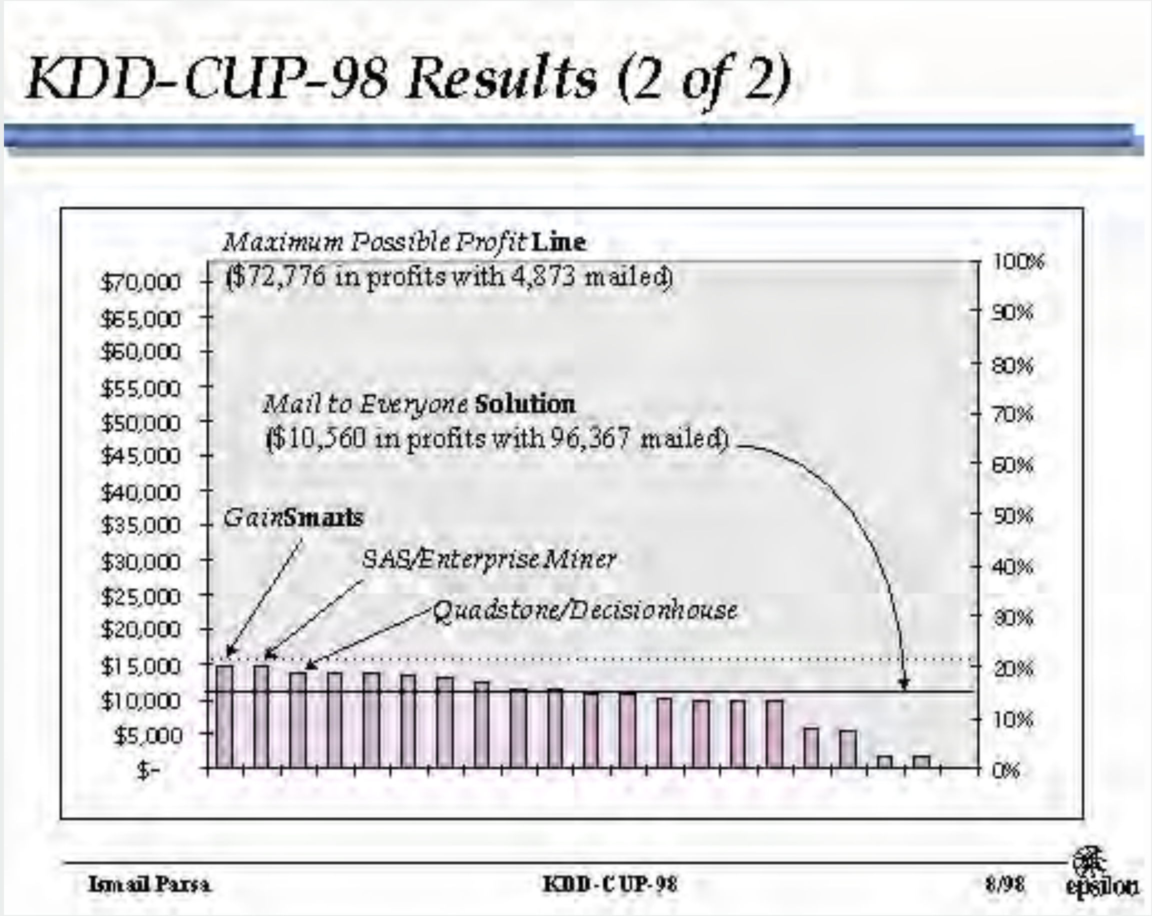


KDD-CUP-98 Results (1 of 2)

Participants	Sum of Actual Profits	Number Mailed	Average Profits
GainSmarts	$ 14,712.24	56,330	0.26
SAS/Enterprise Miner	$ 14,662.43	55,838	0.26
Quadstone/Decisionhouse	$ 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

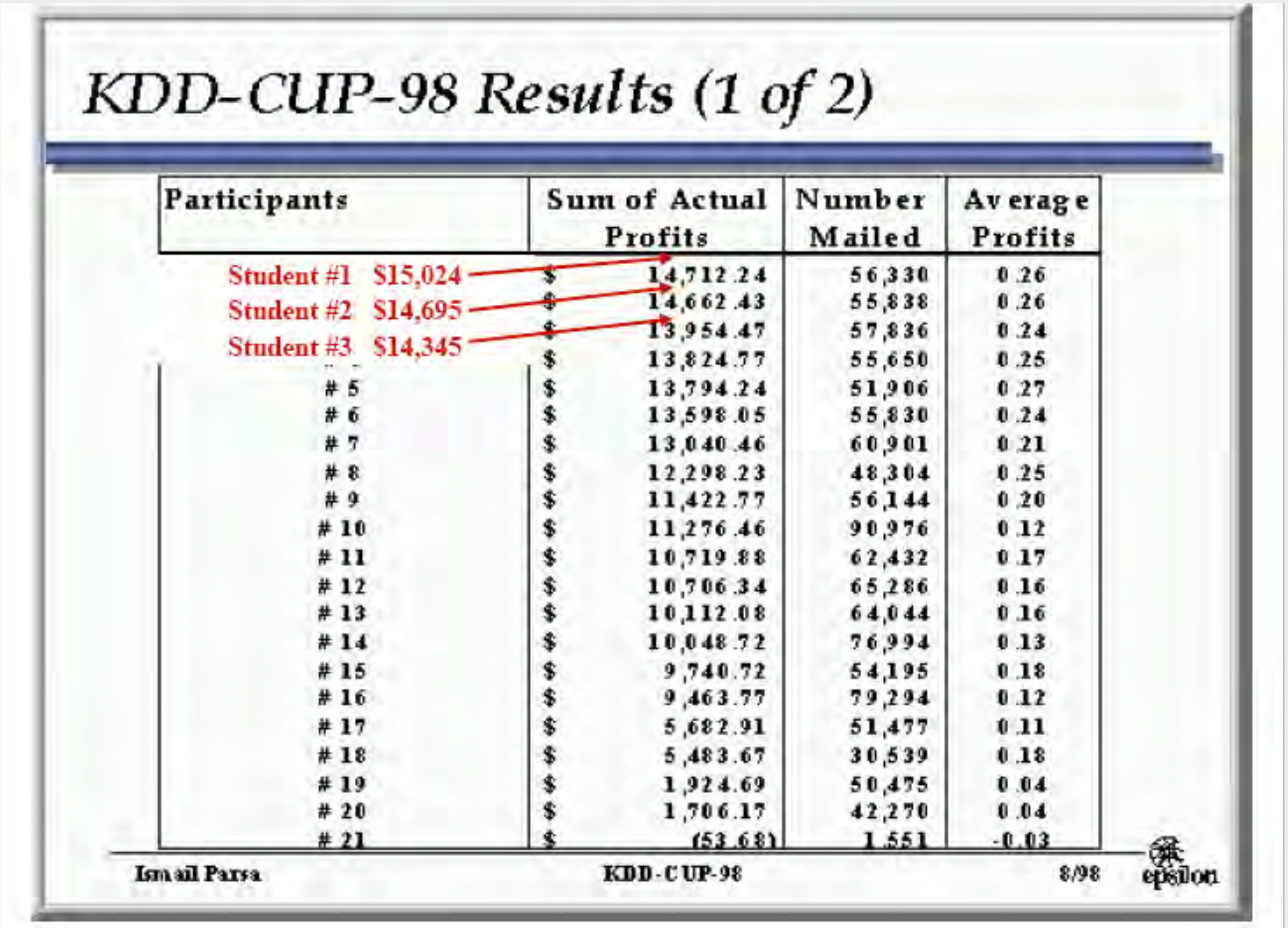
KDD RESULTS -- REDUX



KDD-CUP-98 Results (2 of 2)

Ismail Parsa KDD-CUP-98 8/98 epsilon

STUDENTS IN STAT 442



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
	$ 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

RELATIONAL DATA BASES

- Data are stored in tables

Items

ItemID	ItemName	price
C56621	top hat	34.95
T35691	cane	4.99
RS5292	red shoes	22.95

Shoppers

Person ID	personname	ZIPCODE	item bought
135366	Lyle	19103	T35691
135366	Lyle	19103	C56621
259835	Dick	01267	RS5292

METADATA

The data survey describes the data set contents and characteristics

- Table name
- Description
- Primary key/foreign key relationships
- Collection information: how, where, conditions
- Timeframe: daily, weekly, monthly
- Cosynchronus: every Monday or Tuesday

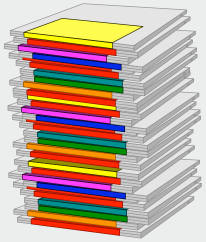
DATA PREPARATION

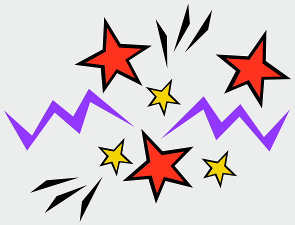
60% to 95% of the time is spent preparing the data

85.32% of all statistics are made up

DATA CHALLENGES

- **Data definitions**
  - Types of variables
- **Data consolidation**
  - Combine data from different sources
  - NASA mars lander
- **Data heterogeneity**
  - Homonyms
  - Synonyms
- **Data quality**





MISSING VALUES

- Credit Card Bank finds that "Income" field is missing
- Wharton Ph.D. Student questionnaire on survey attitudes
- Bowdoin college applicants have mean SAT verbal score above 750
- Clinical Trial of Depression Medication – what does missing mean?

DEPRESSION STUDY

- **Designed to study antidepressant efficacy**
  - Measured via Hamilton Rating Scale

- Side effects
  - Sexual dysfunction
  - Misc safety and tolerability issues

- Late '97 and early '98

- 692 patients

- Two antidepressants + placebo

Have you tried taking medication for your depression?
Are you dissatisfied with your improvement?

If so, you may wish to consider participating in a clinical research study of an investigational medication for people with depression who are not responding to treatment.

Eligible candidates must be between the ages of 18 and 65 and tried other drug therapies and/or combinations of drugs for depression, and not had adequate relief of symptoms.

Qualified participants may receive study-related evaluations and medical care, research medication and laboratory work, at no cost.

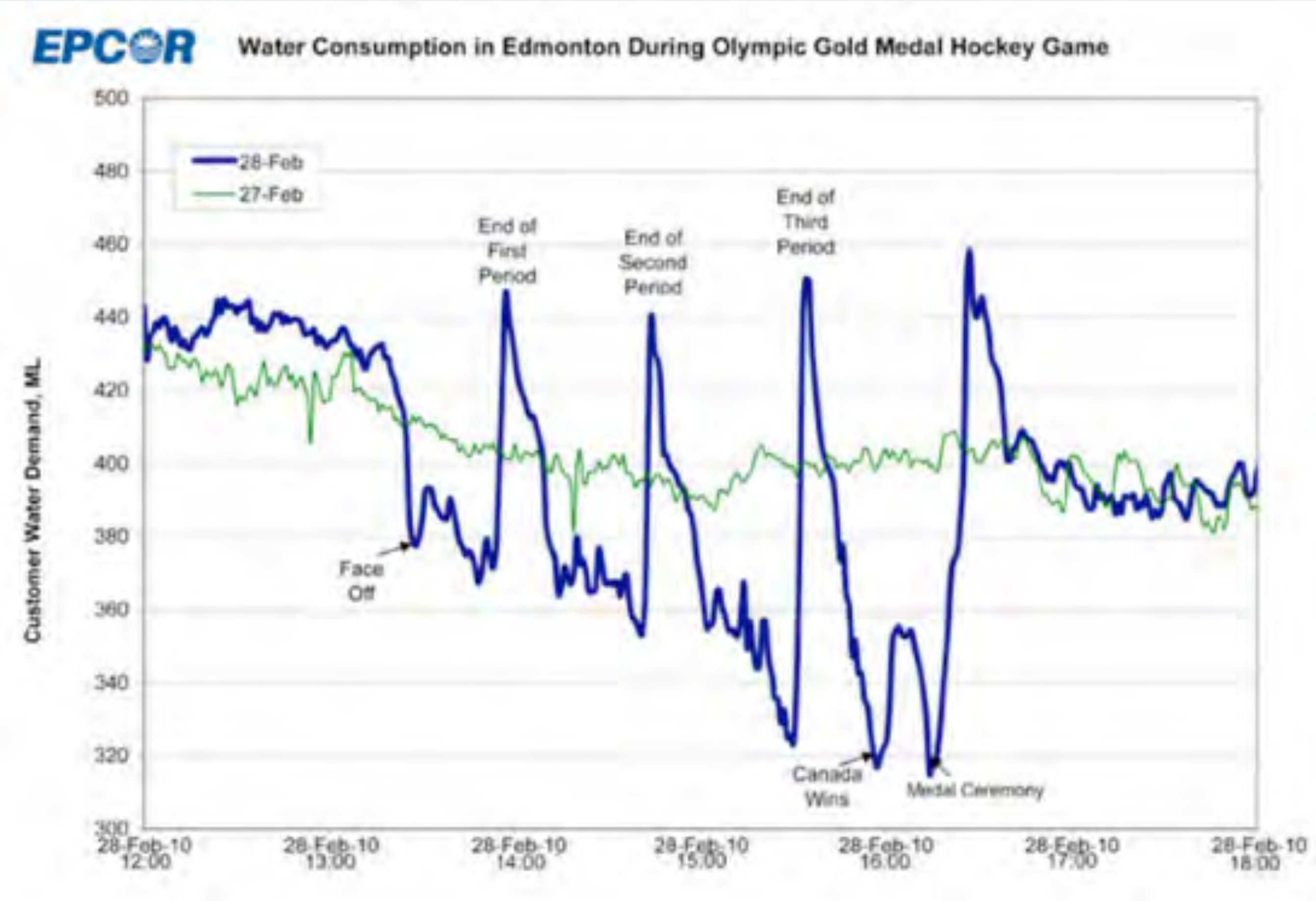
To learn if you may be eligible to participate in this clinical study, please call Frank Posada at 972-283-6286

That's 972-283-6286
Call today.

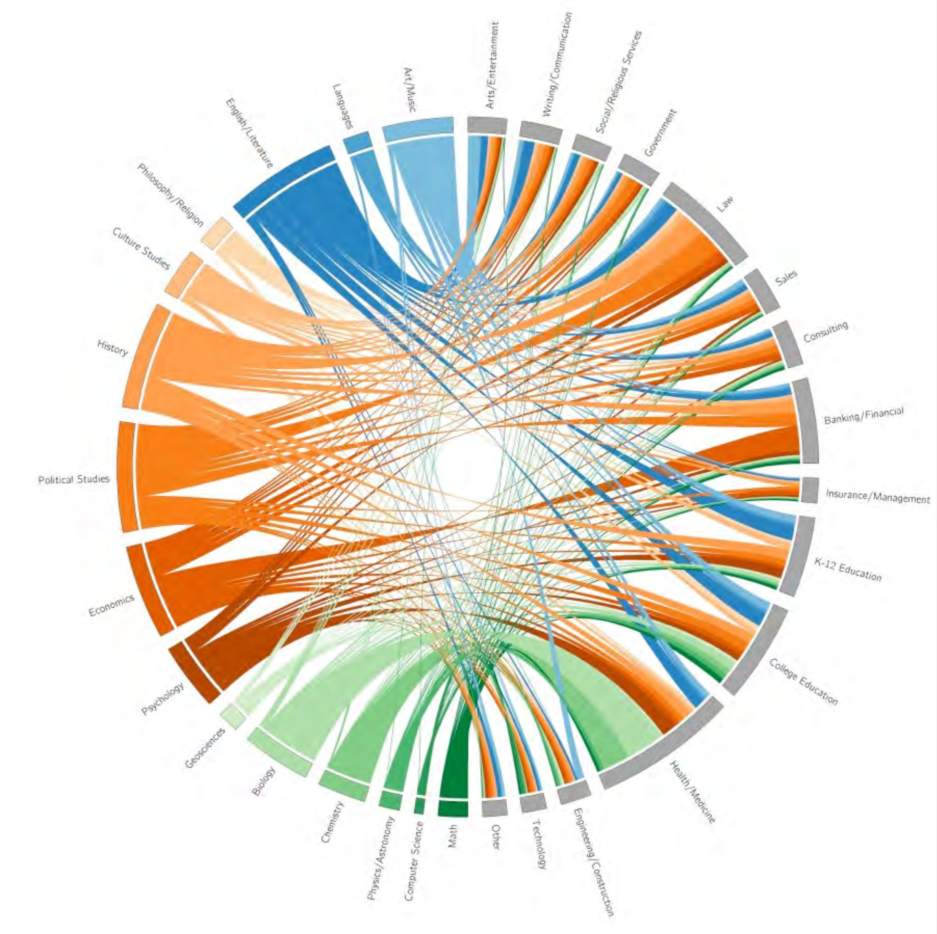
DEPRESSION STUDY DATA

- **Background info**
  - Age
  - Sex
- **Each received either**
  - Placebo
  - Anti depressant 1
  - Anti depressant 2
- **Dosages**
- **At time points 7 and 14 days we also have:**
  - Depression scores
  - Sexual Dysfunction Data
  - **Response indicators**

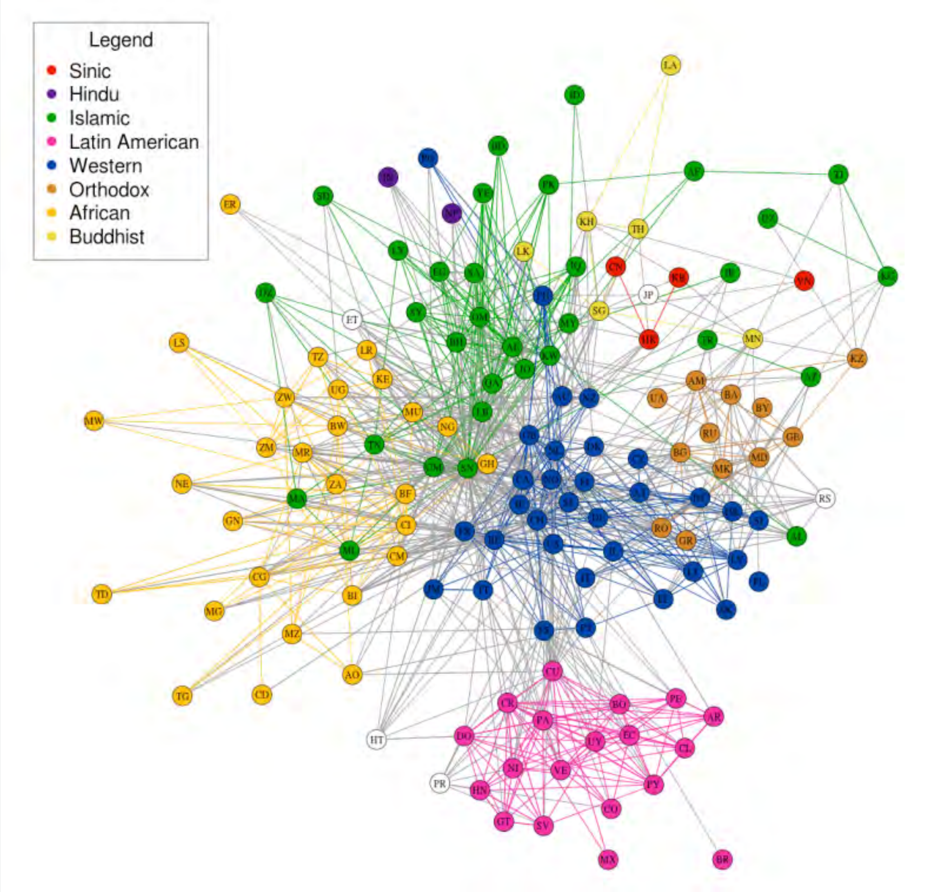
GRAPHICS


EPCOR
Water Consumption in Edmonton During Olympic Gold Medal Hockey Game
Customer Water Demand, ML
500
480
460
440
420
400
380
360
340
320
300
28-Feb
27-Feb
End of First Period
End of Second Period
End of Third Period
Face Off
Canada Wins
Medal Ceremony
28-Feb-10 12:00
28-Feb-10 13:00
28-Feb-10 14:00
28-Feb-10 15:00
28-Feb-10 16:00
28-Feb-10 17:00
28-Feb-10 18:00

15,600 WILLIAMS COLLEGE ALUMS



NETWORKS


Legend
Sinic
Hindu
Islamic
Latin American
Western
Orthodox
African
Buddhist

LESLIE

NameVoyager: Explore name trends letter by letter

Surf the NameVoyager by name endings, letter combos & popularity, plus get more amazing name finding tools with Baby Name W

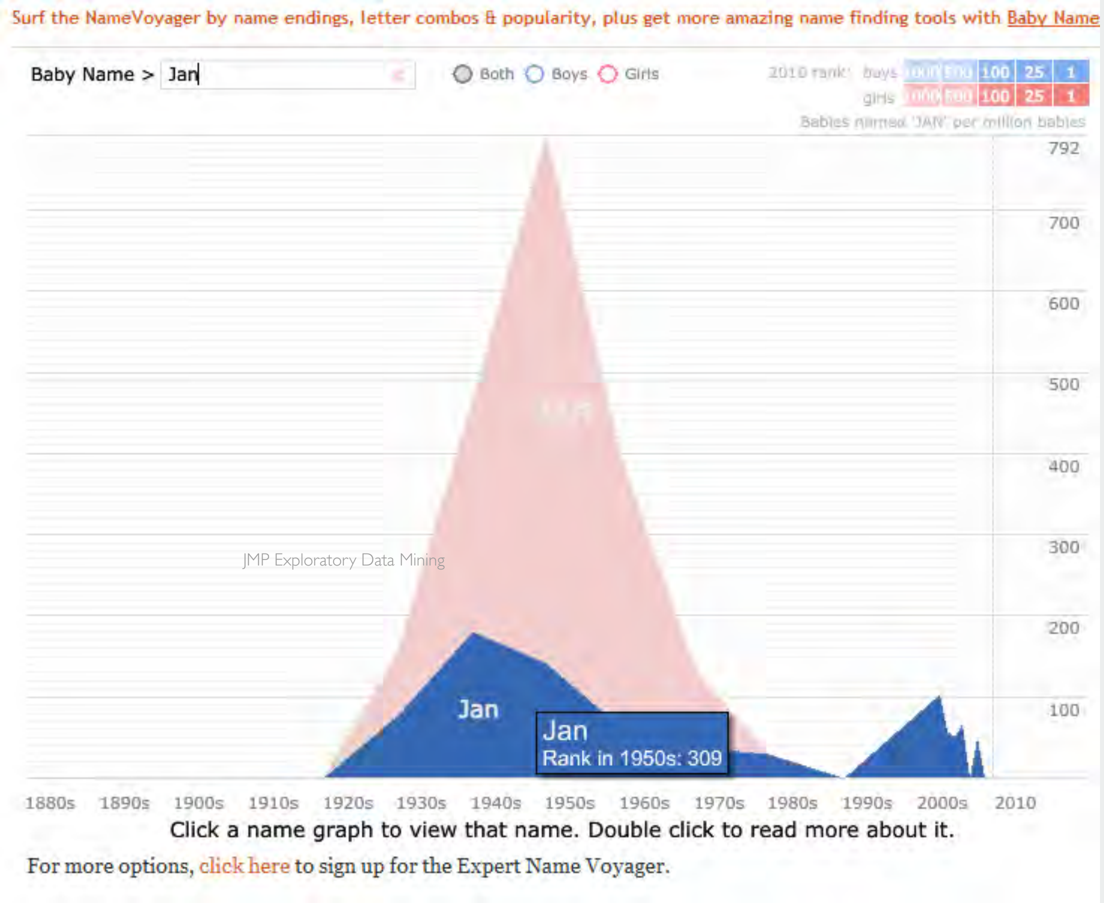
Baby Name > Leslie

Press 'enter' to see exact matches.

Click a name graph to view that name. Double click to read more about it.

For more options, click here to sign up for the Expert Name Voyager.

JAN…



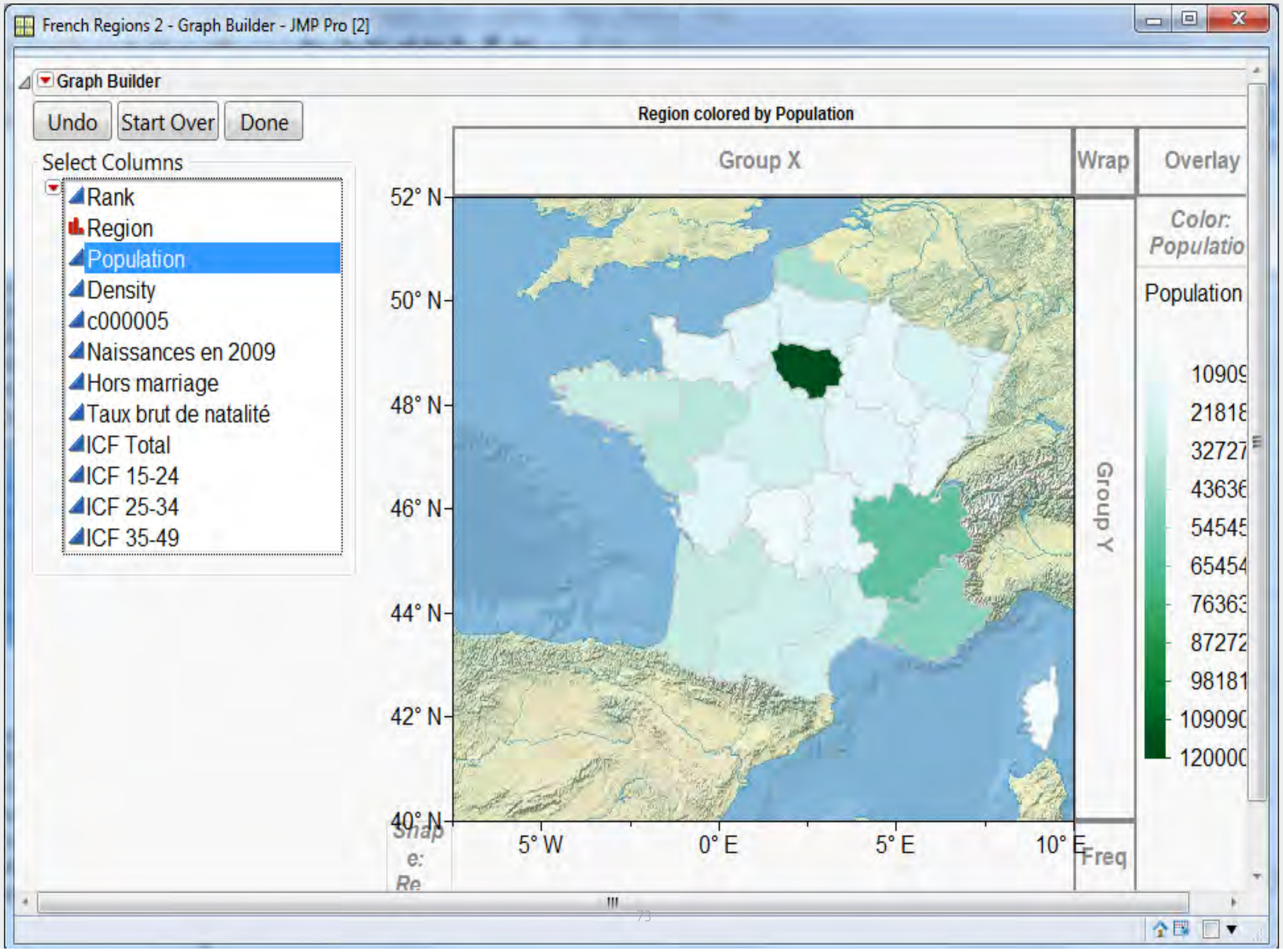
Surf the NameVoyager by name endings, letter combos & popularity, plus get more amazing name finding tools with Baby Name

Baby Name > Jan

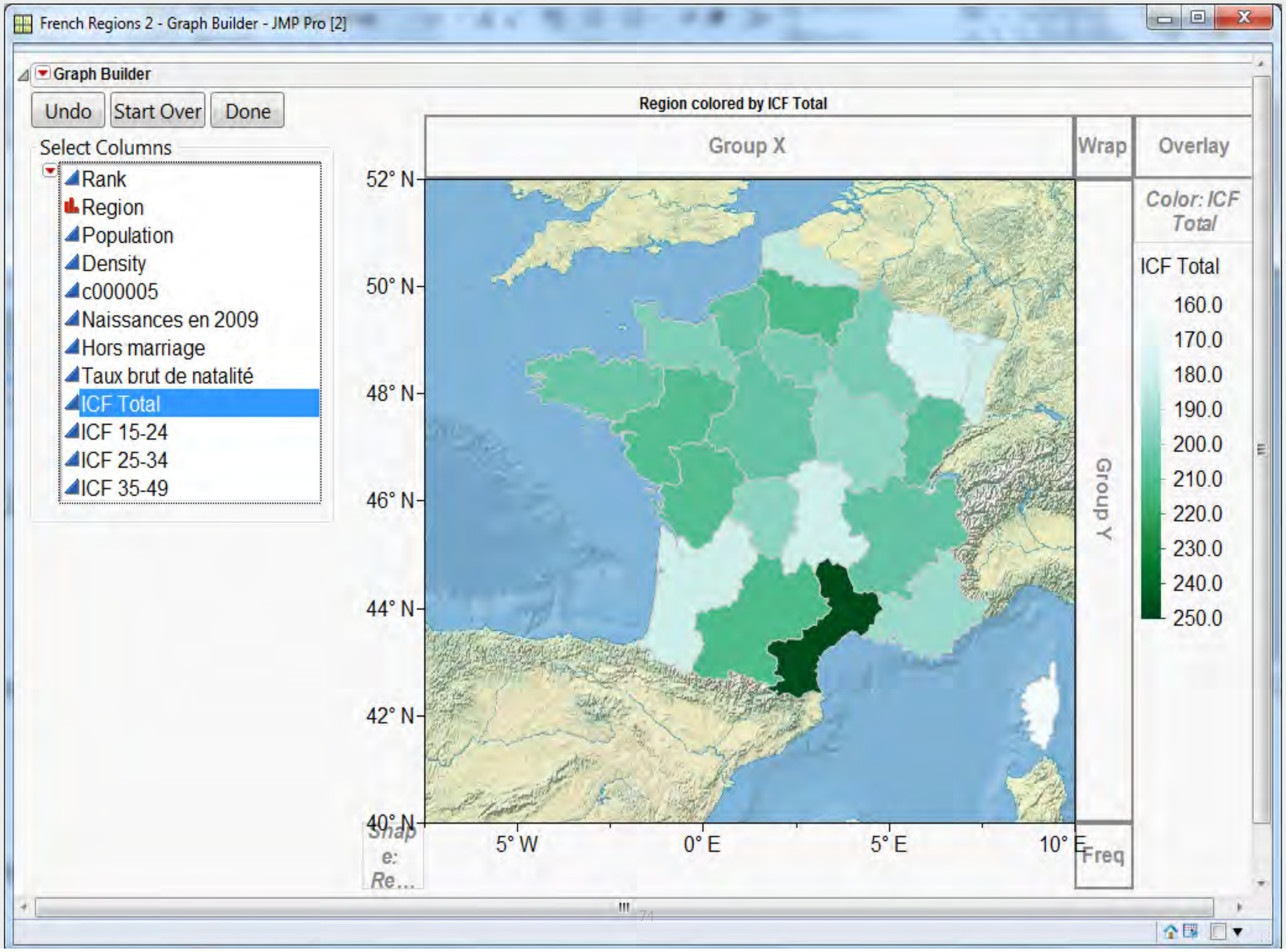
Click a name graph to view that name. Double click to read more about it.

For more options, click here to sign up for the Expert Name Voyager.

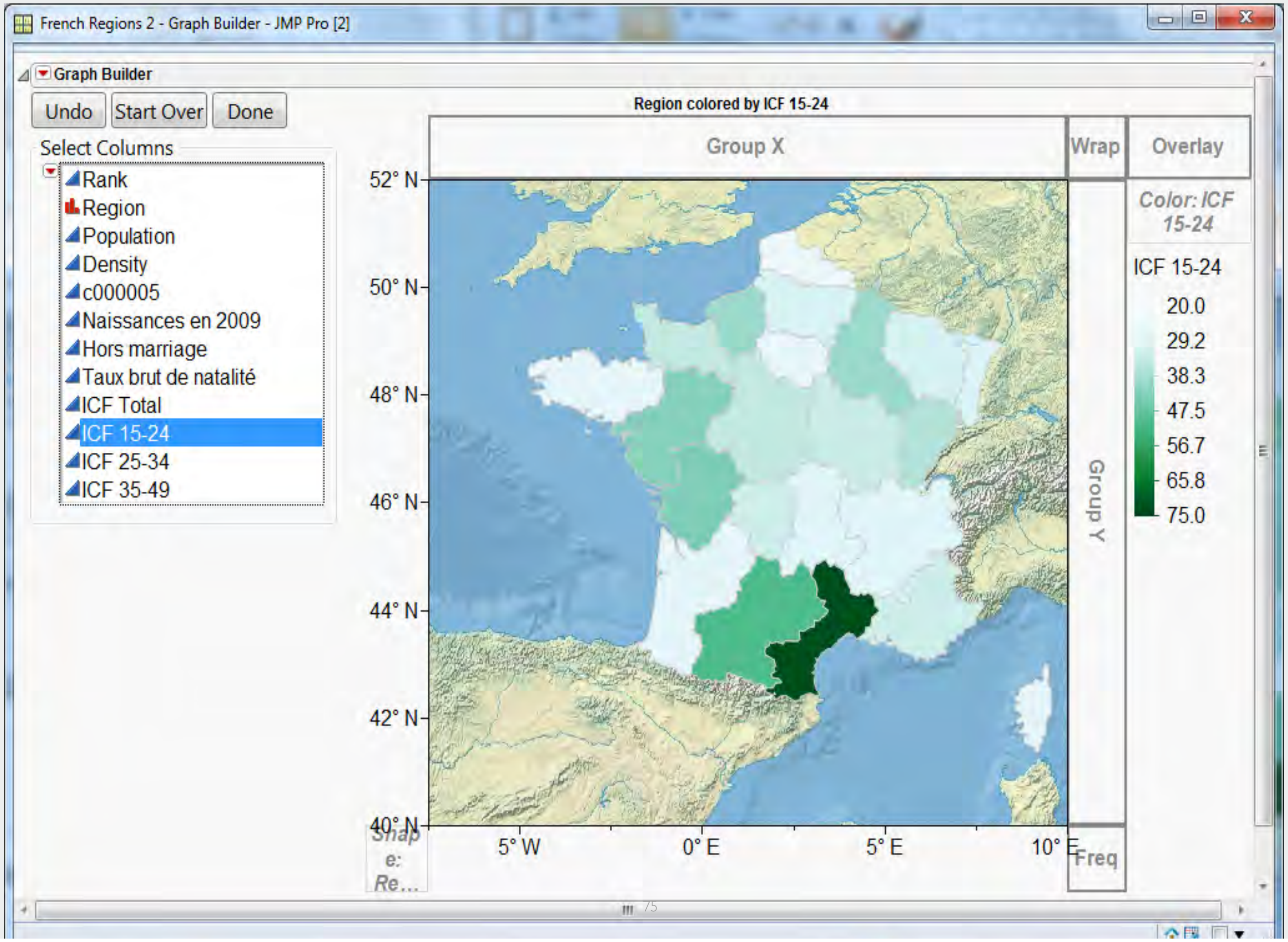
WHERE ARE THE PEOPLE?



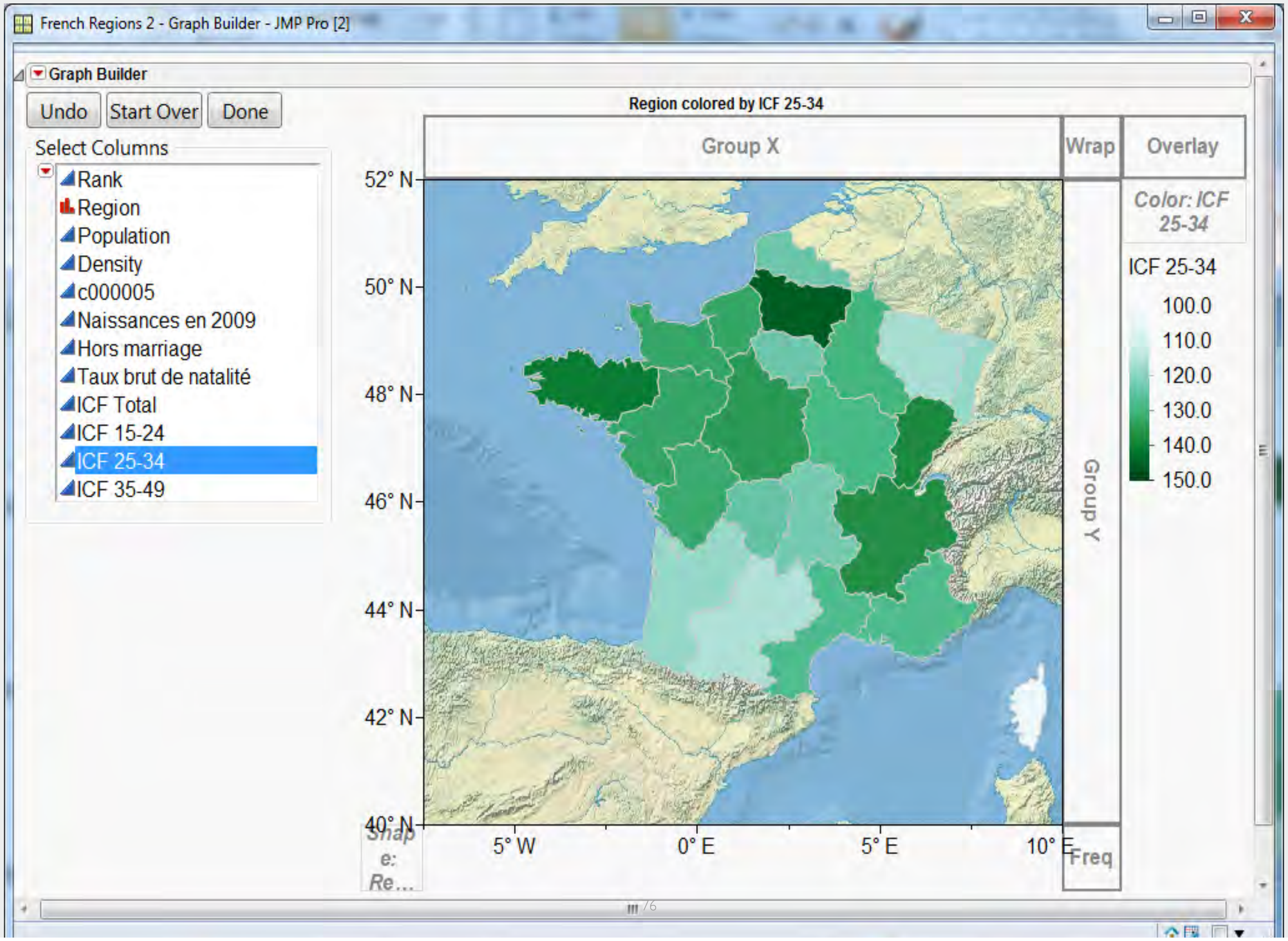
WHERE ARE THE BIRTHS?



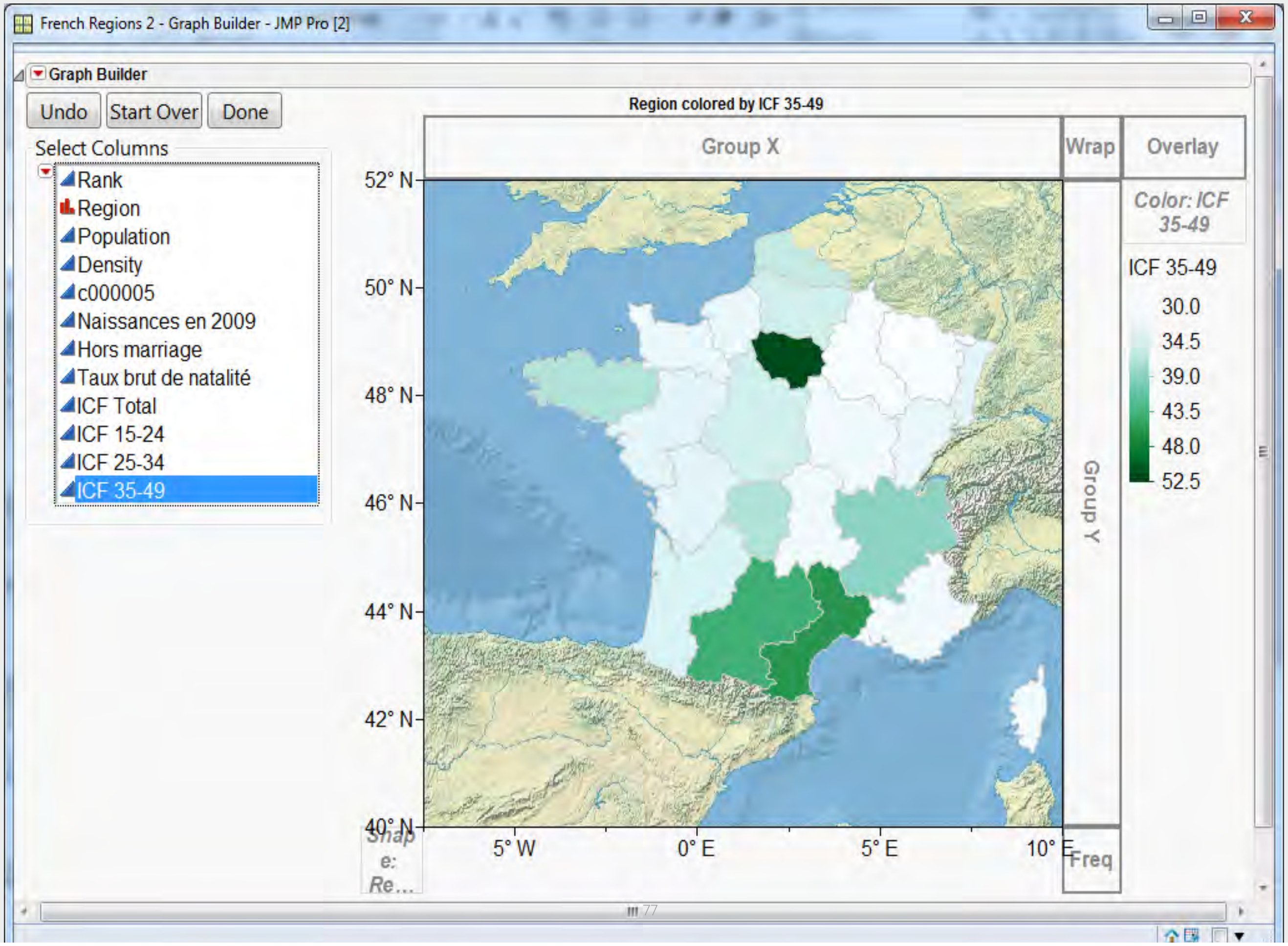
WHERE ARE THE TEENAGE BIRTHS?



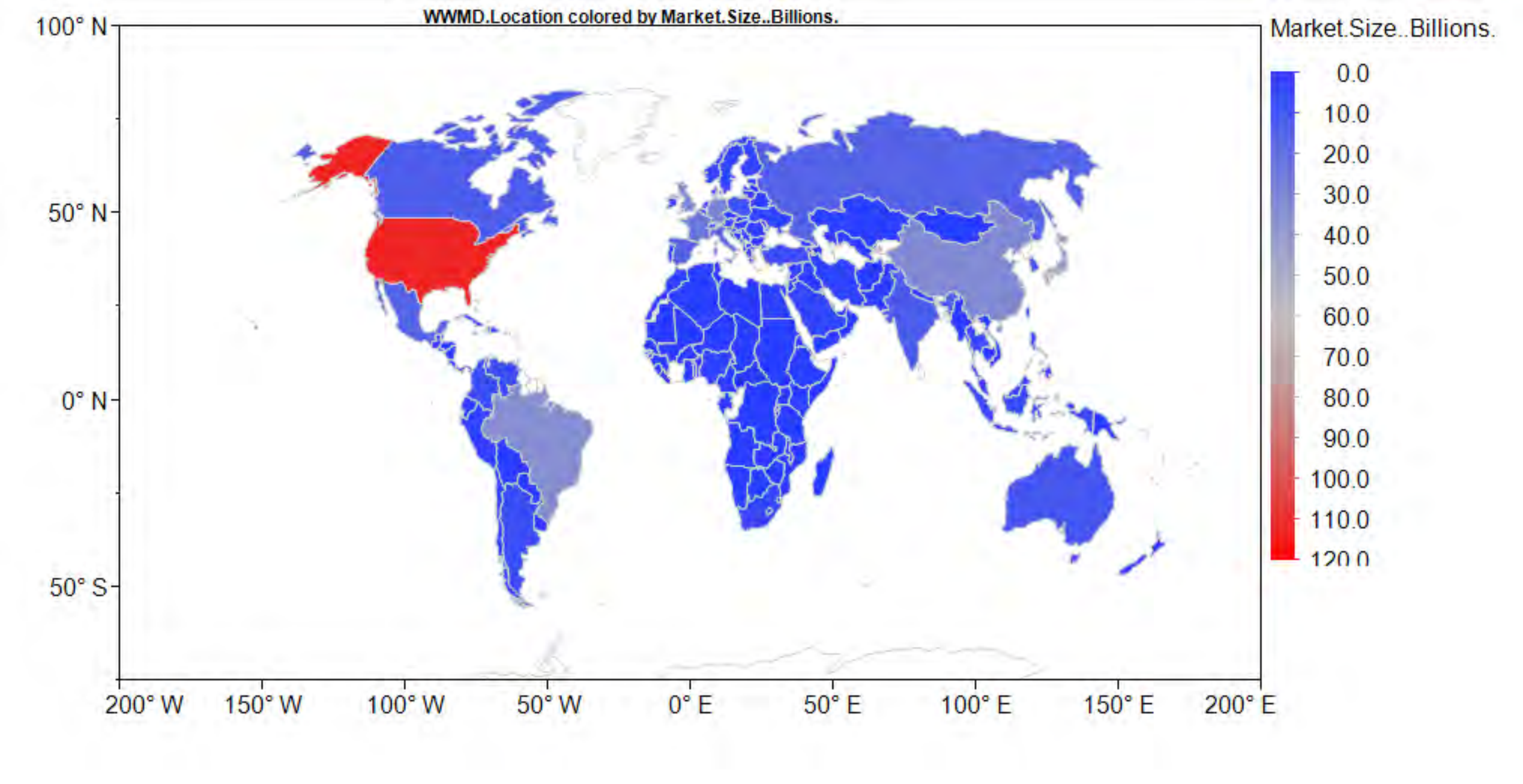
TO WOMEN 25-34?



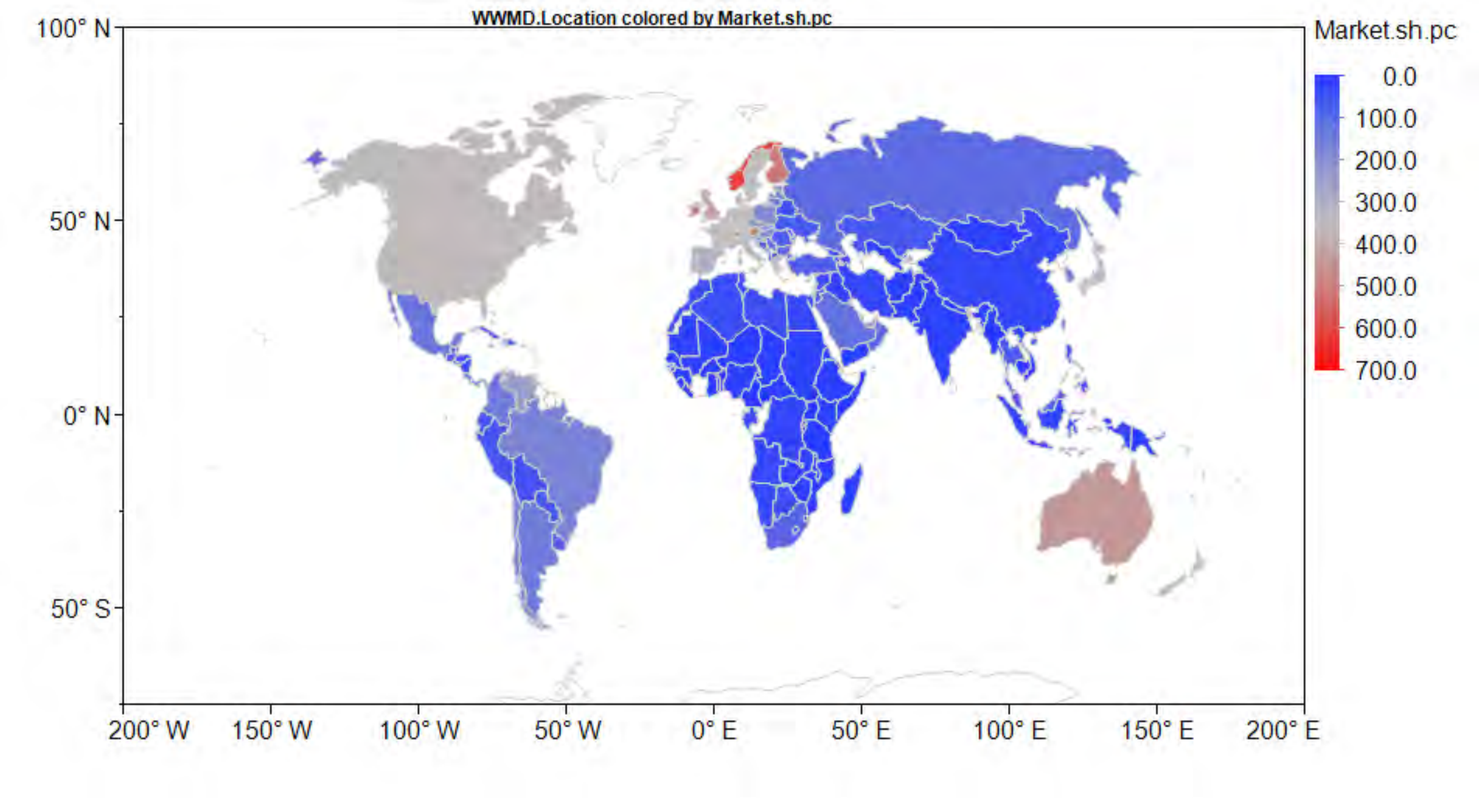
TO WOMEN 35 AND OVER?



WORLD WIDE MARKET SIZE

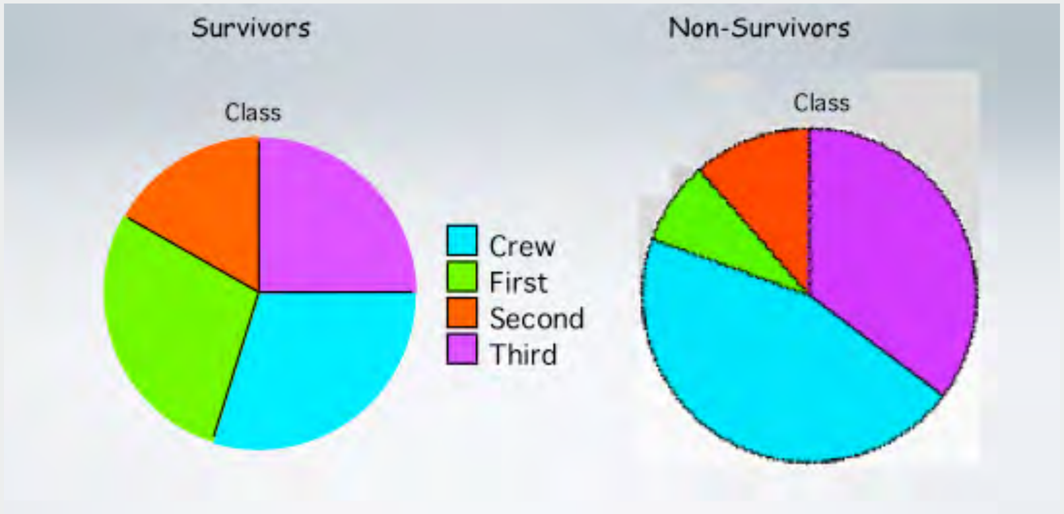


MARKET SIZE PER CAPITA

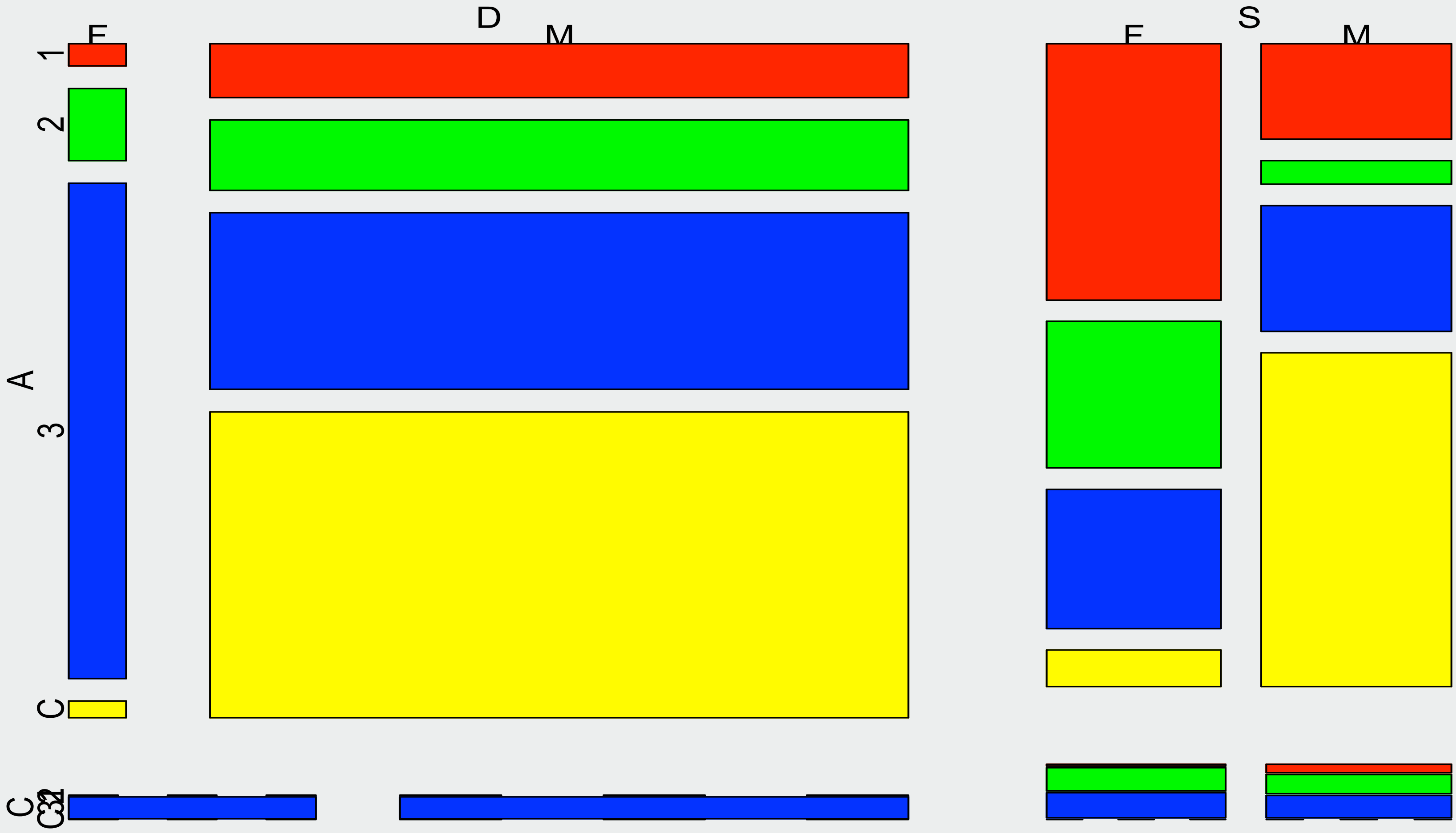


TWO WAY TABLES -- TITANIC

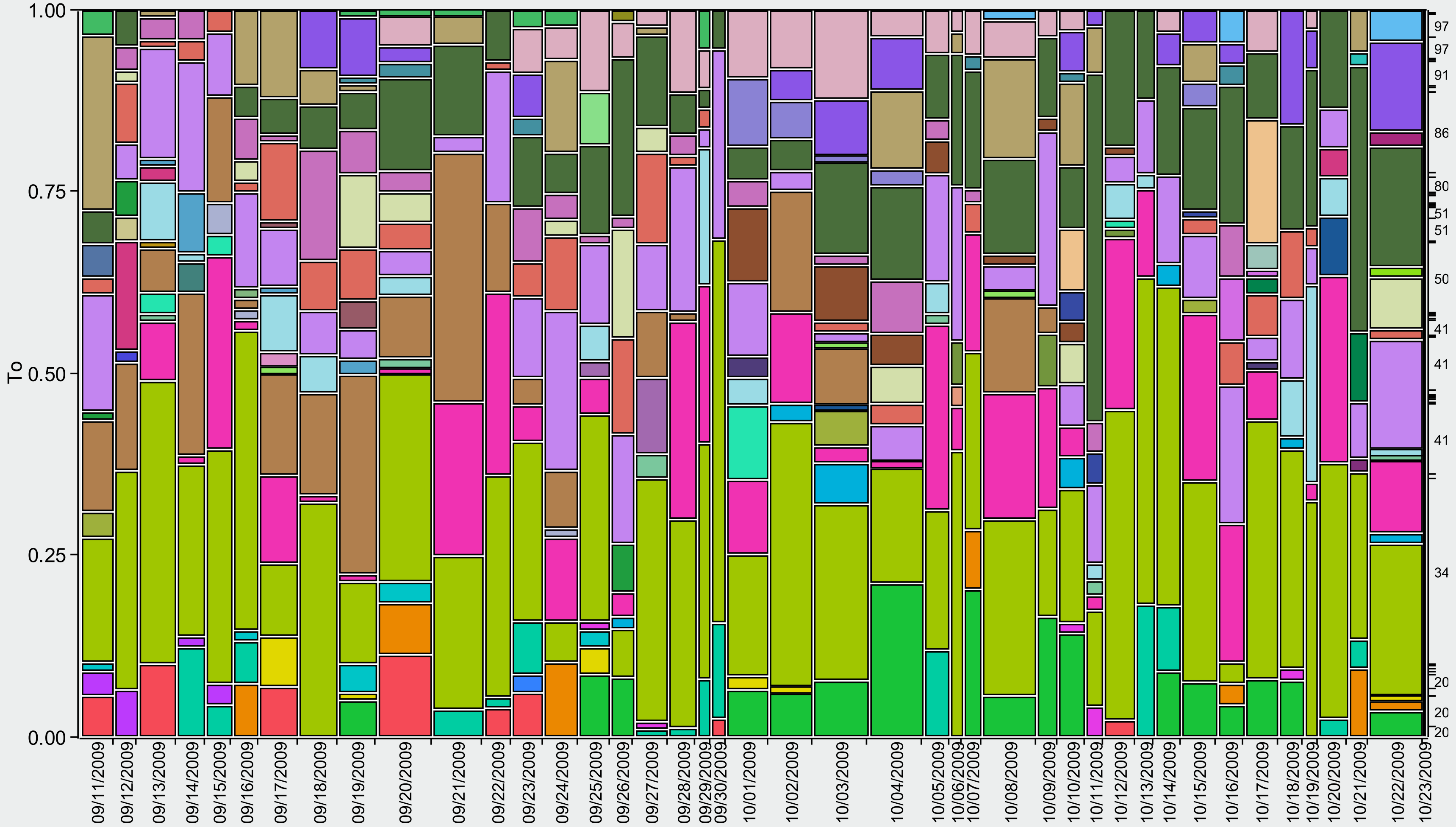
		Ticket Class				
		Crew	First	Second	Third	Total
	Lived	212	202	118	178	**710**
Survival	Died	673	123	167	528	**1491**
	Total	**885**	**325**	**285**	**706**	**2201**



MOSAIC PLOT

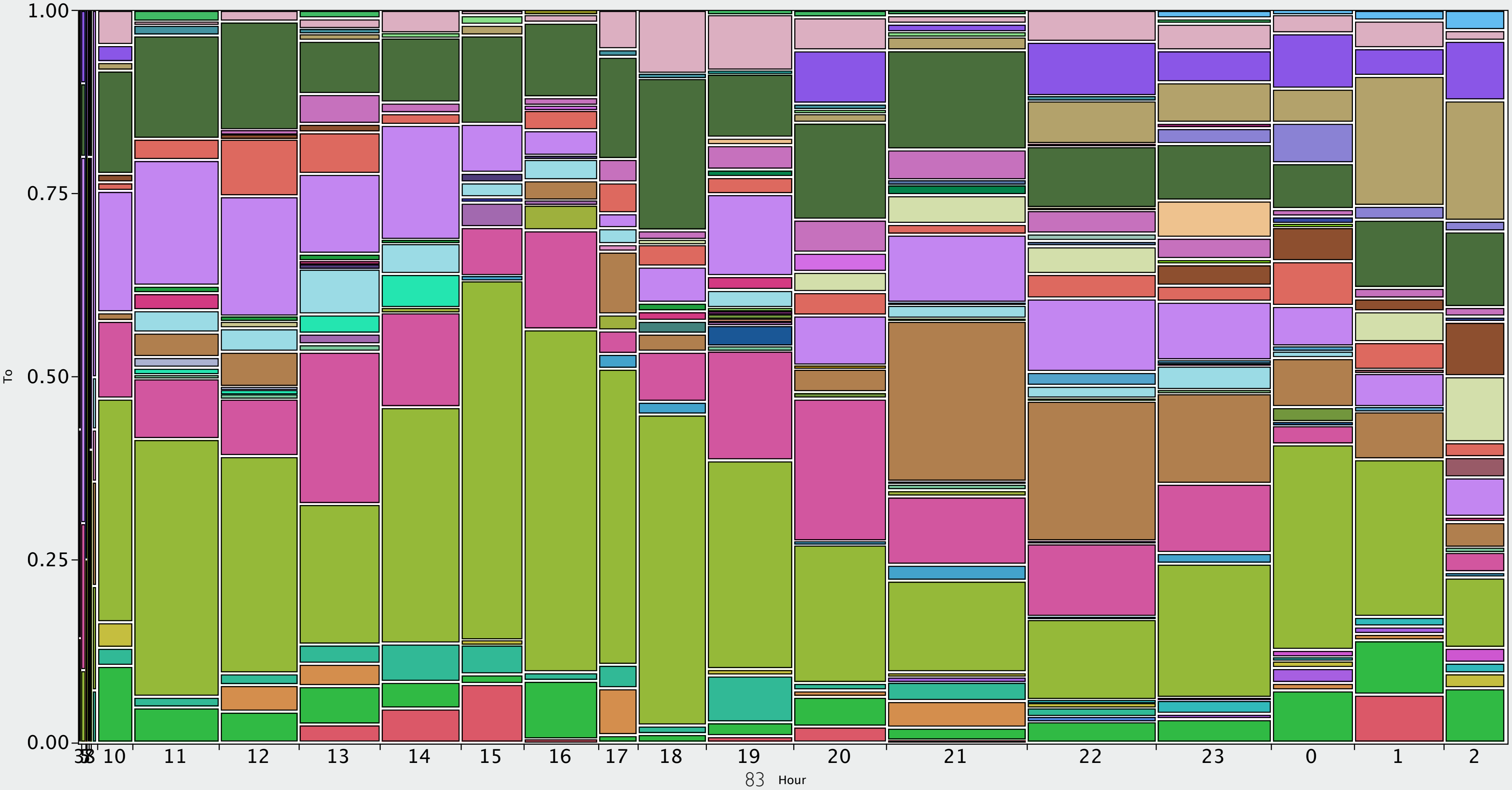


MOSAIC PLOT



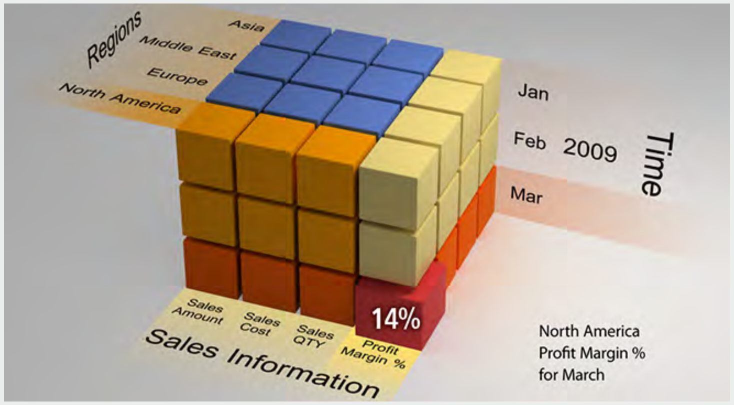
1.00
0.75
0.50
0.25
0.00
To
09/11/2009
09/12/2009
09/13/2009
09/14/2009
09/15/2009
09/16/2009
09/17/2009
09/18/2009
09/19/2009
09/20/2009
09/21/2009
09/22/2009
09/23/2009
09/24/2009
09/25/2009
09/26/2009
09/27/2009
09/28/2009
09/29/2009
09/30/2009
10/01/2009
10/02/2009
10/03/2009
10/04/2009
10/05/2009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

TEXTING BY HOUR



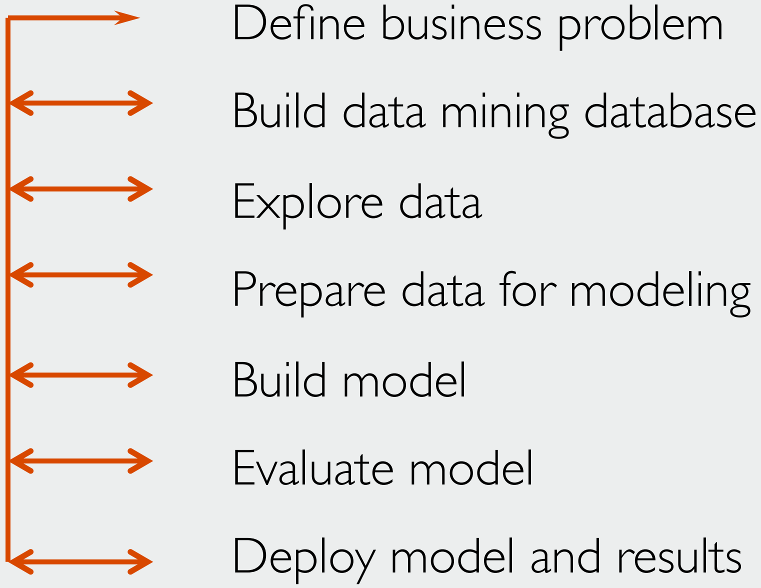
ANALYTICS AND OLAP

- **On-line analytical processing (OLAP)**
  - Slicing and Dicing
  - Descriptive, not predictive



- **Predictive Analytics:**
  - Uses data to inductively find patterns – models!
- **Associations?**
  - Most associated variables in the census
  - Most associated variables in a supermarket

KNOWLEDGE DISCOVERY PROCESS

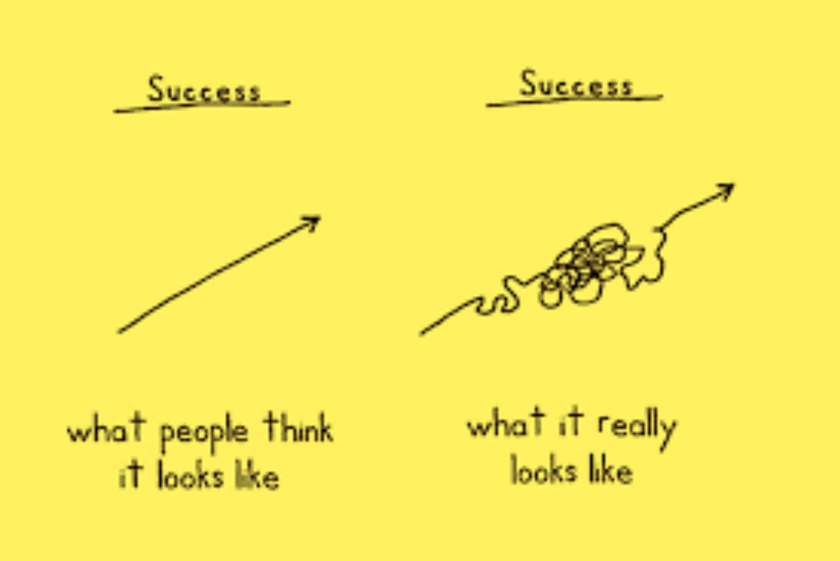


- Define business problem
- Build data mining database
- Explore data
- Prepare data for modeling
- Build model
- Evaluate model
- Deploy model and results

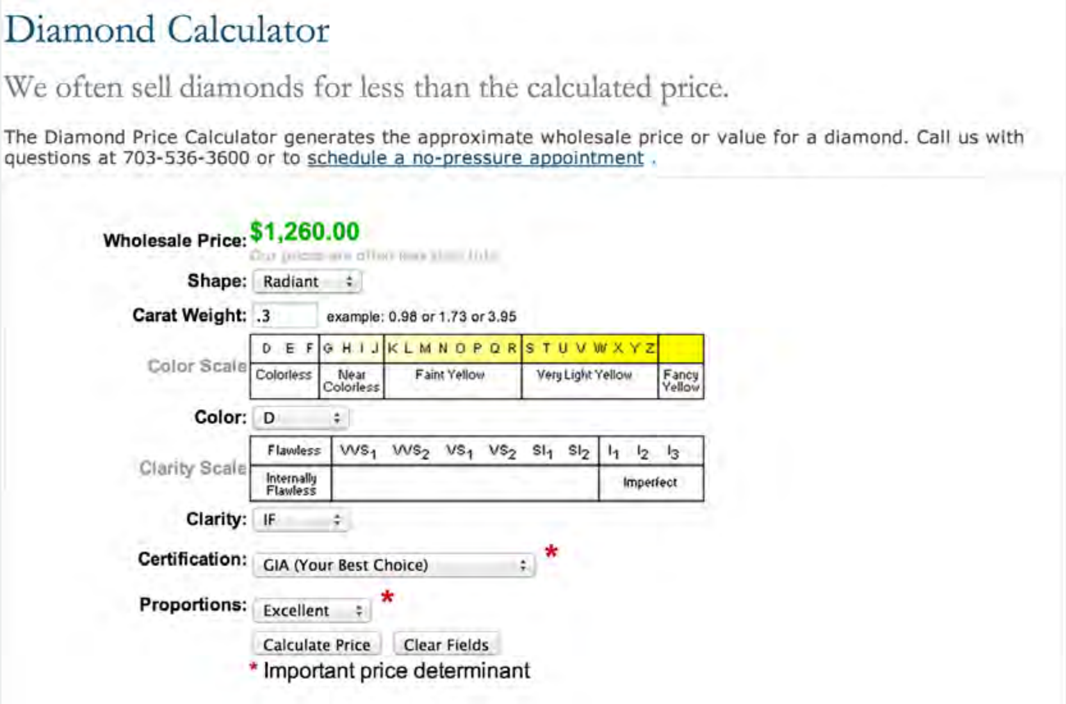
Note: This process model is adapted from
CRISP-DM: CRoss Industry Standard Process for Data Mining

KEYS TO SUCCESSFUL DATA MINING

- The keys to success:
  - Formulating the problem
  - Using the right data
  - Flexibility in modeling
  - Acting on results
- Success depends more on the way you mine the data rather than the specific tool



DIAMOND PRICES



Diamond Calculator

We often sell diamonds for less than the calculated price.

The Diamond Price Calculator generates the approximate wholesale price or value for a diamond. Call us with questions at 703-536-3600 or to schedule a no-pressure appointment .

Wholesale Price: $1,260.00

Shape: Radiant

Carat Weight: .3 example: 0.98 or 1.73 or 3.95

Color Scale

D E F	G H I J	K L M N O P Q R	S T U V W X Y Z	
Colorless	Near Colorless	Faint Yellow	Very Light Yellow	Fancy Yellow

Color: D

Clarity Scale

Flawless	VVS_1 VVS_2 VS_1 VS_2 SI_1 SI_2	I_1 I_2 I_3
Internally Flawless		Imperfect

Clarity: IF

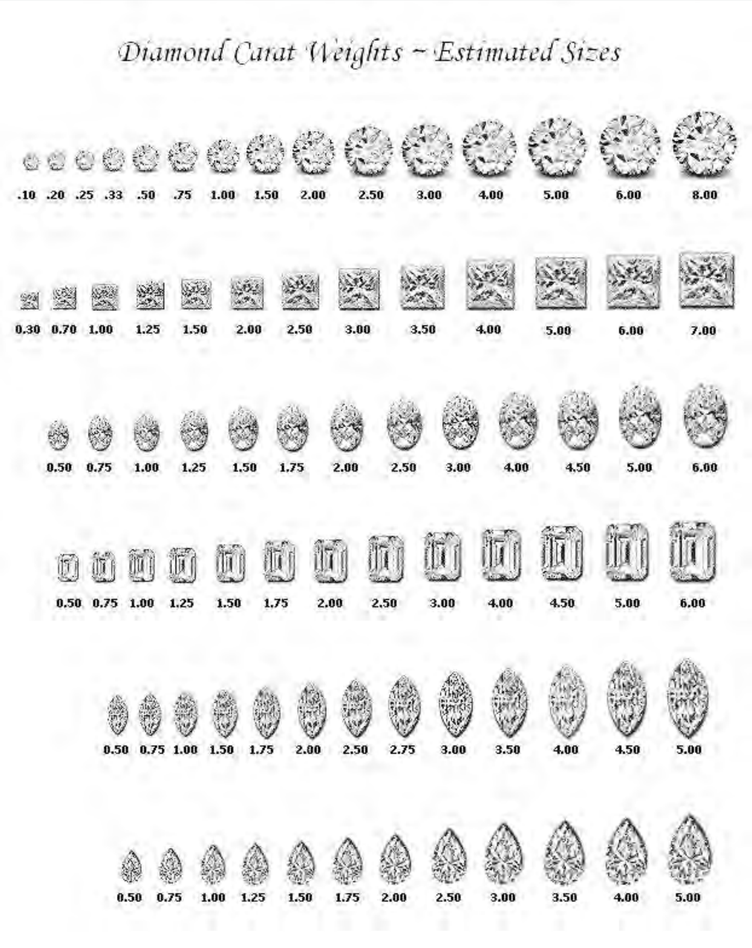
Certification: GIA (Your Best Choice) *

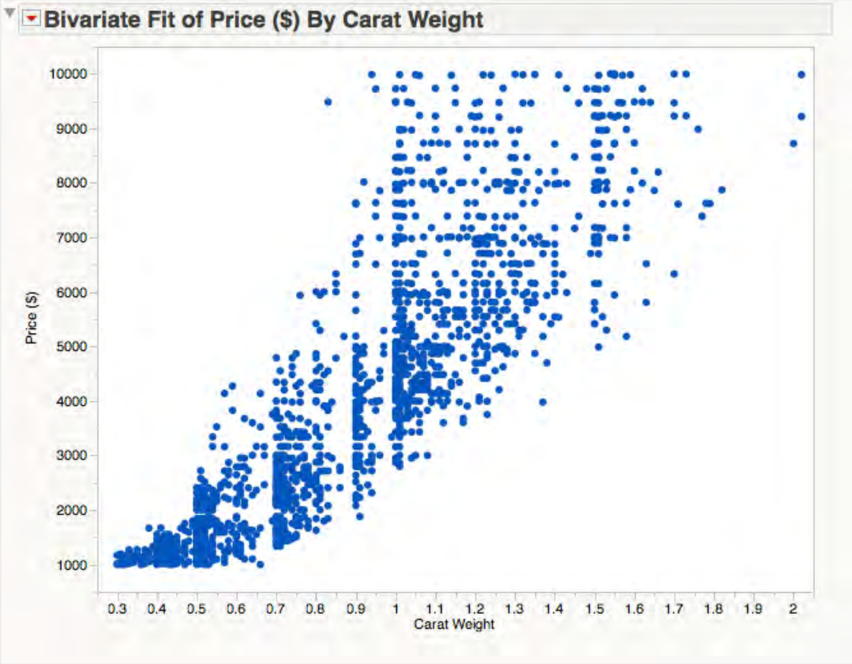
Proportions: Excellent *

* Important price determinant

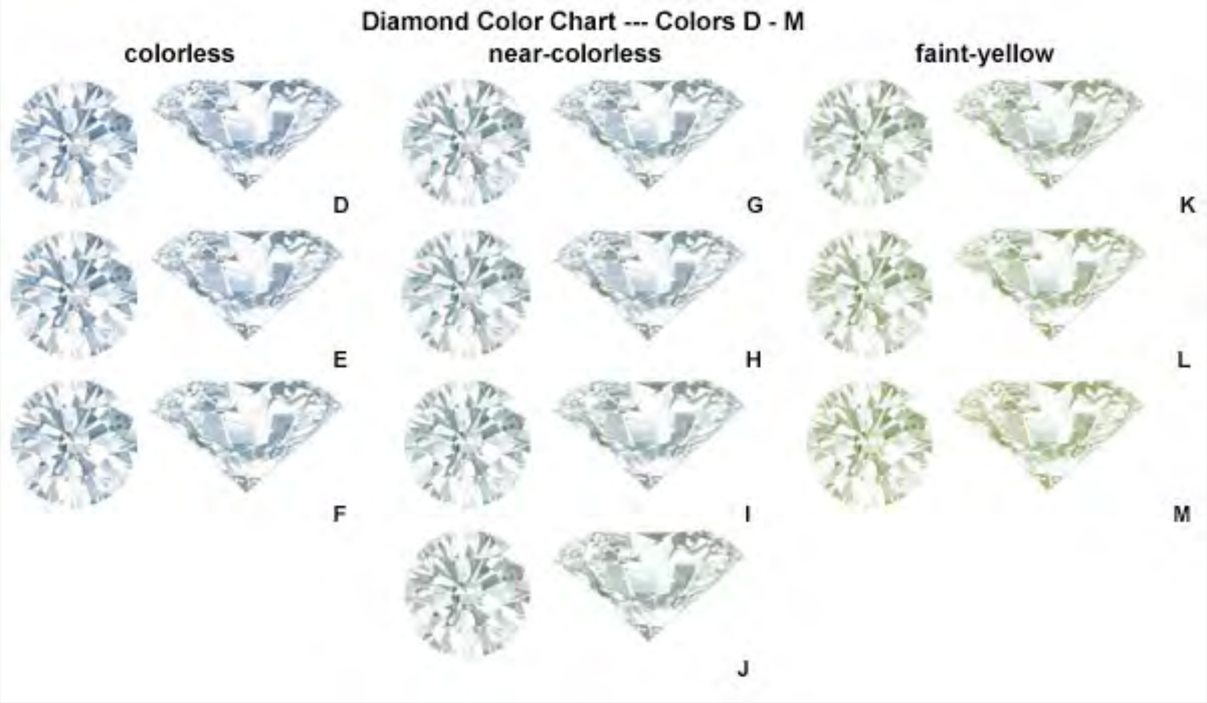
- Carat
- Color
- Cut
- Clarity

CARAT

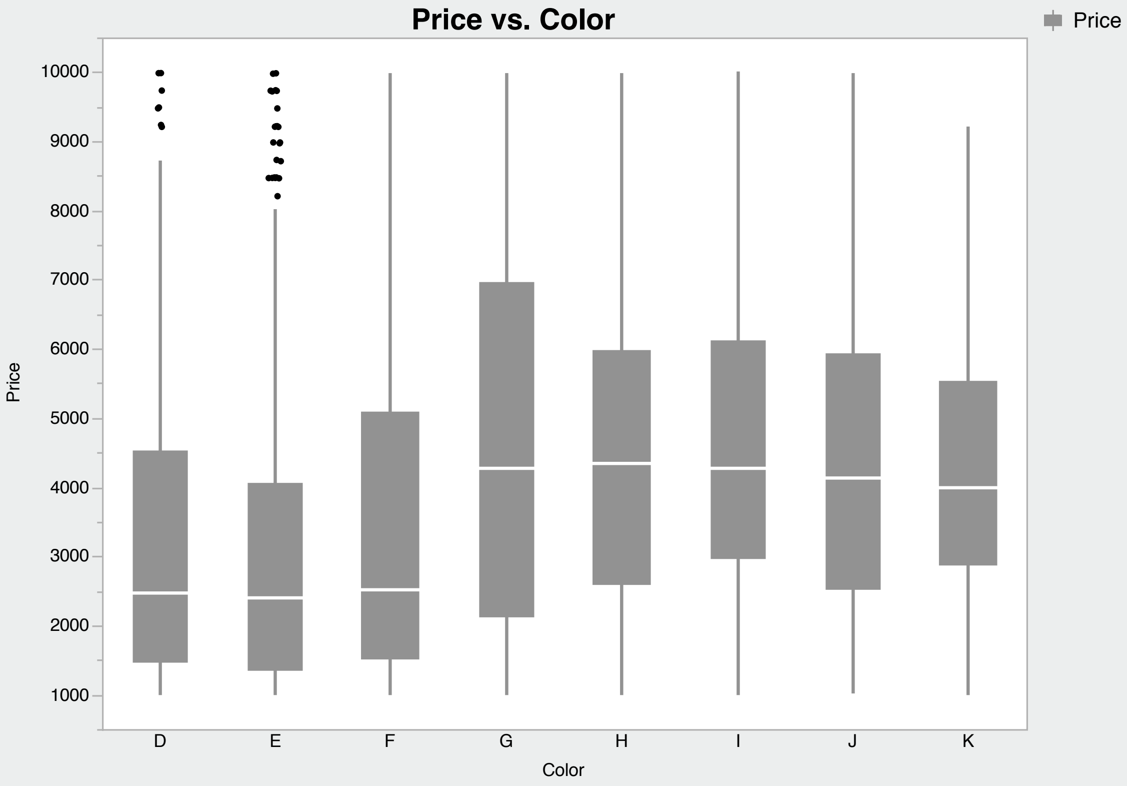

Diamond Carat Weights ~ Estimated Sizes
.10 .20 .25 .33 .50 .75 1.00 1.50 2.00 2.50 3.00 4.00 5.00 6.00 8.00
0.30 0.70 1.00 1.25 1.50 2.00 2.50 3.00 3.50 4.00 5.00 6.00 7.00
0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 4.00 4.50 5.00 6.00
0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 4.00 4.50 5.00 6.00
0.50 0.75 1.00 1.50 1.75 2.00 2.50 2.75 3.00 3.50 4.00 4.50 5.00
0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 3.50 4.00 5.00


Bivariate Fit of Price ($) By Carat Weight
Price ($)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2
Carat Weight

COLOR

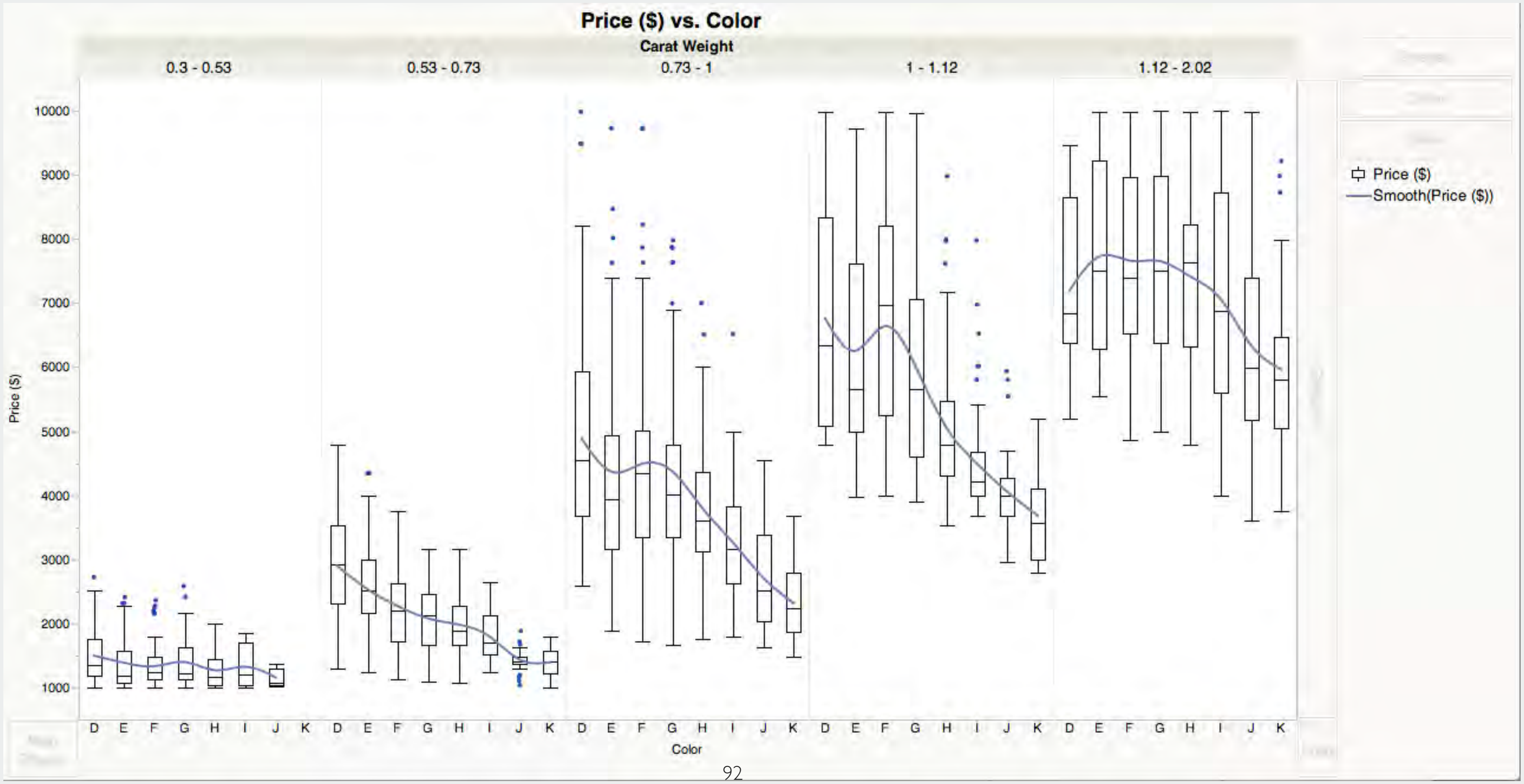

Diamond Color Chart --- Colors D - M
colorless
near-colorless
faint-yellow
D
E
F
G
H
I
J
K
L
M

PRICE VS. COLOR

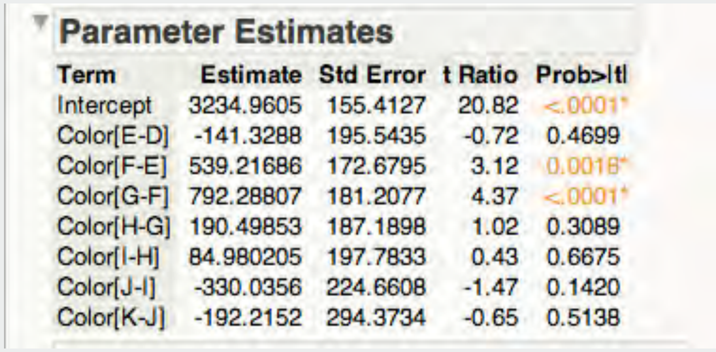


Need to think multivariately!!

ADD SIZE TO THE MODEL

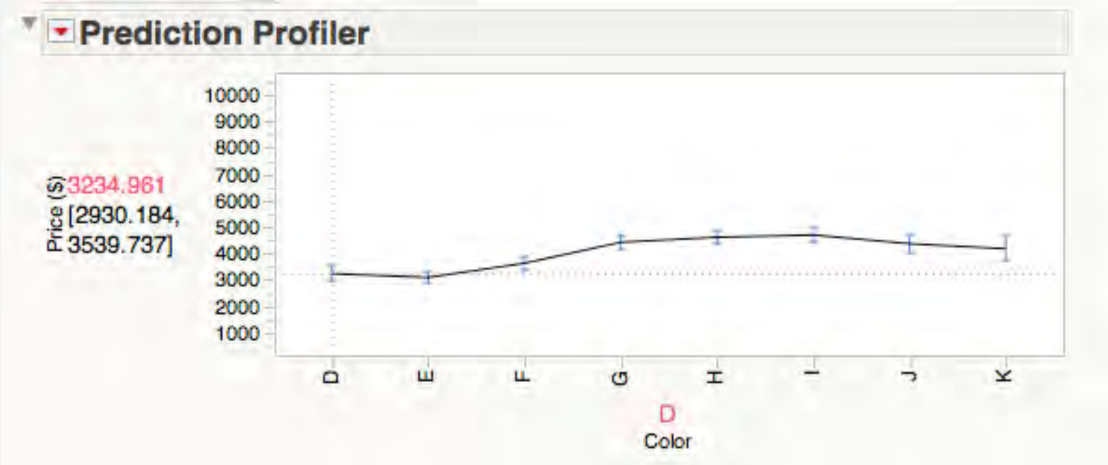


REGRESSION ON COLOR ALONE

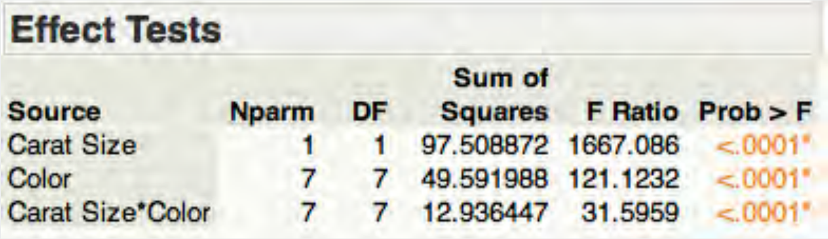


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3234.9605	155.4127	20.82	<.0001*
Color[E-D]	-141.3288	195.5435	-0.72	0.4699
Color[F-E]	539.21686	172.6795	3.12	0.0018*
Color[G-F]	792.28807	181.2077	4.37	<.0001*
Color[H-G]	190.49853	187.1898	1.02	0.3089
Color[I-H]	84.980205	197.7833	0.43	0.6675
Color[J-I]	-330.0356	224.6608	-1.47	0.1420
Color[K-J]	-192.2152	294.3734	-0.65	0.5138

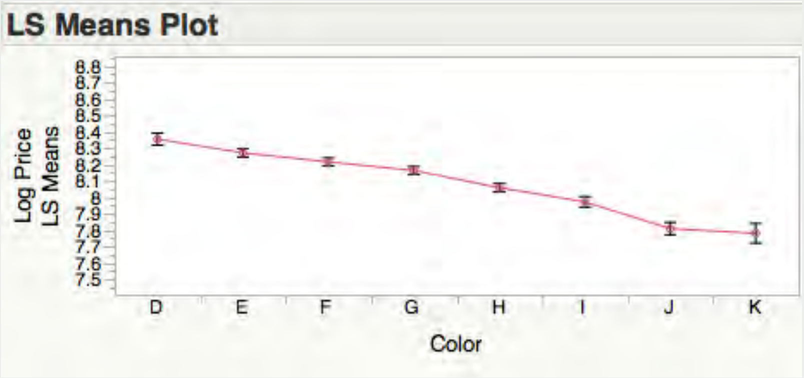


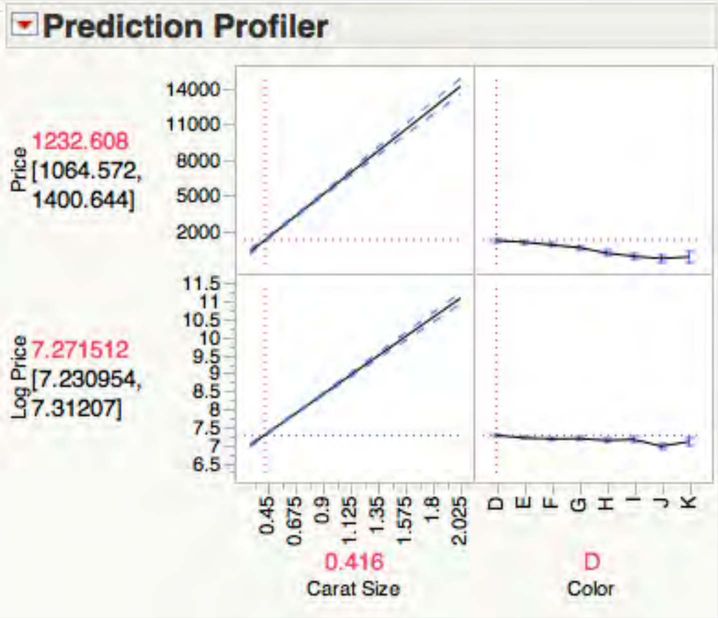
ADD CARAT SIZE TO THE MODEL



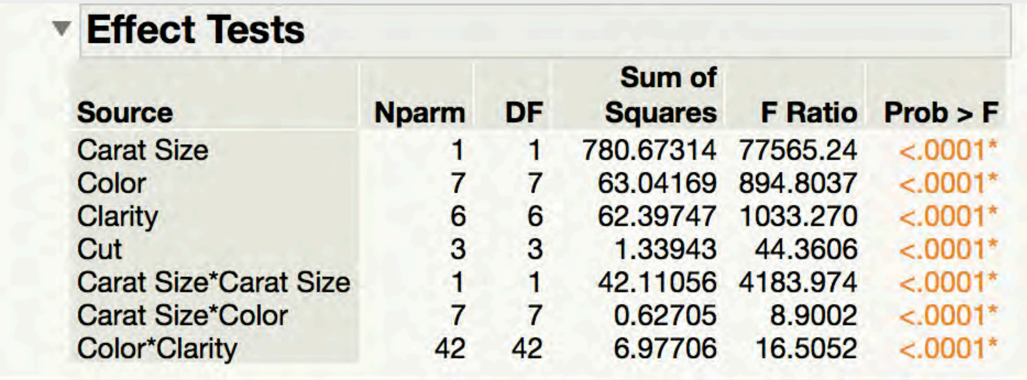
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Carat Size	1	1	97.508872	1667.086	<.0001*
Color	7	7	49.591988	121.1232	<.0001*
Carat Size*Color	7	7	12.936447	31.5959	<.0001*



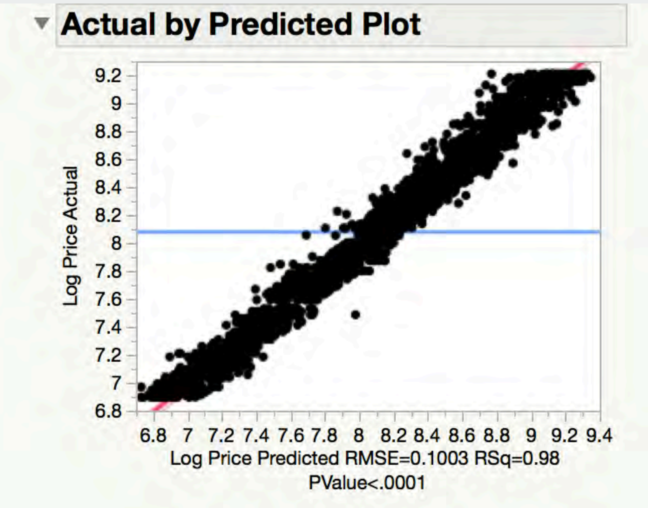


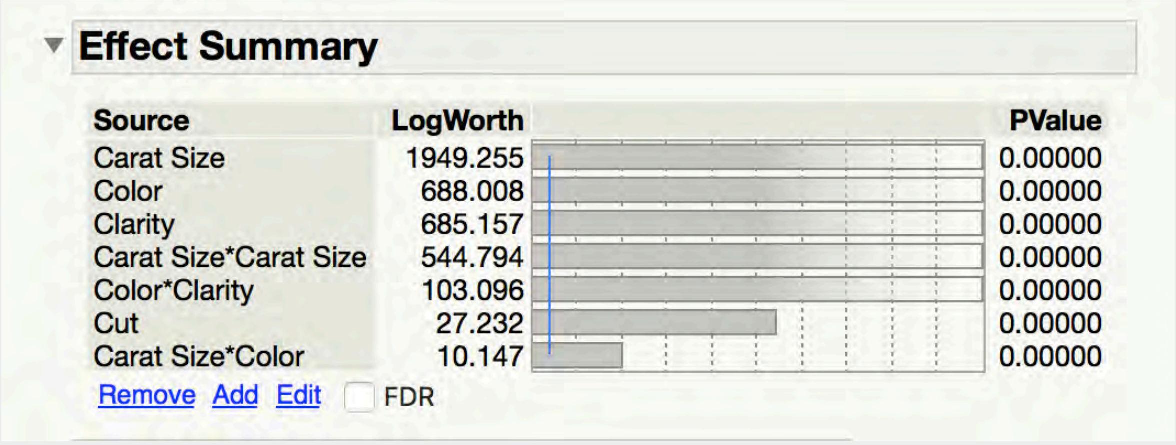
SECOND ORDER MODEL



Effect Tests

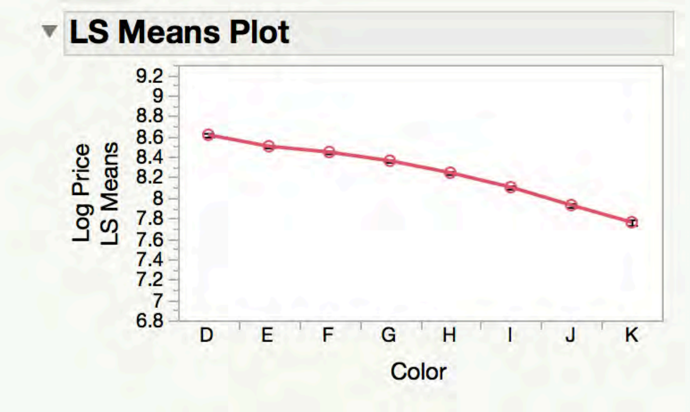
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Carat Size	1	1	780.67314	77565.24	<.0001*
Color	7	7	63.04169	894.8037	<.0001*
Clarity	6	6	62.39747	1033.270	<.0001*
Cut	3	3	1.33943	44.3606	<.0001*
Carat Size*Carat Size	1	1	42.11056	4183.974	<.0001*
Carat Size*Color	7	7	0.62705	8.9002	<.0001*
Color*Clarity	42	42	6.97706	16.5052	<.0001*





Effect Summary

Source	LogWorth	PValue
Carat Size	1949.255	0.00000
Color	688.008	0.00000
Clarity	685.157	0.00000
Carat Size*Carat Size	544.794	0.00000
Color*Clarity	103.096	0.00000
Cut	27.232	0.00000
Carat Size*Color	10.147	0.00000



PROFILE

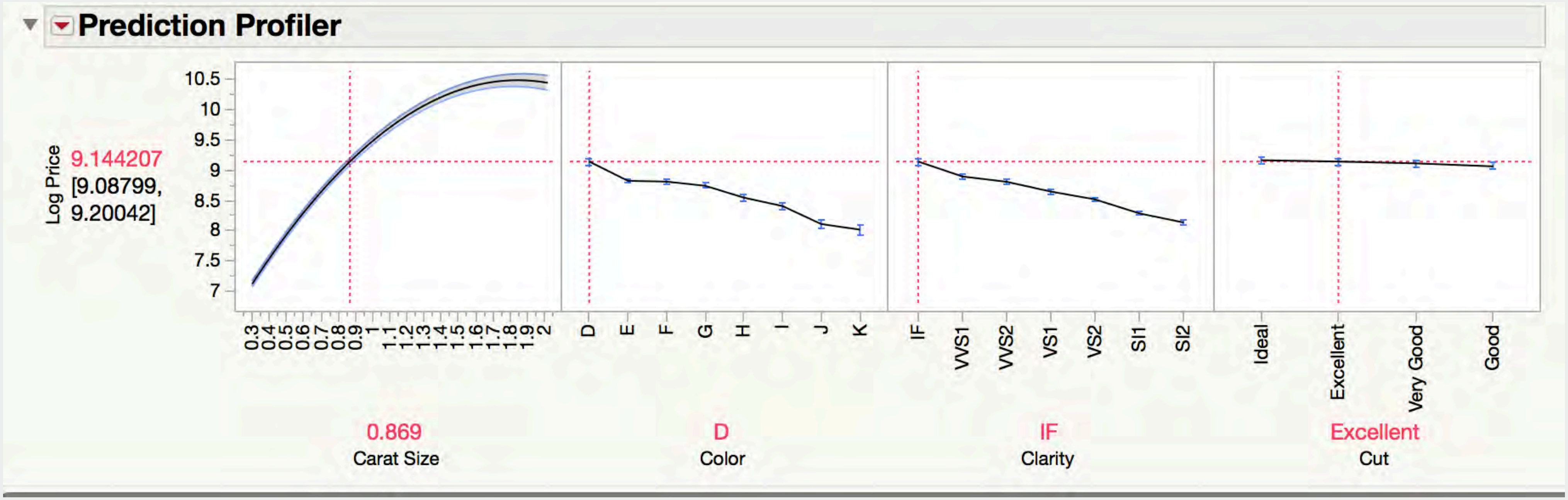
0.869
Carat Size

D
Color

IF
Clarity

Excellent
Cut

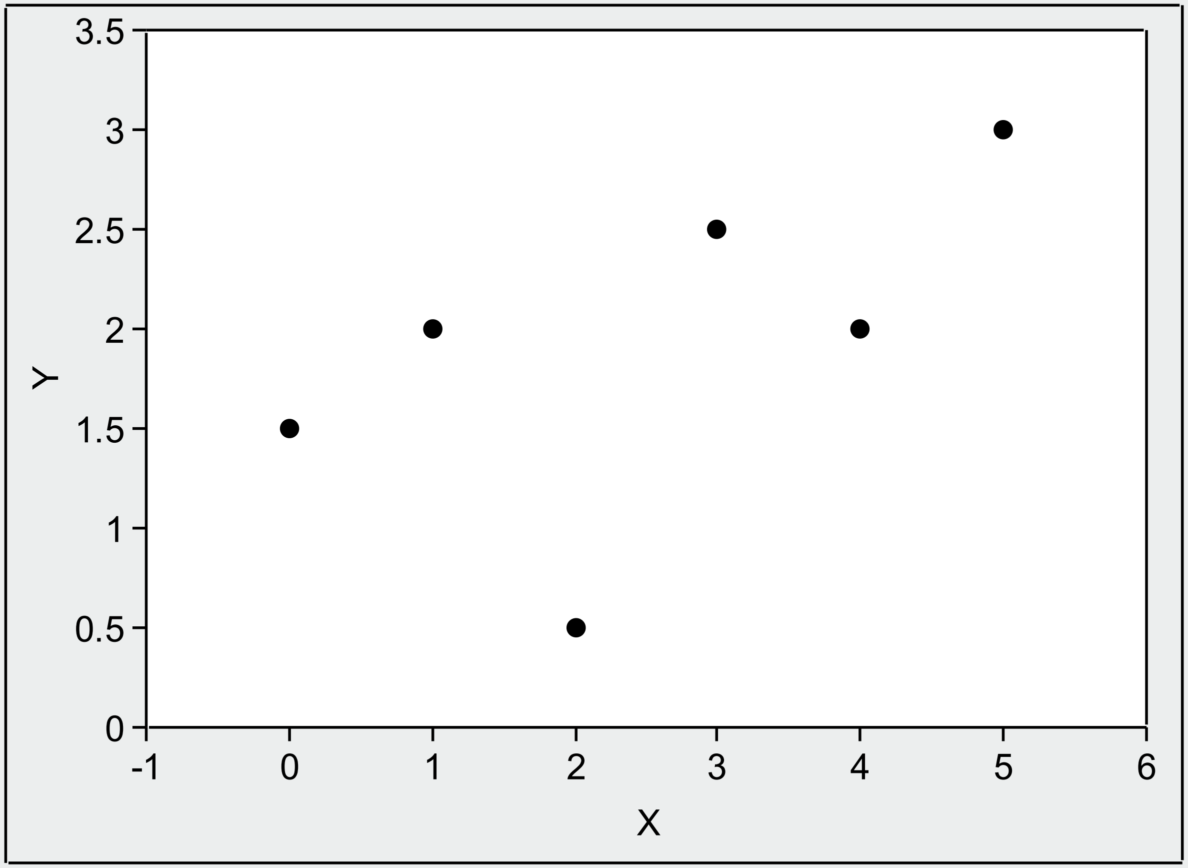
VALUE ORDERING



MODEL BUILDING

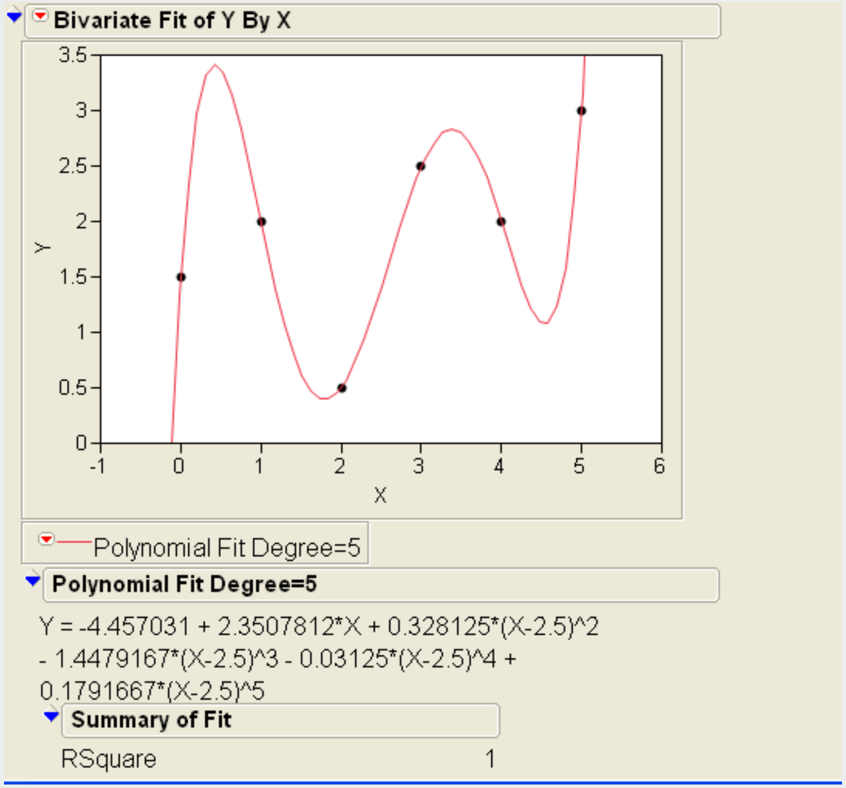
- Model building cycle
  - **Train**
    - On one data set
  - **Test**
    - On another
  - **Evaluate**
    - On a third data set, or at a later time

OVERFITTING



OVERFITTING IN REGRESSION

- Classical overfitting:
  - Fit 5th order polynomial to 6 data points
  - “Great” R^2



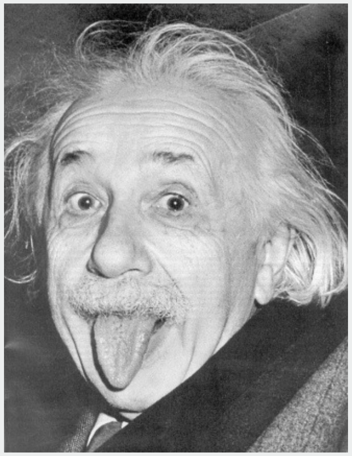
```
Y = -4.457031 + 2.3507812*X + 0.328125*(X-2.5)^2
- 1.4479167*(X-2.5)^3 - 0.03125*(X-2.5)^4 +
0.1791667*(X-2.5)^5
```

```
RSquare                                    1
```

WHY DO MODELS OVERFIT?

- Fitting non-explanatory variables to data
- Overfitting is the result of:
  - Including too many predictor variables
  - Lack of regularizing the model, for example:
    - Decision tree too deep
    - Too many variables in regression
    - Neural net not tuned

HOW TO AVOID OVERFITTING



- Avoiding overfitting is a balancing act – Occam's Razor
  - Fit fewer variables rather than more
  - Have a reason for including a variable (other than it is in the database)
  - Regularize (don't overtrain)
  - Know your field.

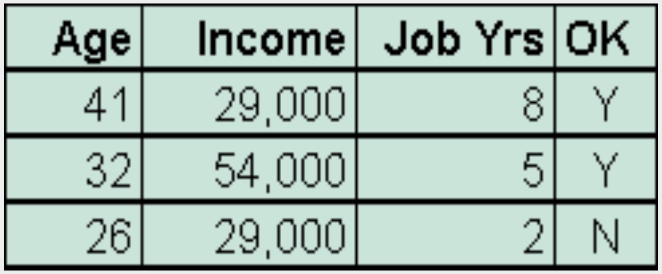
All models should be as simple as possible
but no simpler than necessary
Albert Einstein

VALIDATION

- *Method* : split data into a training data set and a testing data set. A third data set for validation may also be used
- *Advantages:* easy to use and understand. Good estimate of prediction error for reasonably large data sets
- *Disadvantages*: lose up to 20%-30% of data from model building

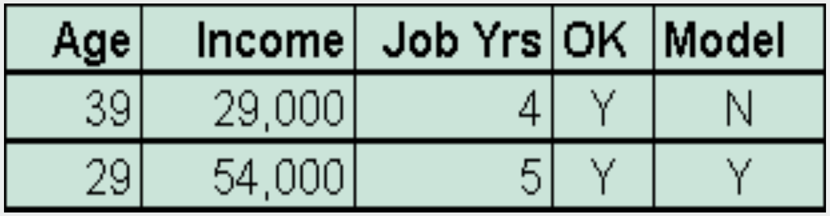
TRAIN VS. TEST

Train



Age	Income	Job Yrs	OK
41	29,000	8	Y
32	54,000	5	Y
26	29,000	2	N

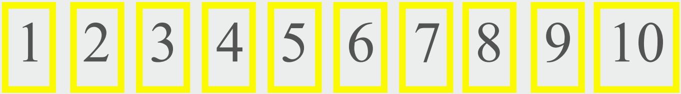
Test



Age	Income	Job Yrs	OK	Model
39	29,000	4	Y	N
29	54,000	5	Y	Y

N-FOLD CROSS VALIDATION

- Divide the data into N equal sized groups and build a model on the data with one group left out.
- Repeat for either systematic or random subgroups
- For very small data sets, N can be 1 (jackknife)



1 2 3 4 5 6 7 8 9 10

INTERNAL VS. EXTERNAL CV

- Internal

  Many methods use internal cross-validation to choose size of model

  - Trees – depth
  - Regression – number of variables
  - Neural Network – size and type of architecture

- External

  Used for model comparison

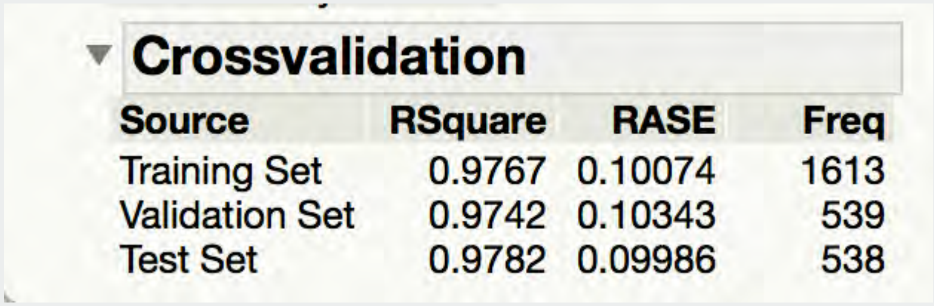
  - Compute R2 on the test set
    - Just the correlation of predicted and actual
  - Compute confusion matrix on the test set

REGRESSION - SUMMARY

- Often works well
- Easy to use
- Theory gives prediction and confidence intervals
- Key is variable selection with interactions and transformations
- Use logistic regression for binary data

CROSS-VALIDATE DIAMONDS

Carat Size	Color	Clarity	Depth	Table	Cut	Report	Price	Log Price	Table-Depth	Table/Depth	Test
0.3	E	VVS1	60	59	Excellent	GIA	1000	6.90775528	-1	0.98333333	0
0.44	E	VS2	61.9	58	Excellent	GIA	1000	6.90775528	-3.9	0.93699515	1
0.31	E	VVS1	61.3	58	Excellent	GIA	1000	6.90775528	-3.3	0.94616639	0
0.66	K	SI1	62.8	57	Excellent	GIA	1000	6.90775528	-5.8	0.90764331	2
0.47	H	VS2	59.1	64	Very Good	GIA	1000	6.90775528	4.9	1.08291032	2
0.4	G	VS1	62	59	Excellent	GIA	1000	6.90775528	-3	0.9516129	2
0.36	D	VS2	61.3	57	Excellent	GIA	1000	6.90775528	-4.3	0.92985318	0



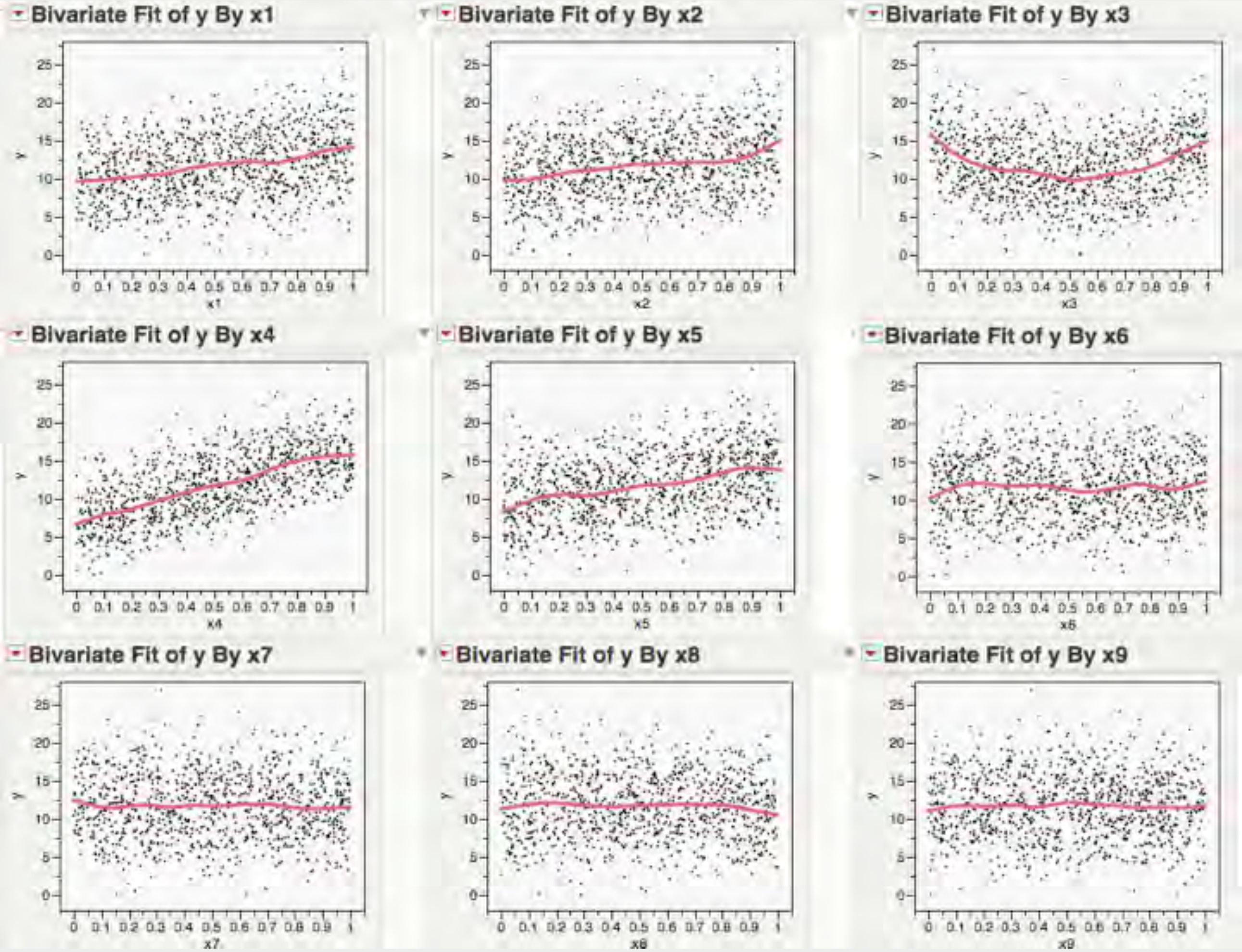
Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.9767	0.10074	1613
Validation Set	0.9742	0.10343	539
Test Set	0.9782	0.09986	538

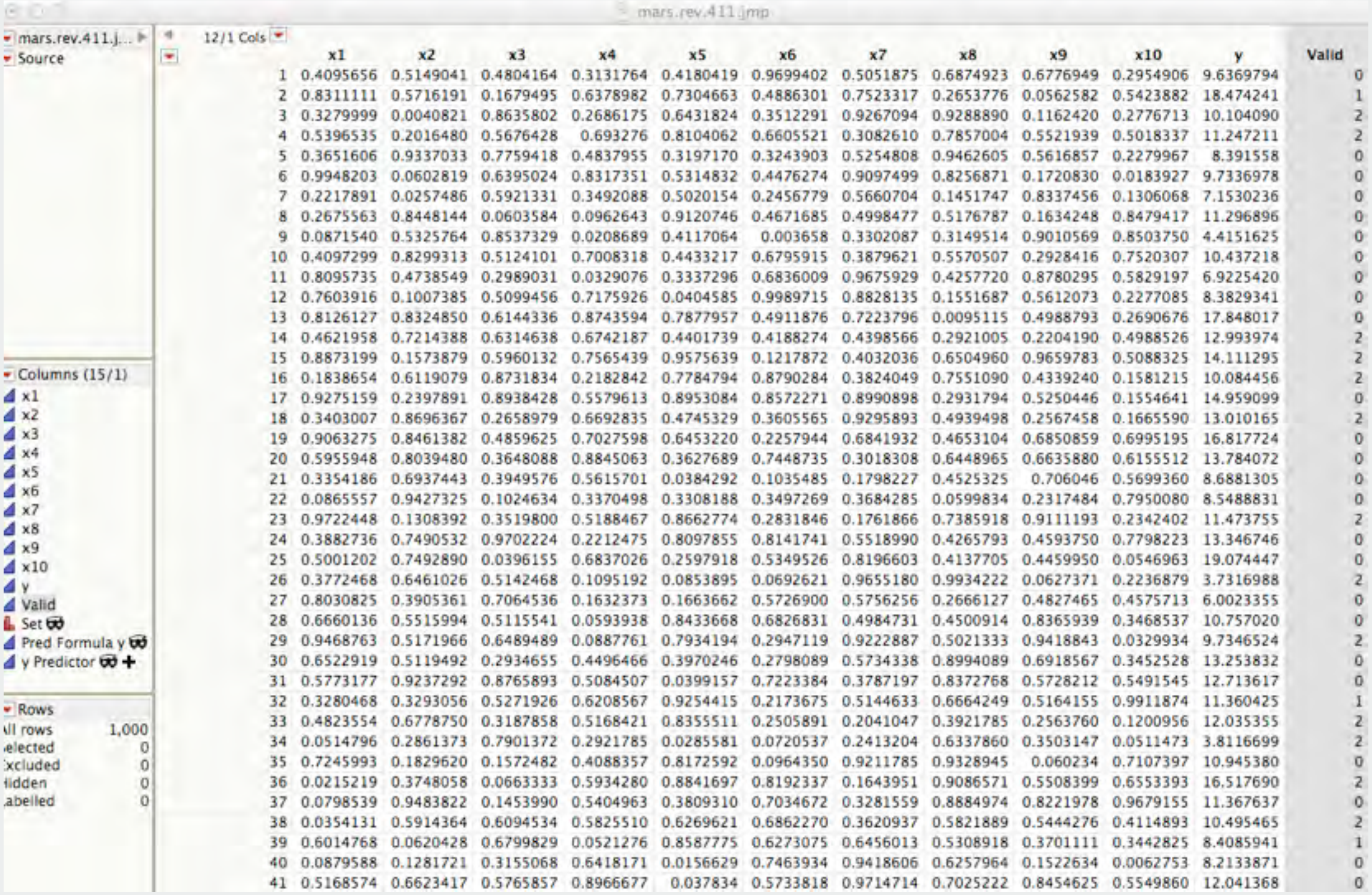
REGULARIZATION

- A model can be built to closely fit the training set but not the real data.
- Symptom: the errors in the training set are reduced, but increased in the test or validation sets.
- Regularization minimizes the residual sum of squares adjusted for model complexity.
- Accomplished by using a smaller decision tree or by pruning it. In neural nets, avoiding over-training.

TOY PROBLEM

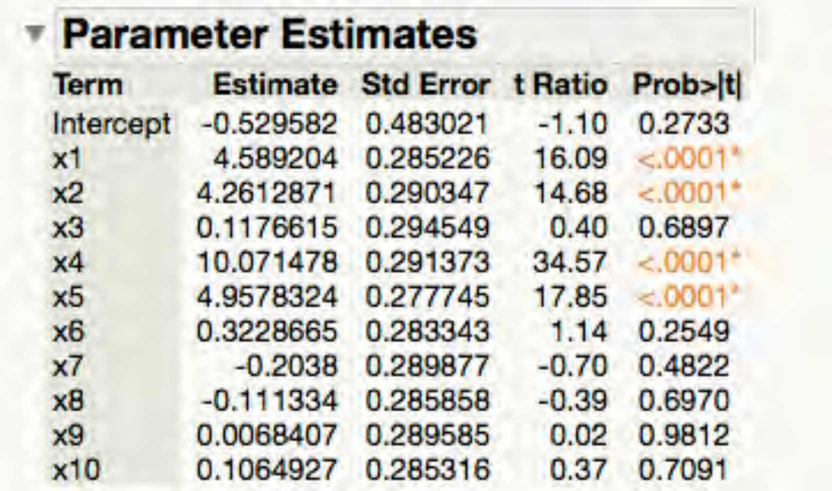

Bivariate Fit of y By x1
Bivariate Fit of y By x2
Bivariate Fit of y By x3
Bivariate Fit of y By x4
Bivariate Fit of y By x5
Bivariate Fit of y By x6
Bivariate Fit of y By x7
Bivariate Fit of y By x8
Bivariate Fit of y By x9
y
0
5
10
15
20
25
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
x1
x2
x3
x4
x5
x6
x7
x8
x9

GENERATING SETS



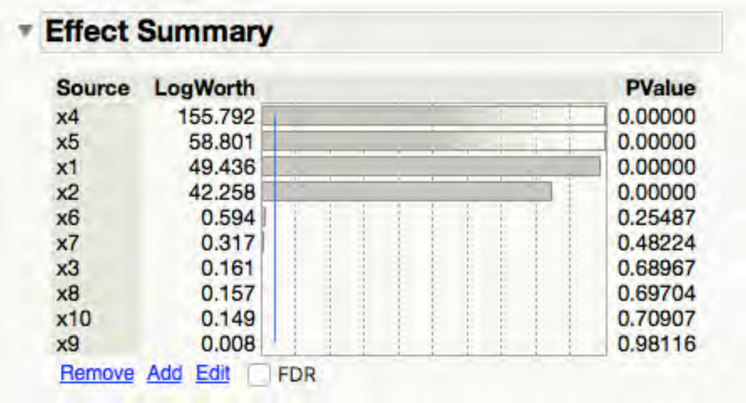
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y	Valid
1	0.4095656	0.5149041	0.4804164	0.3131764	0.4180419	0.9699402	0.5051875	0.6874923	0.6776949	0.2954906	9.6369794	0
2	0.8311111	0.5716191	0.1679495	0.6378982	0.7304663	0.4886301	0.7523317	0.2653776	0.0562582	0.5423882	18.474241	1
3	0.3279999	0.0040821	0.8635802	0.2686175	0.6431824	0.3512291	0.9267094	0.9288890	0.1162420	0.2776713	10.104090	2
4	0.5396535	0.2016480	0.5676428	0.693276	0.8104062	0.6605521	0.3082610	0.7857004	0.5521939	0.5018337	11.247211	2
5	0.3651606	0.9337033	0.7759418	0.4837955	0.3197170	0.3243903	0.5254808	0.9462605	0.5616857	0.2279967	8.391558	0
6	0.9948203	0.0602819	0.6395024	0.8317351	0.5314832	0.4476274	0.9097499	0.8256871	0.1720830	0.0183927	9.7336978	0
7	0.2217891	0.0257486	0.5921331	0.3492088	0.5020154	0.2456779	0.5660704	0.1451747	0.8337456	0.1306068	7.1530236	0
8	0.2675563	0.8448144	0.0603584	0.0962643	0.9120746	0.4671685	0.4998477	0.5176787	0.1634248	0.8479417	11.296896	0
9	0.0871540	0.5325764	0.8537329	0.0208689	0.4117064	0.003658	0.3302087	0.3149514	0.9010569	0.8503750	4.4151625	0
10	0.4097299	0.8299313	0.5124101	0.7008318	0.4433217	0.6795915	0.3879621	0.5570507	0.2928416	0.7520307	10.437218	0
11	0.8095735	0.4738549	0.2989031	0.0329076	0.3337296	0.6836009	0.9675929	0.4257720	0.8780295	0.5829197	6.9225420	0
12	0.7603916	0.1007385	0.5099456	0.7175926	0.0404585	0.9989715	0.8828135	0.1551687	0.5612073	0.2277085	8.3829341	0
13	0.8126127	0.8324850	0.6144336	0.8743594	0.7877957	0.4911876	0.7223796	0.0095115	0.4988793	0.2690676	17.848017	0
14	0.4621958	0.7214388	0.6314638	0.6742187	0.4401739	0.4188274	0.4398566	0.2921005	0.2204190	0.4988526	12.993974	2
15	0.8873199	0.1573879	0.5960132	0.7565439	0.9575639	0.1217872	0.4032036	0.6504960	0.9659783	0.5088325	14.111295	2
16	0.1838654	0.6119079	0.8731834	0.2182842	0.7784794	0.8790284	0.3824049	0.7551090	0.4339240	0.1581215	10.084456	2
17	0.9275159	0.2397891	0.8938428	0.5579613	0.8953084	0.8572271	0.8990898	0.2931794	0.5250446	0.1554641	14.959099	0
18	0.3403007	0.8696367	0.2658979	0.6692835	0.4745329	0.3605565	0.9295893	0.4939498	0.2567458	0.1665590	13.010165	2
19	0.9063275	0.8461382	0.4859625	0.7027598	0.6453220	0.2257944	0.6841932	0.4653104	0.6850859	0.6995195	16.817724	0
20	0.5955948	0.8039480	0.3648088	0.8845063	0.3627689	0.7448735	0.3018308	0.6448965	0.6635880	0.6155512	13.784072	0
21	0.3354186	0.6937443	0.3949576	0.5615701	0.0384292	0.1035485	0.1798227	0.4525325	0.706046	0.5699360	8.6881305	0
22	0.0865557	0.9427325	0.1024634	0.3370498	0.3308188	0.3497269	0.3684285	0.0599834	0.2317484	0.7950080	8.5488831	0
23	0.9722448	0.1308392	0.3519800	0.5188467	0.8662774	0.2831846	0.1761866	0.7385918	0.9111193	0.2342402	11.473755	2
24	0.3882736	0.7490532	0.9702224	0.2212475	0.8097855	0.8141741	0.5518990	0.4265793	0.4593750	0.7798223	13.346746	0
25	0.5001202	0.7492890	0.0396155	0.6837026	0.2597918	0.5349526	0.8196603	0.4137705	0.4459950	0.0546963	19.074447	0
26	0.3772468	0.6461026	0.5142468	0.1095192	0.0853895	0.0692621	0.9655180	0.9934222	0.0627371	0.2236879	3.7316988	2
27	0.8030825	0.3905361	0.7064536	0.1632373	0.1663662	0.5726900	0.5756256	0.2666127	0.4827465	0.4575713	6.0023355	0
28	0.6660136	0.5515994	0.5115541	0.0593938	0.8433668	0.6826831	0.4984731	0.4500914	0.8365939	0.3468537	10.757020	0
29	0.9468763	0.5171966	0.6489489	0.0887761	0.7934194	0.2947119	0.9222887	0.5021333	0.9418843	0.0329934	9.7346524	2
30	0.6522919	0.5119492	0.2934655	0.4496466	0.3970246	0.2798089	0.5734338	0.8994089	0.6918567	0.3452528	13.253832	0
31	0.5773177	0.9237292	0.8765893	0.5084507	0.0399157	0.7223384	0.3787197	0.8372768	0.5728212	0.5491545	12.713617	0
32	0.3280468	0.3293056	0.5271926	0.6208567	0.9254415	0.2173675	0.5144633	0.6664249	0.5164155	0.9911874	11.360425	1
33	0.4823554	0.6778750	0.3187858	0.5168421	0.8355511	0.2505891	0.2041047	0.3921785	0.2563760	0.1200956	12.035355	2
34	0.0514796	0.2861373	0.7901372	0.2921785	0.0285581	0.0720537	0.2413204	0.6337860	0.3503147	0.0511473	3.8116699	2
35	0.7245993	0.1829620	0.1572482	0.4088357	0.8172592	0.0964350	0.9211785	0.9328945	0.060234	0.7107397	10.945380	0
36	0.0215219	0.3748058	0.0663333	0.5934280	0.8841697	0.8192337	0.1643951	0.9086571	0.5508399	0.6553393	16.517690	2
37	0.0798539	0.9483822	0.1453990	0.5404963	0.3809310	0.7034672	0.3281559	0.8884974	0.8221978	0.9679155	11.367637	0
38	0.0354131	0.5914364	0.6094534	0.5825510	0.6269621	0.6862270	0.3620937	0.5821889	0.5444276	0.4114893	10.495465	2
39	0.6014768	0.0620428	0.6799829	0.0521276	0.8587775	0.6273075	0.6456013	0.5308918	0.3701111	0.3442825	8.4085941	1
40	0.0879588	0.1281721	0.3155068	0.6418171	0.0156629	0.7463934	0.9418606	0.6257964	0.1522634	0.0062753	8.2133871	0
41	0.5168574	0.6623417	0.5765857	0.8966677	0.037834	0.5733818	0.9714714	0.7025222	0.8454625	0.5549860	12.041368	0

MULTIPLE REGRESSION



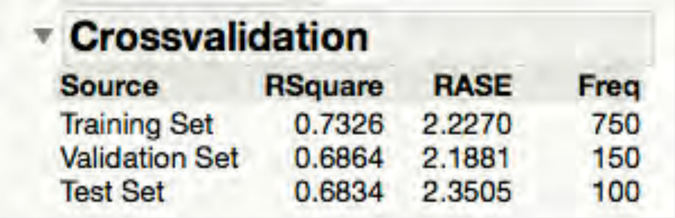
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.529582	0.483021	-1.10	0.2733
x1	4.589204	0.285226	16.09	<.0001*
x2	4.2612871	0.290347	14.68	<.0001*
x3	0.1176615	0.294549	0.40	0.6897
x4	10.071478	0.291373	34.57	<.0001*
x5	4.9578324	0.277745	17.85	<.0001*
x6	0.3228665	0.283343	1.14	0.2549
x7	-0.2038	0.289877	-0.70	0.4822
x8	-0.111334	0.285858	-0.39	0.6970
x9	0.0068407	0.289585	0.02	0.9812
x10	0.1064927	0.285316	0.37	0.7091



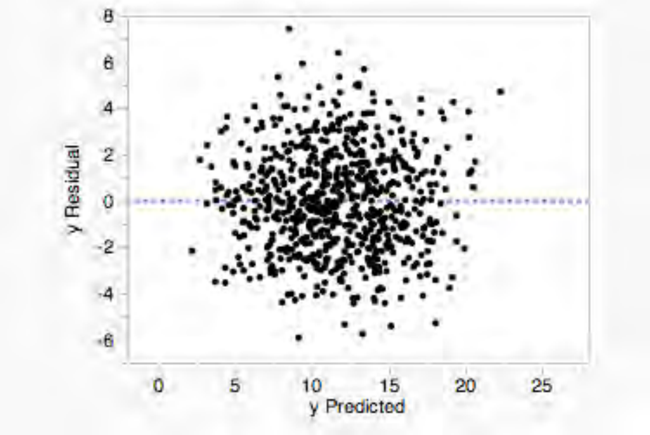
Effect Summary

Source	LogWorth	PValue
x4	155.792	0.00000
x5	58.801	0.00000
x1	49.436	0.00000
x2	42.258	0.00000
x6	0.594	0.25487
x7	0.317	0.48224
x3	0.161	0.68967
x8	0.157	0.69704
x10	0.149	0.70907
x9	0.008	0.98116



Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7326	2.2270	750
Validation Set	0.6864	2.1881	150
Test Set	0.6834	2.3505	100



IN **R**

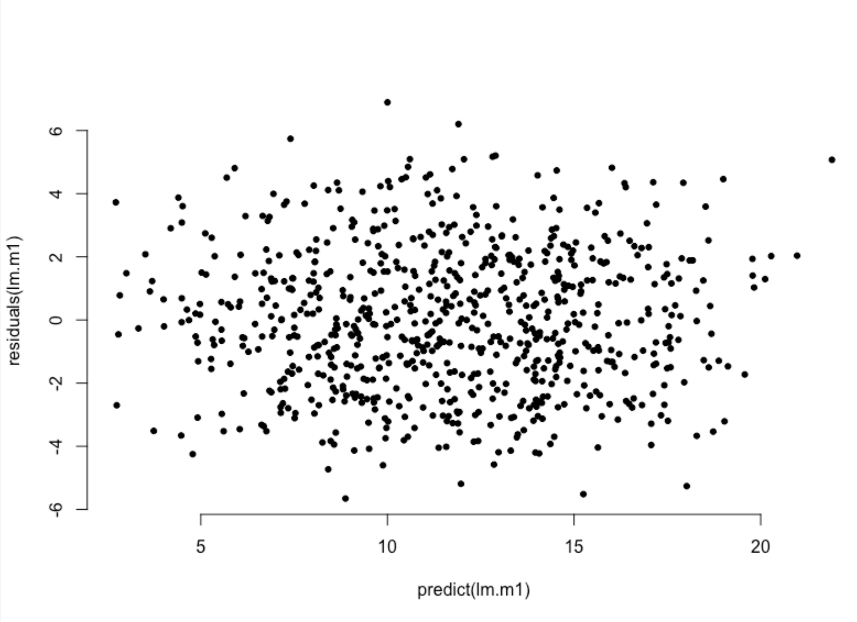
```
> summary(lm.m1)

Call:
lm(formula = y ~ . - Test, data = subset(marsd, Test == 0))

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6551 -1.5676 -0.1085  1.5394  6.8917 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.529582   0.483021  -1.096    0.273    
x1           4.589204   0.285226  16.090   <2e-16 ***
x2           4.261287   0.290347  14.677   <2e-16 ***
x3           0.117661   0.294549   0.399    0.690    
x4          10.071478   0.291373  34.566   <2e-16 ***
x5           4.957832   0.277745  17.850   <2e-16 ***
x6           0.322867   0.283343   1.139    0.255    
x7          -0.203800   0.289877  -0.703    0.482    
x8          -0.111334   0.285858  -0.389    0.697    
x9           0.006841   0.289585   0.024    0.981    
x10          0.106493   0.285316   0.373    0.709    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.244 on 739 degrees of freedom
Multiple R-squared:  0.7326,  Adjusted R-squared:  0.729 
F-statistic: 202.5 on 10 and 739 DF,  p-value: < 2.2e-16
```



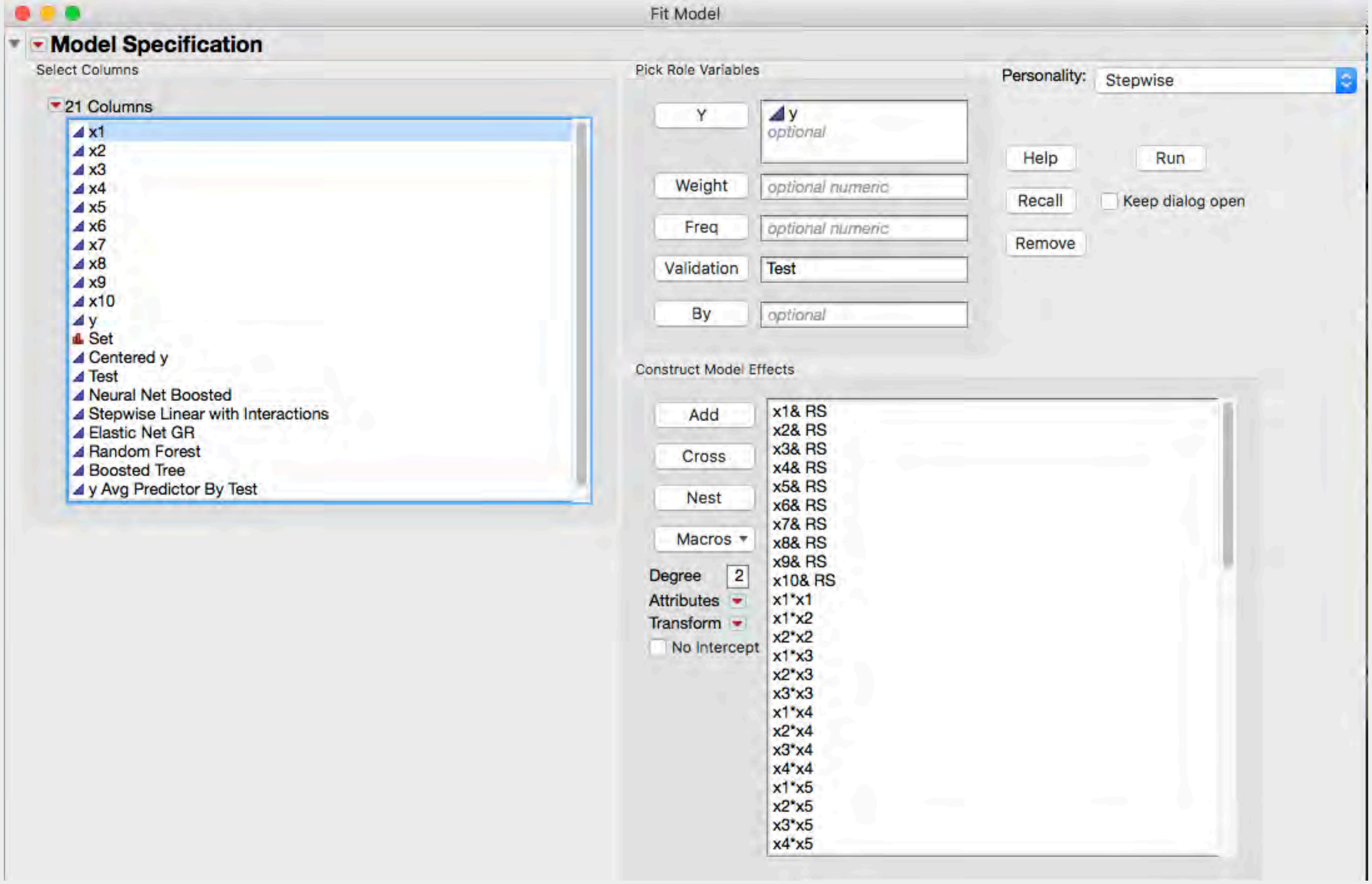
```
plot(residuals(lm.m1)~predict(lm.m1),pch=19,bty="n",cex=.7)
```

CROSSVALIDATION IN **R**

```
> cor(predict(lm.m1,new=subset(marsd,Test==1)),marsd[marsd$Test==1,11])^2
[1] 0.6932686
```

```
> cor(predict(lm.m1,new=subset(marsd,Test==2)),marsd[marsd$Test==2,11])^2
[1] 0.6843684
```

SECOND ORDER MODEL



Personality: Stepwise

Y: y

Weight: optional numeric

Freq: optional numeric

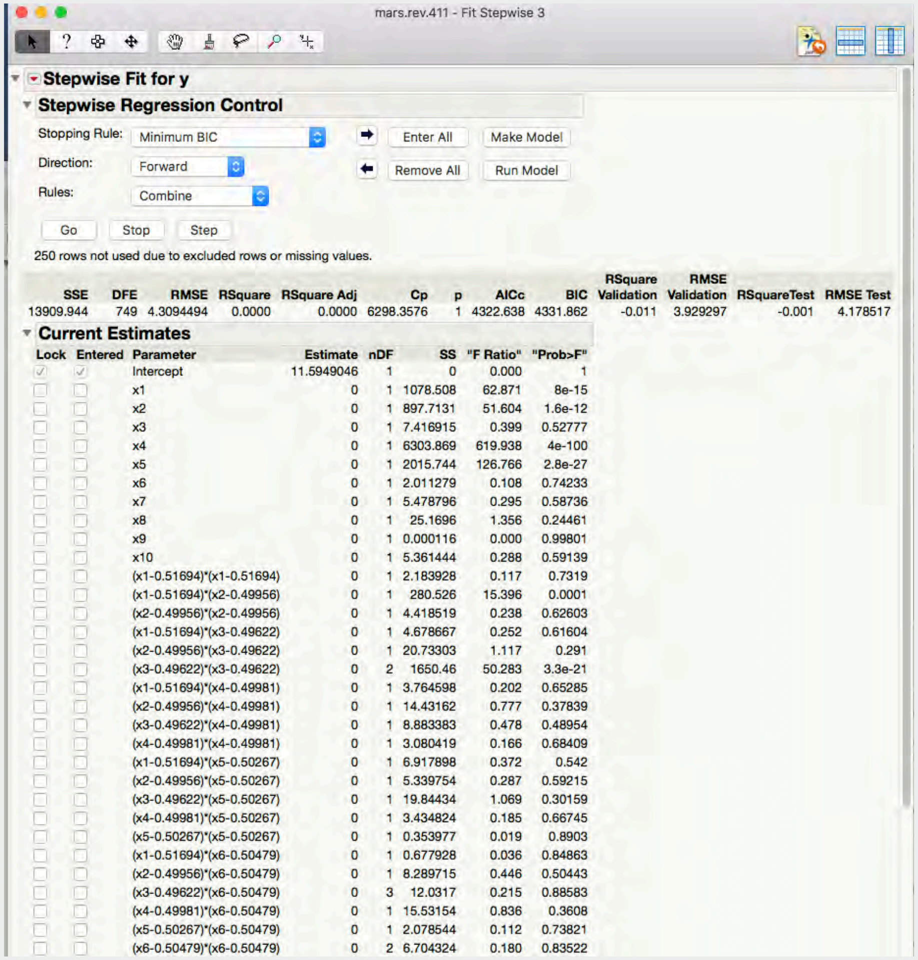
Validation: Test

By: optional

Degree: 2

x1& RS
x2& RS
x3& RS
x4& RS
x5& RS
x6& RS
x7& RS
x8& RS
x9& RS
x10& RS
x1*x1
x1*x2
x2*x2
x1*x3
x2*x3
x3*x3
x1*x4
x2*x4
x3*x4
x4*x4
x1*x5
x2*x5
x3*x5
x4*x5

STEPWISE



Stepwise Fit for y

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Forward

Rules: Combine

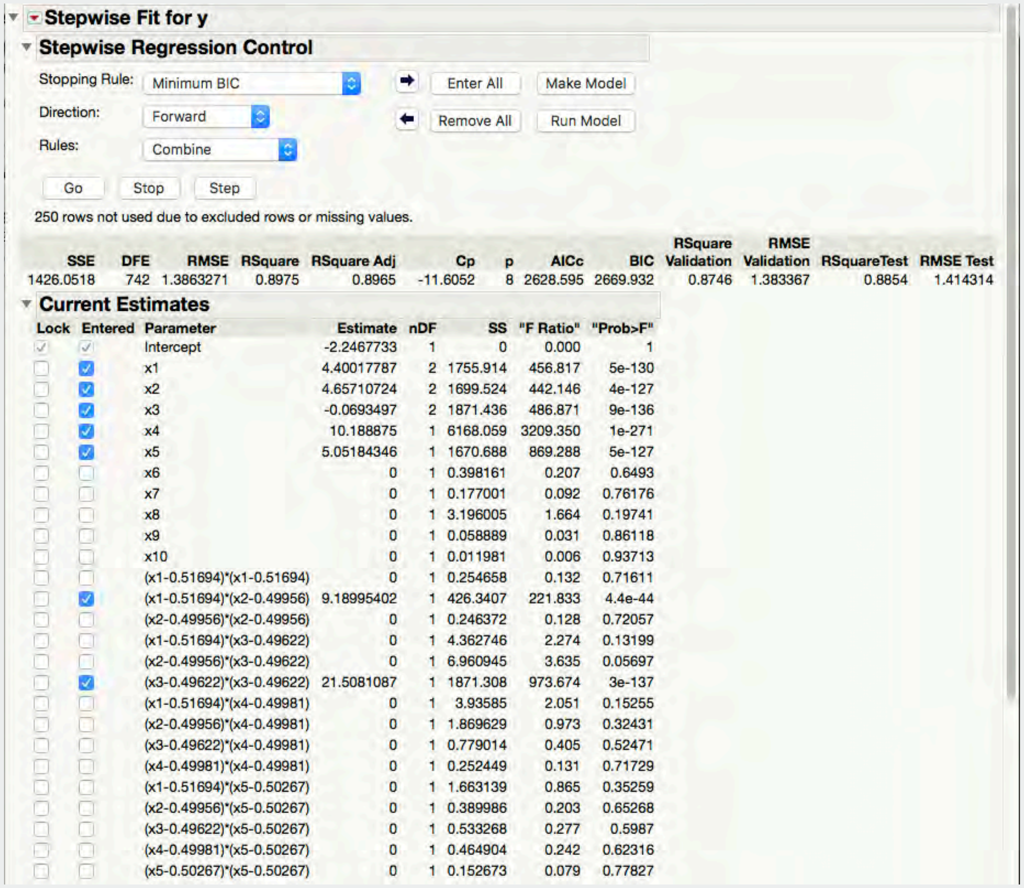
250 rows not used due to excluded rows or missing values.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	RSquare Validation	RMSE Validation	RSquareTest	RMSE Test
13909.944	749	4.3094494	0.0000	0.0000	6298.3576	1	4322.638	4331.862	-0.011	3.929297	-0.001	4.178517

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	11.5949046	1	0	0.000	1
☐	☐	x1	0	1	1078.508	62.871	8e-15
☐	☐	x2	0	1	897.7131	51.604	1.6e-12
☐	☐	x3	0	1	7.416915	0.399	0.52777
☐	☐	x4	0	1	6303.869	619.938	4e-100
☐	☐	x5	0	1	2015.744	126.766	2.8e-27
☐	☐	x6	0	1	2.011279	0.108	0.74233
☐	☐	x7	0	1	5.478796	0.295	0.58736
☐	☐	x8	0	1	25.1696	1.356	0.24461
☐	☐	x9	0	1	0.000116	0.000	0.99801
☐	☐	x10	0	1	5.361444	0.288	0.59139
☐	☐	(x1-0.51694)*(x1-0.51694)	0	1	2.183928	0.117	0.7319
☐	☐	(x1-0.51694)*(x2-0.49956)	0	1	280.526	15.396	0.0001
☐	☐	(x2-0.49956)*(x2-0.49956)	0	1	4.418519	0.238	0.62603
☐	☐	(x1-0.51694)*(x3-0.49622)	0	1	4.678667	0.252	0.61604
☐	☐	(x2-0.49956)*(x3-0.49622)	0	1	20.73303	1.117	0.291
☐	☐	(x3-0.49622)*(x3-0.49622)	0	2	1650.46	50.283	3.3e-21
☐	☐	(x1-0.51694)*(x4-0.49981)	0	1	3.764598	0.202	0.65285
☐	☐	(x2-0.49956)*(x4-0.49981)	0	1	14.43162	0.777	0.37839
☐	☐	(x3-0.49622)*(x4-0.49981)	0	1	8.883383	0.478	0.48954
☐	☐	(x4-0.49981)*(x4-0.49981)	0	1	3.080419	0.166	0.68409
☐	☐	(x1-0.51694)*(x5-0.50267)	0	1	6.917898	0.372	0.542
☐	☐	(x2-0.49956)*(x5-0.50267)	0	1	5.339754	0.287	0.59215
☐	☐	(x3-0.49622)*(x5-0.50267)	0	1	19.84434	1.069	0.30159
☐	☐	(x4-0.49981)*(x5-0.50267)	0	1	3.434824	0.185	0.66745
☐	☐	(x5-0.50267)*(x5-0.50267)	0	1	0.353977	0.019	0.8903
☐	☐	(x1-0.51694)*(x6-0.50479)	0	1	0.677928	0.036	0.84863
☐	☐	(x2-0.49956)*(x6-0.50479)	0	1	8.289715	0.446	0.50443
☐	☐	(x3-0.49622)*(x6-0.50479)	0	3	12.0317	0.215	0.88583
☐	☐	(x4-0.49981)*(x6-0.50479)	0	1	15.53154	0.836	0.3608
☐	☐	(x5-0.50267)*(x6-0.50479)	0	1	2.078544	0.112	0.73821
☐	☐	(x6-0.50479)*(x6-0.50479)	0	2	6.704324	0.180	0.83522

SECOND ORDER MODEL



Stepwise Fit for y

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Forward

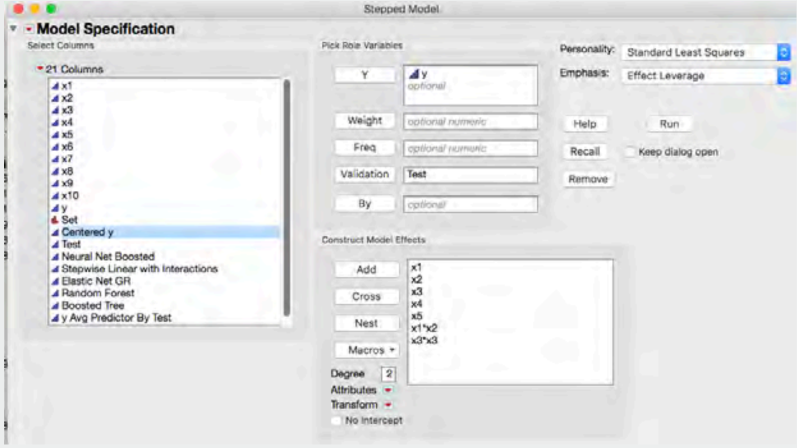
Rules: Combine

250 rows not used due to excluded rows or missing values.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	RSquare Validation	RMSE Validation	RSquareTest	RMSE Test
1426.0518	742	1.3863271	0.8975	0.8965	-11.6052	8	2628.595	2669.932	0.8746	1.383367	0.8854	1.414314

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	-2.2467733	1	0	0.000	1
☐	☑	x1	4.40017787	2	1755.914	456.817	5e-130
☐	☑	x2	4.65710724	2	1699.524	442.146	4e-127
☐	☑	x3	-0.0693497	2	1871.436	486.871	9e-136
☐	☑	x4	10.188875	1	6168.059	3209.350	1e-271
☐	☑	x5	5.05184346	1	1670.688	869.288	5e-127
☐	☐	x6	0	1	0.398161	0.207	0.6493
☐	☐	x7	0	1	0.177001	0.092	0.76176
☐	☐	x8	0	1	3.196005	1.664	0.19741
☐	☐	x9	0	1	0.058889	0.031	0.86118
☐	☐	x10	0	1	0.011981	0.006	0.93713
☐	☐	(x1-0.51694)*(x1-0.51694)	0	1	0.254658	0.132	0.71611
☐	☑	(x1-0.51694)*(x2-0.49956)	9.18995402	1	426.3407	221.833	4.4e-44
☐	☐	(x2-0.49956)*(x2-0.49956)	0	1	0.246372	0.128	0.72057
☐	☐	(x1-0.51694)*(x3-0.49622)	0	1	4.362746	2.274	0.13199
☐	☐	(x2-0.49956)*(x3-0.49622)	0	1	6.960945	3.635	0.05697
☐	☑	(x3-0.49622)*(x3-0.49622)	21.5081087	1	1871.308	973.674	3e-137
☐	☐	(x1-0.51694)*(x4-0.49981)	0	1	3.93585	2.051	0.15255
☐	☐	(x2-0.49956)*(x4-0.49981)	0	1	1.869629	0.973	0.32431
☐	☐	(x3-0.49622)*(x4-0.49981)	0	1	0.779014	0.405	0.52471
☐	☐	(x4-0.49981)*(x4-0.49981)	0	1	0.252449	0.131	0.71729
☐	☐	(x1-0.51694)*(x5-0.50267)	0	1	1.663139	0.865	0.35259
☐	☐	(x2-0.49956)*(x5-0.50267)	0	1	0.389986	0.203	0.65268
☐	☐	(x3-0.49622)*(x5-0.50267)	0	1	0.533268	0.277	0.5987
☐	☐	(x4-0.49981)*(x5-0.50267)	0	1	0.464904	0.242	0.62316
☐	☐	(x5-0.50267)*(x5-0.50267)	0	1	0.152673	0.079	0.77827



Stepped Model

Model Specification

Select Columns

21 Columns
- x1
- x2
- x3
- x4
- x5
- x6
- x7
- x8
- x9
- x10
- y
- Set
- Centered y
- Test
- Neural Net Boosted
- Stepwise Linear with Interactions
- Elastic Net GR
- Random Forest
- Boosted Tree
- y Avg Predictor By Test

Pick Role Variables

Y: y

Weight: optional numeric

Freq: optional numeric

Validation: Test

By: optional

Personality: Standard Least Squares

Emphasis: Effect Leverage

☐ Keep dialog open

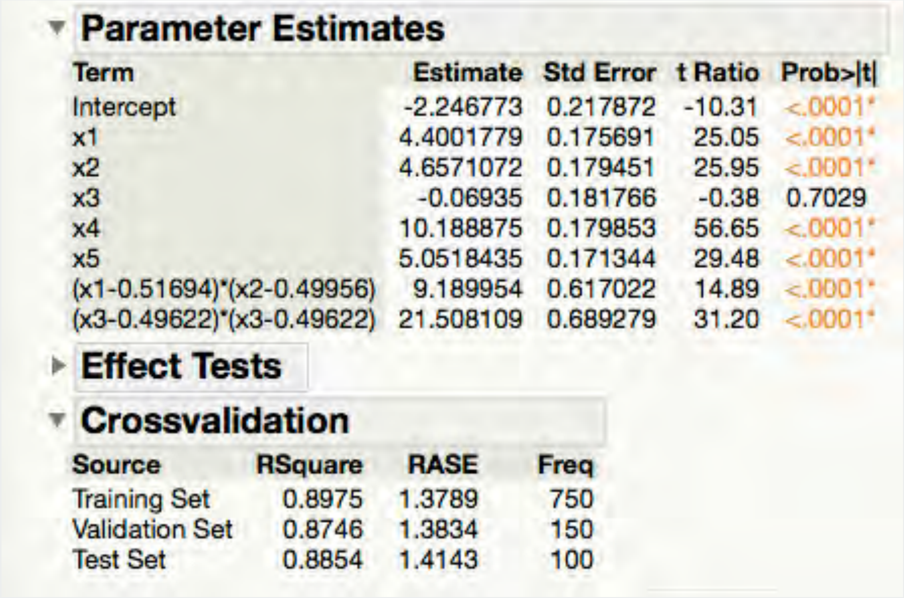
Construct Model Effects

x1
x2
x3
x4
x5
x1*x2
x3*x3

Degree 2

☐ No Intercept

RESULTS



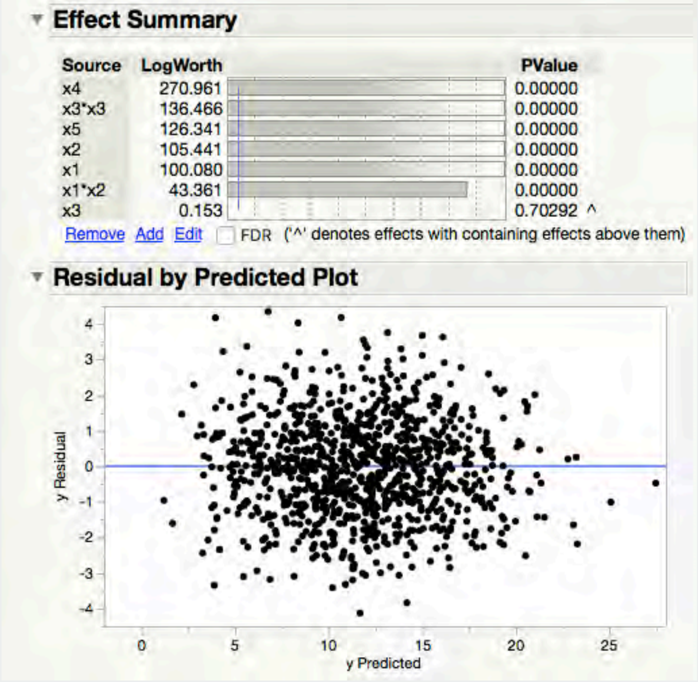
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.246773	0.217872	-10.31	<.0001*
x1	4.4001779	0.175691	25.05	<.0001*
x2	4.6571072	0.179451	25.95	<.0001*
x3	-0.06935	0.181766	-0.38	0.7029
x4	10.188875	0.179853	56.65	<.0001*
x5	5.0518435	0.171344	29.48	<.0001*
(x1-0.51694)*(x2-0.49956)	9.189954	0.617022	14.89	<.0001*
(x3-0.49622)*(x3-0.49622)	21.508109	0.689279	31.20	<.0001*

Effect Tests

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.8975	1.3789	750
Validation Set	0.8746	1.3834	150
Test Set	0.8854	1.4143	100



Effect Summary

Source	LogWorth	PValue
x4	270.961	0.00000
x3*x3	136.466	0.00000
x5	126.341	0.00000
x2	105.441	0.00000
x1	100.080	0.00000
x1*x2	43.361	0.00000
x3	0.153	0.70292 ^

('^' denotes effects with containing effects above them)

Residual by Predicted Plot

WHERE TO GO FROM HERE?

- Regression Pros:
  - Powerful transparent model
  - Flexible using transformations
- Regression Cons:
  - Model selection
  - Scalability
- Generalized Regression
  - Modify Least Squares — Similar to Ridge Regression
- Logistic Regression

PENALIZED REGRESSION

$$L(\beta, x, y) = \sum (y_i - \hat{y}_i)^2$$

- Least squares

PENALIZED REGRESSION

$$L(\beta,x,y) = \sum (y_i - \hat{y}_i)^2$$

- Least squares

$$L(\beta,x,y) = \sum (y_i - \hat{y}_i)^2 + \lambda \sum b_j^2$$

- Ridge Regression

PENALIZED REGRESSION

$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2$$

- ✦ Least squares

$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2+\lambda\sum b_j^2$$

- ✦ Ridge Regression

$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2+\lambda\sum\left|b_j\right|$$

- ✦ Lasso

PENALIZED REGRESSION

$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2$$

- ✦ Least squares

$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2+\lambda\sum b_j^2$$

- ✦ Ridge Regression

$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2+\lambda\sum\left|b_j\right|$$

- ✦ Lasso

$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2+\lambda_1\sum\left|b_j\right|+\lambda_2\sum b_j^2$$

- ✦ Elastic Net

PENALIZED REGRESSION

$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2$$

- ✦ Least squares

$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2+\lambda\sum b_j^2$$

- ✦ Ridge Regression

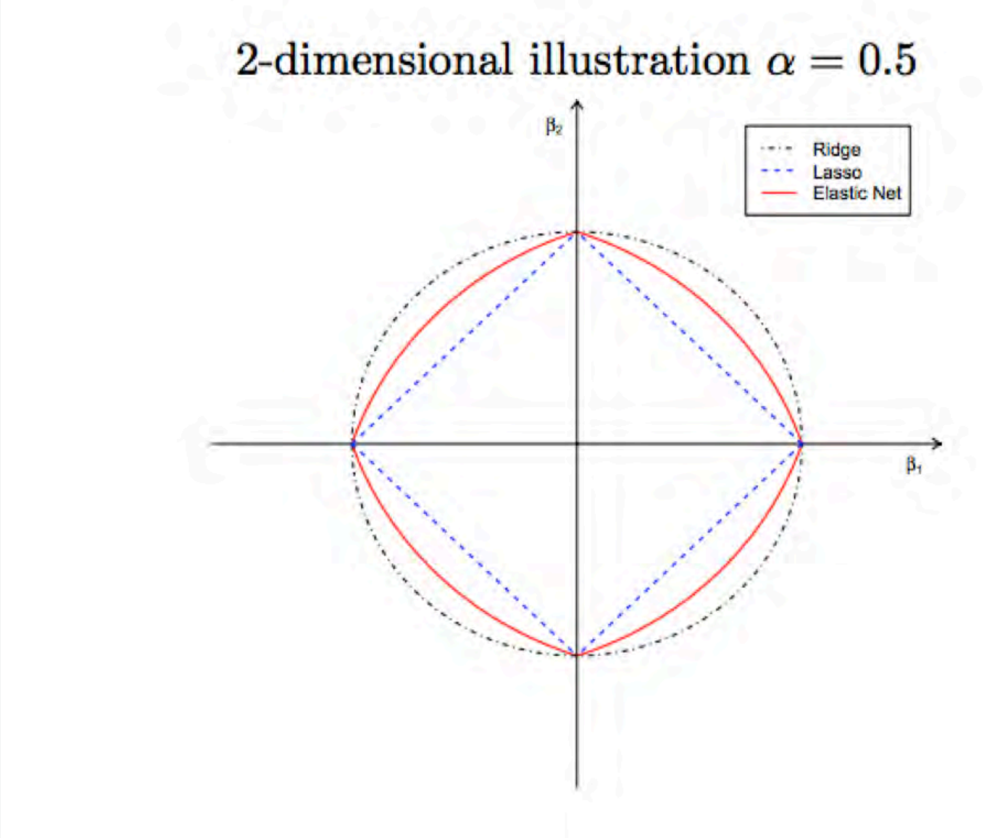
$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2+\lambda\sum\left|b_j\right|$$

- ✦ Lasso

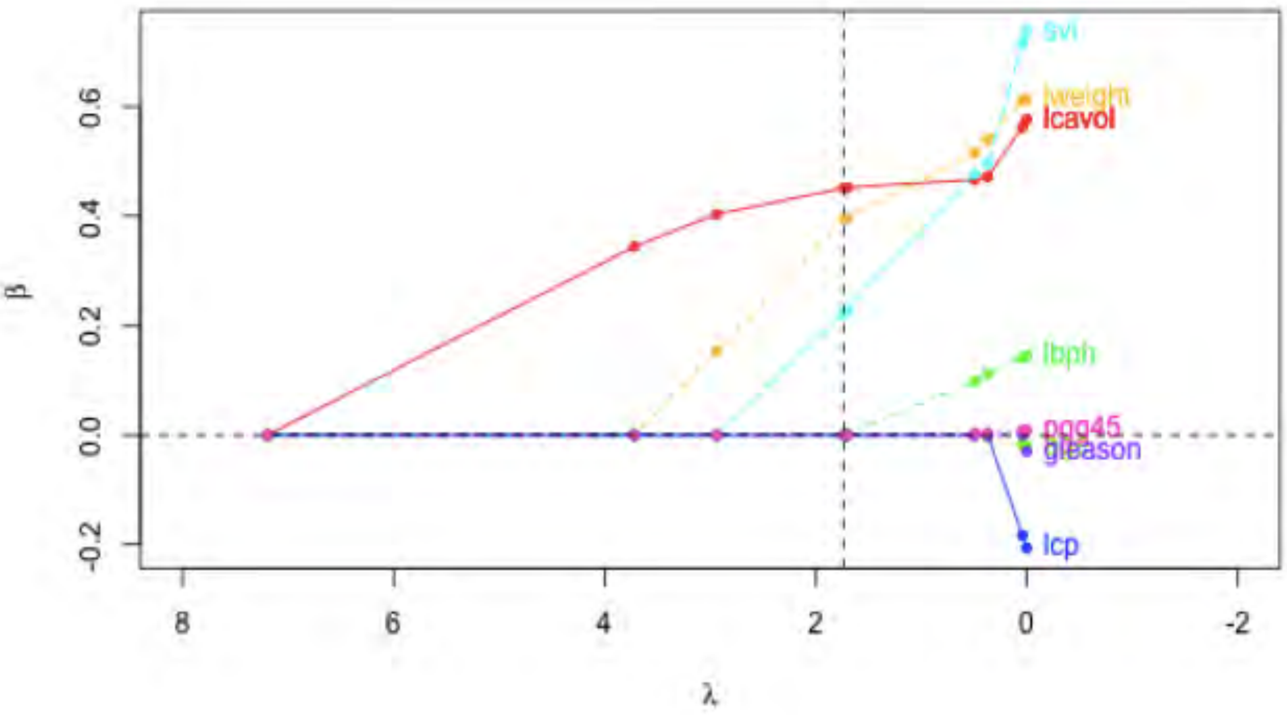
$$L(\beta,x,y)=\sum(y_i-\hat{y}_i)^2+\lambda_1\sum\left|b_j\right|+\lambda_2\sum b_j^2$$

- ✦ Elastic Net

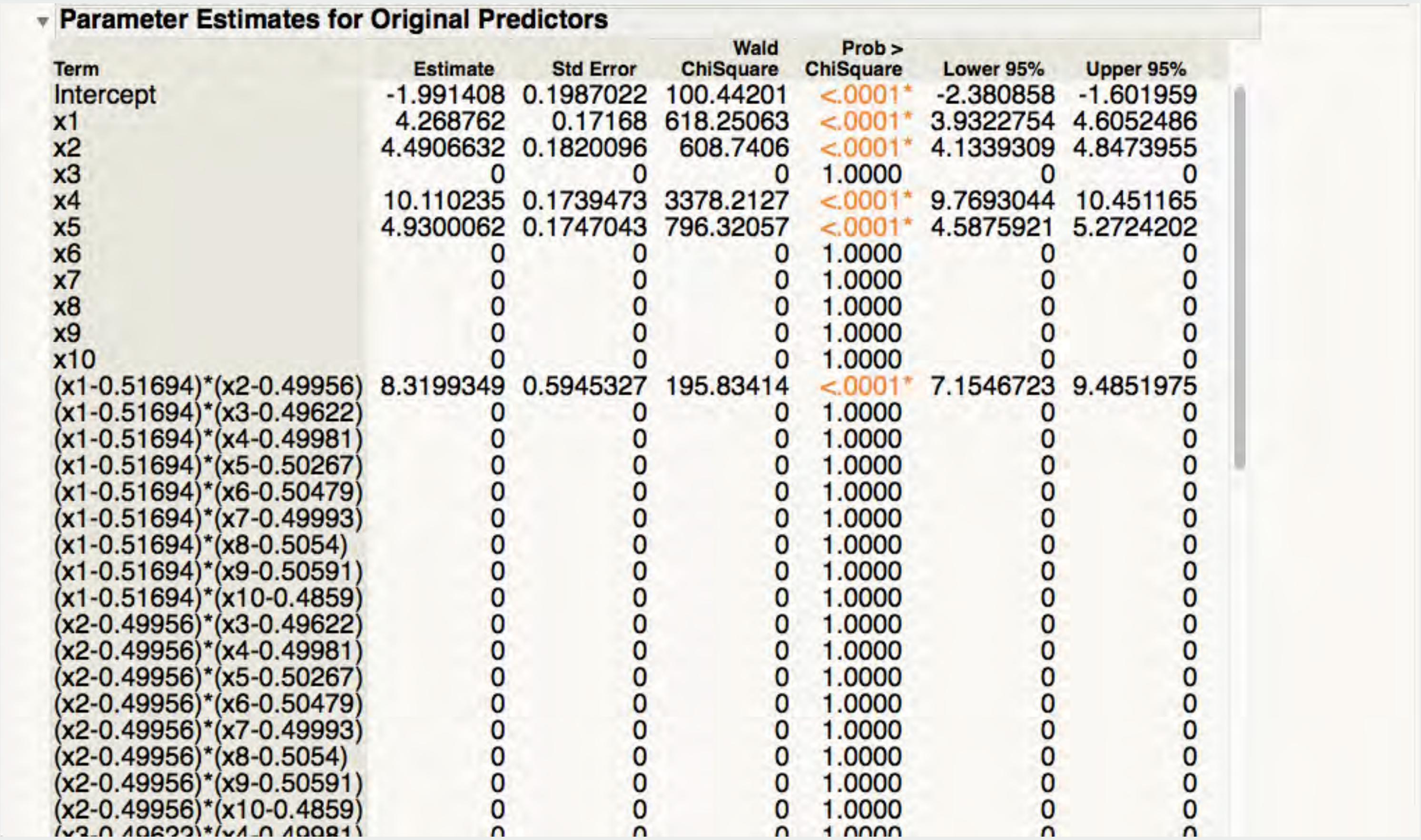
RIDGE, LASSO, ELASTIC NET



LASSO AS STAGEWISE REGRESSION



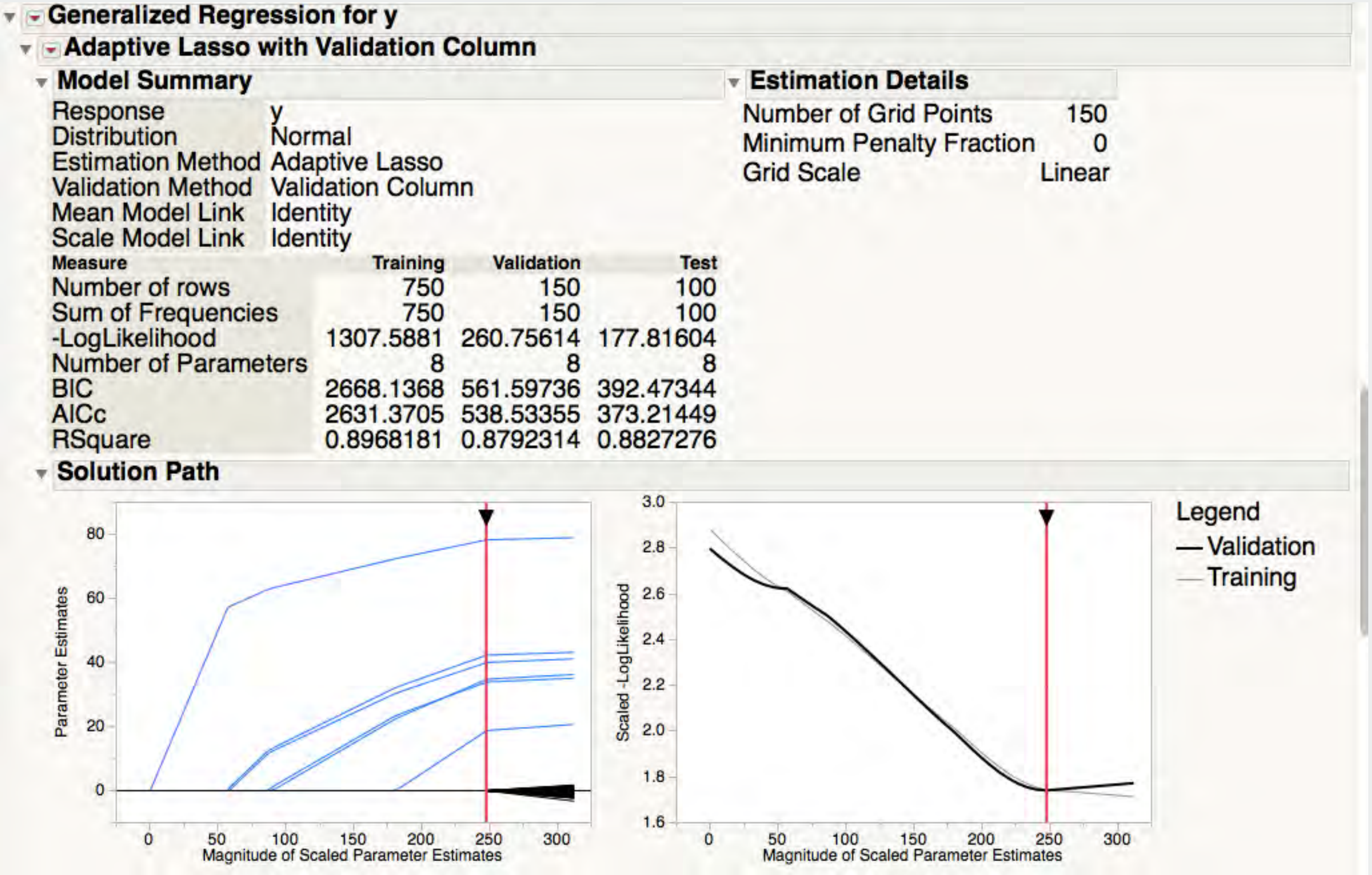
LASSO COEFFICIENTS



Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-1.991408	0.1987022	100.44201	<.0001*	-2.380858	-1.601959
x1	4.268762	0.17168	618.25063	<.0001*	3.9322754	4.6052486
x2	4.4906632	0.1820096	608.7406	<.0001*	4.1339309	4.8473955
x3	0	0	0	1.0000	0	0
x4	10.110235	0.1739473	3378.2127	<.0001*	9.7693044	10.451165
x5	4.9300062	0.1747043	796.32057	<.0001*	4.5875921	5.2724202
x6	0	0	0	1.0000	0	0
x7	0	0	0	1.0000	0	0
x8	0	0	0	1.0000	0	0
x9	0	0	0	1.0000	0	0
x10	0	0	0	1.0000	0	0
(x1-0.51694)*(x2-0.49956)	8.3199349	0.5945327	195.83414	<.0001*	7.1546723	9.4851975
(x1-0.51694)*(x3-0.49622)	0	0	0	1.0000	0	0
(x1-0.51694)*(x4-0.49981)	0	0	0	1.0000	0	0
(x1-0.51694)*(x5-0.50267)	0	0	0	1.0000	0	0
(x1-0.51694)*(x6-0.50479)	0	0	0	1.0000	0	0
(x1-0.51694)*(x7-0.49993)	0	0	0	1.0000	0	0
(x1-0.51694)*(x8-0.5054)	0	0	0	1.0000	0	0
(x1-0.51694)*(x9-0.50591)	0	0	0	1.0000	0	0
(x1-0.51694)*(x10-0.4859)	0	0	0	1.0000	0	0
(x2-0.49956)*(x3-0.49622)	0	0	0	1.0000	0	0
(x2-0.49956)*(x4-0.49981)	0	0	0	1.0000	0	0
(x2-0.49956)*(x5-0.50267)	0	0	0	1.0000	0	0
(x2-0.49956)*(x6-0.50479)	0	0	0	1.0000	0	0
(x2-0.49956)*(x7-0.49993)	0	0	0	1.0000	0	0
(x2-0.49956)*(x8-0.5054)	0	0	0	1.0000	0	0
(x2-0.49956)*(x9-0.50591)	0	0	0	1.0000	0	0
(x2-0.49956)*(x10-0.4859)	0	0	0	1.0000	0	0
(x3-0.49622)*(x4-0.49981)	0	0	0	1.0000	0	0

LASSO



Generalized Regression for y

Adaptive Lasso with Validation Column

Model Summary

Response	y
Distribution	Normal
Estimation Method	Adaptive Lasso
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

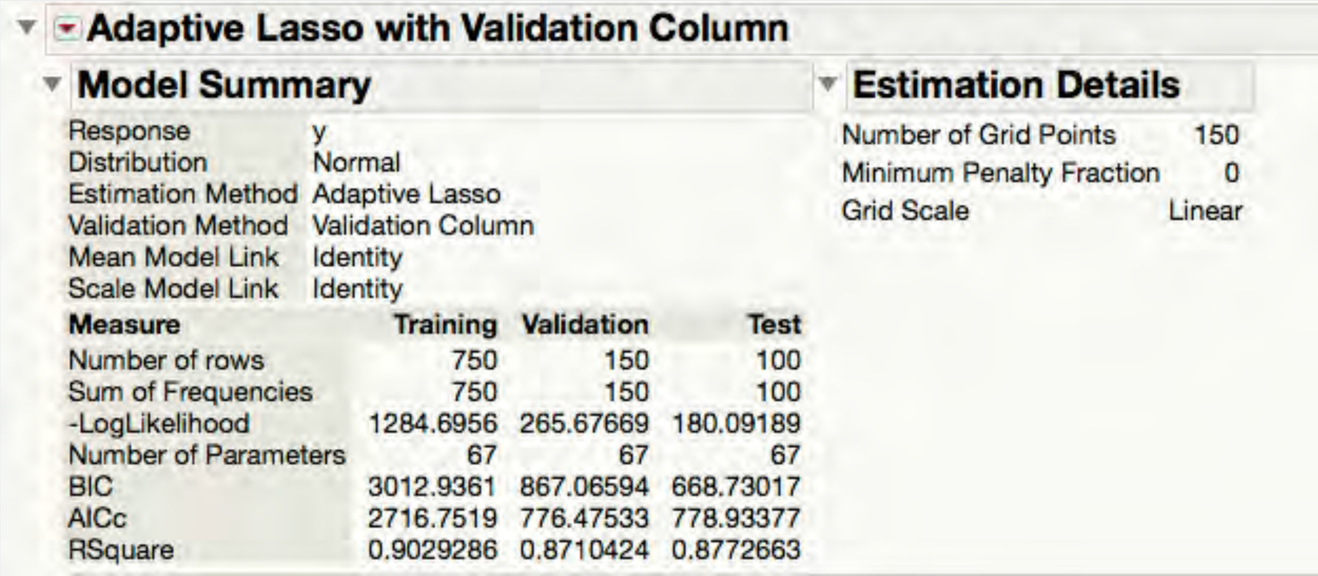
Measure	Training	Validation	Test
Number of rows	750	150	100
Sum of Frequencies	750	150	100
-LogLikelihood	1307.5881	260.75614	177.81604
Number of Parameters	8	8	8
BIC	2668.1368	561.59736	392.47344
AICc	2631.3705	538.53355	373.21449
RSquare	0.8968181	0.8792314	0.8827276

Estimation Details

Number of Grid Points	150
Minimum Penalty Fraction	0
Grid Scale	Linear

Solution Path

LASSO RESULTS



Adaptive Lasso with Validation Column

Model Summary

Response	y
Distribution	Normal
Estimation Method	Adaptive Lasso
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

Measure	Training	Validation	Test
Number of rows	750	150	100
Sum of Frequencies	750	150	100
-LogLikelihood	1284.6956	265.67669	180.09189
Number of Parameters	67	67	67
BIC	3012.9361	867.06594	668.73017
AICc	2716.7519	776.47533	778.93377
RSquare	0.9029286	0.8710424	0.8772663

Estimation Details

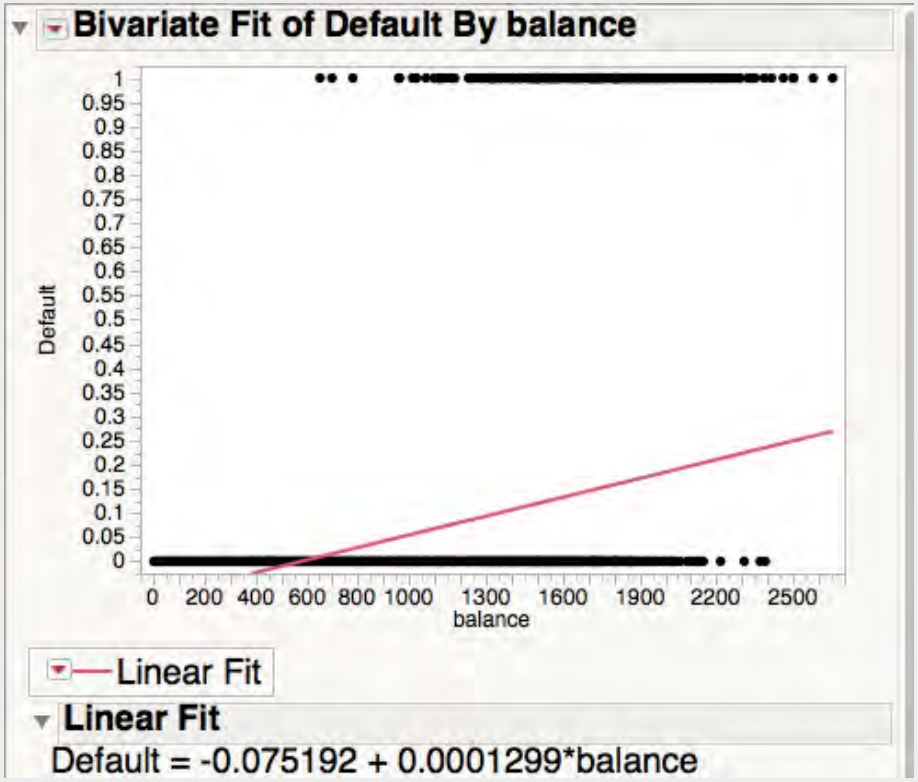
Number of Grid Points	150
Minimum Penalty Fraction	0
Grid Scale	Linear

TYPES OF PROBLEMS

- Descriptions
- Classification (categorical or discrete values)
- Regression (continuous values)
  - Time series (continuous values)
- Clustering
- Association

LOGISTIC REGRESSION

What happens if we use linear regression on 1-0 (yes/no) data?

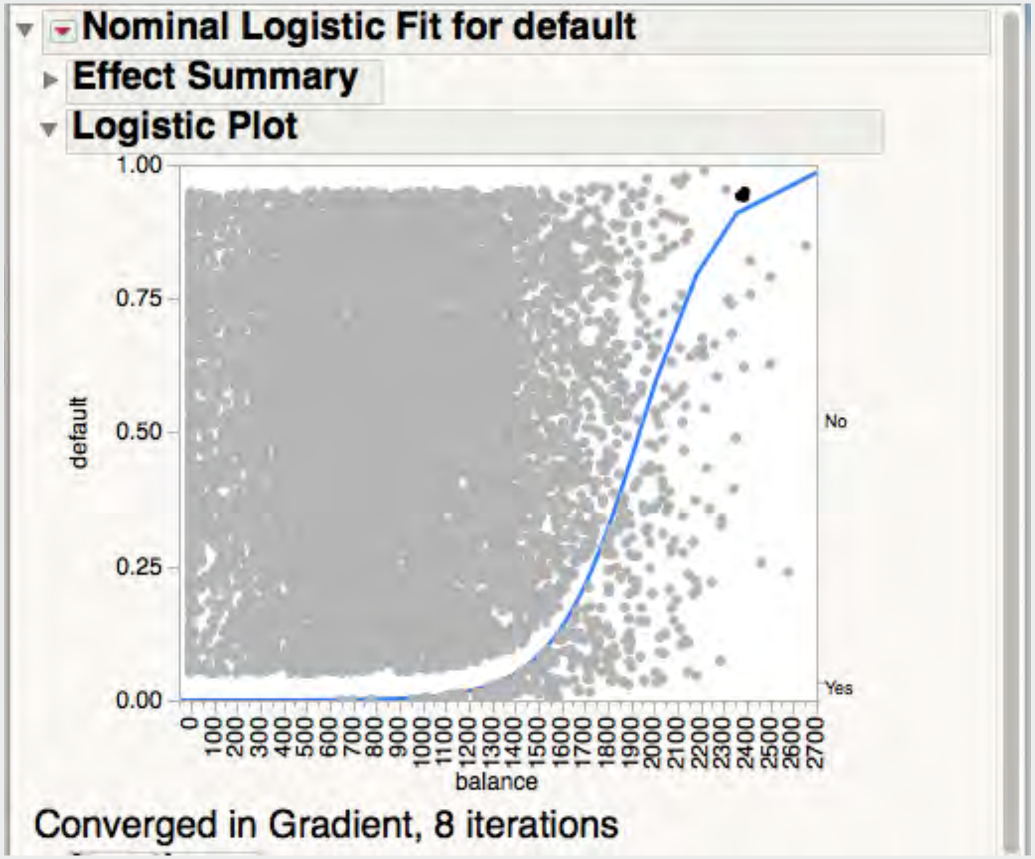


Bivariate Fit of Default By balance

Linear Fit

Default = -0.075192 + 0.0001299*balance

LOGISTIC REGRESSION

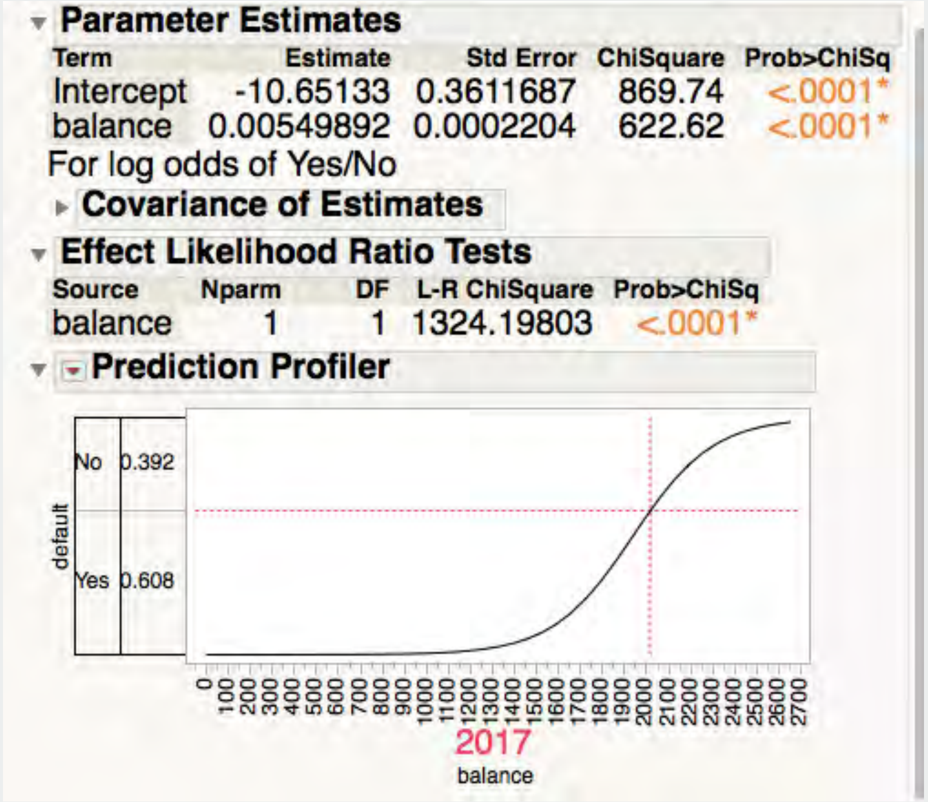


Nominal Logistic Fit for default

Effect Summary

Logistic Plot

Converged in Gradient, 8 iterations



Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-10.65133	0.3611687	869.74	<.0001*
balance	0.00549892	0.0002204	622.62	<.0001*

For log odds of Yes/No

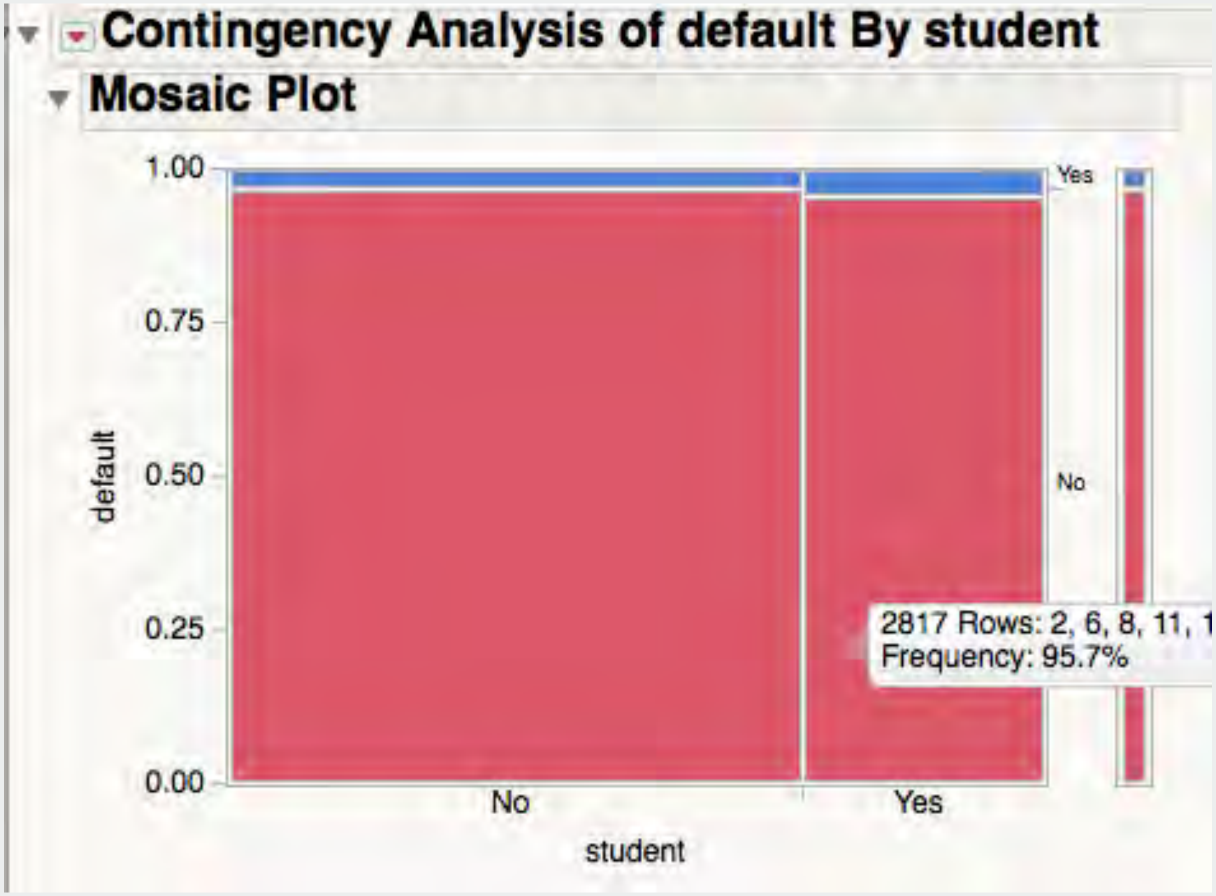
Covariance of Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
balance	1	1	1324.19803	<.0001*

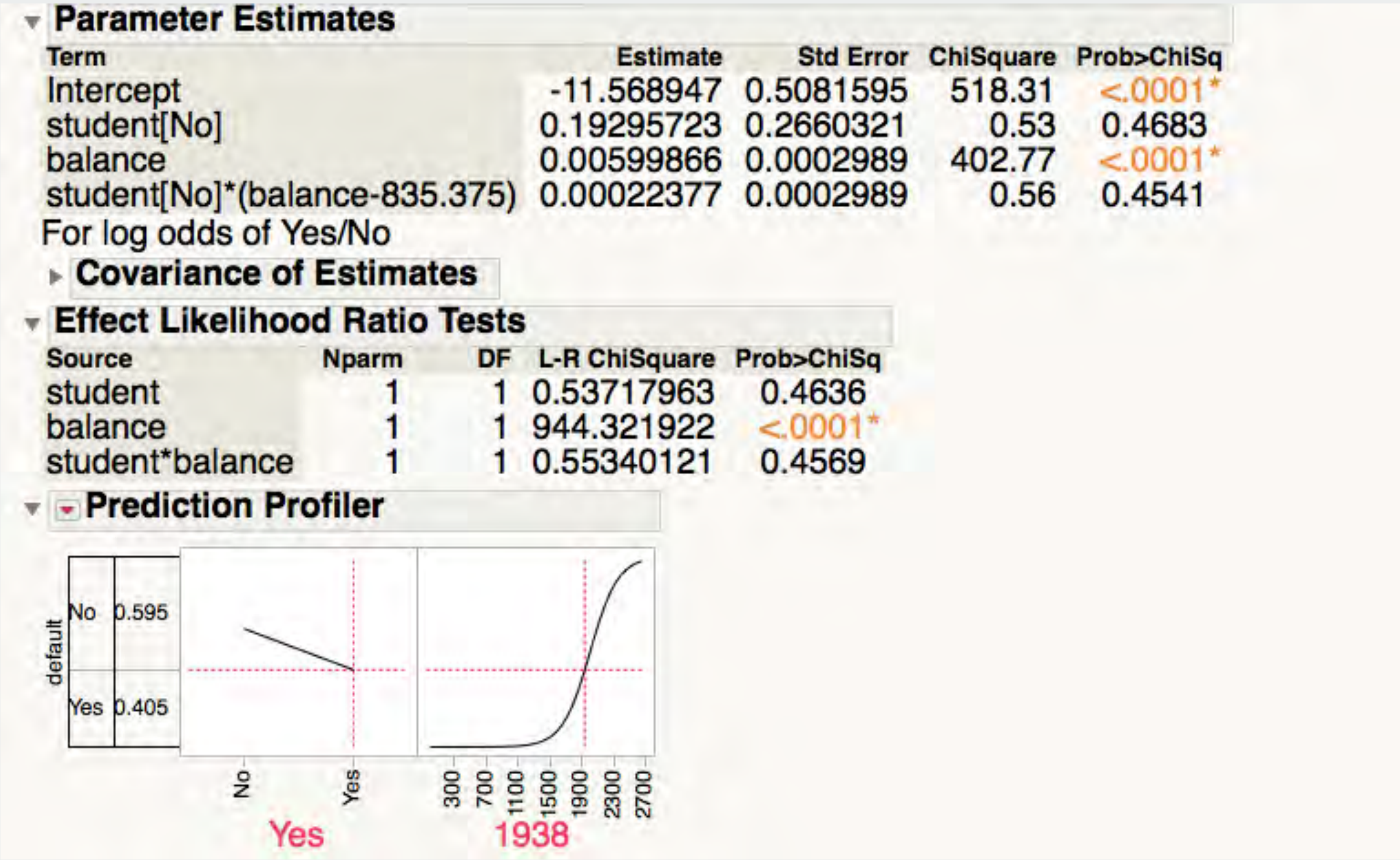
Prediction Profiler

LOGISTIC REGRESSION



2817 Rows: 2, 6, 8, 11, 1
Frequency: 95.7%

LOGISTIC REGRESSION



▾ **Parameter Estimates**

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-11.568947	0.5081595	518.31	<.0001*
student[No]	0.19295723	0.2660321	0.53	0.4683
balance	0.00599866	0.0002989	402.77	<.0001*
student[No]*(balance-835.375)	0.00022377	0.0002989	0.56	0.4541

For log odds of Yes/No

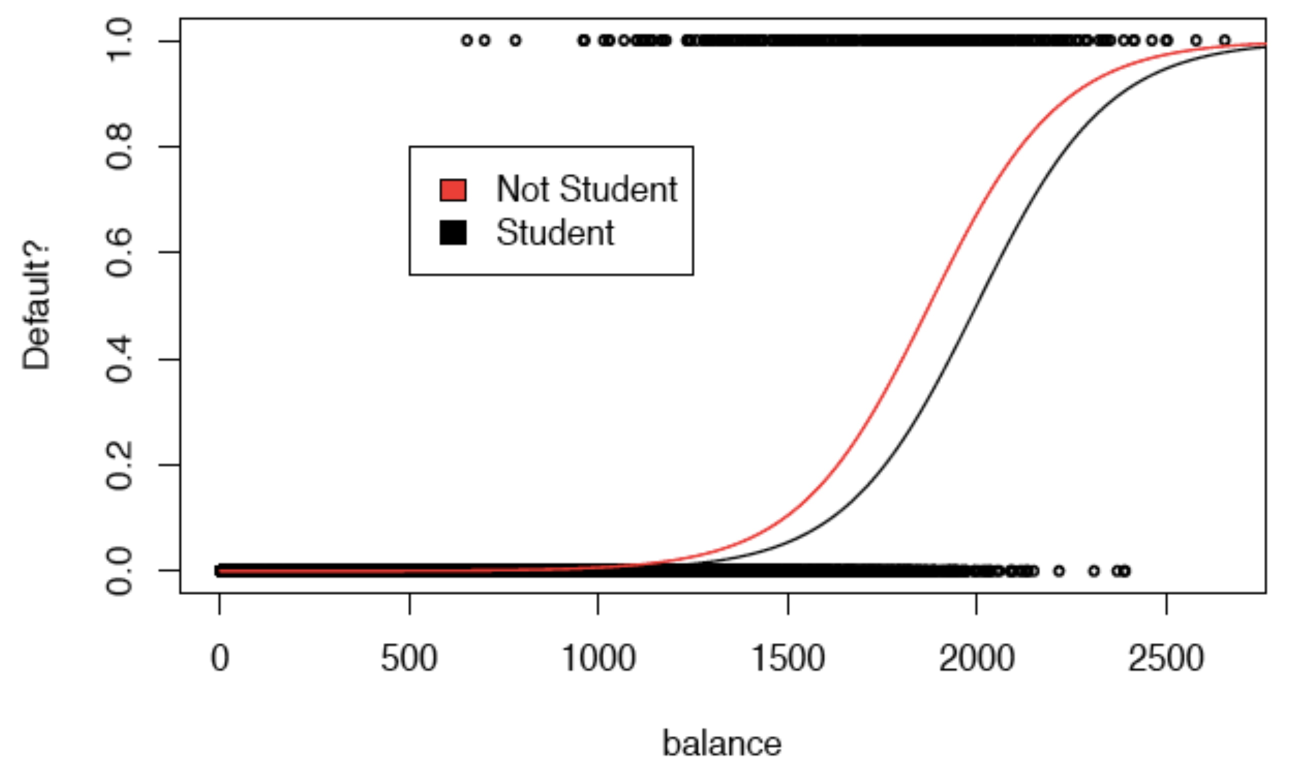
▸ **Covariance of Estimates**

▾ **Effect Likelihood Ratio Tests**

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
student	1	1	0.53717963	0.4636
balance	1	1	944.321922	<.0001*
student*balance	1	1	0.55340121	0.4569

▾ **Prediction Profiler**

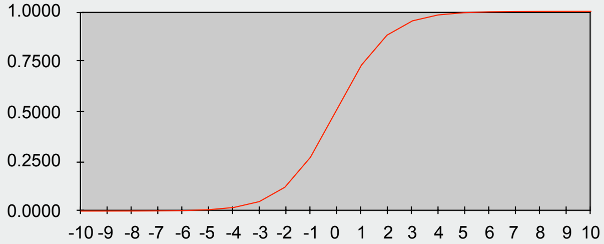
LOGISTIC REGRESSION — IN **R**



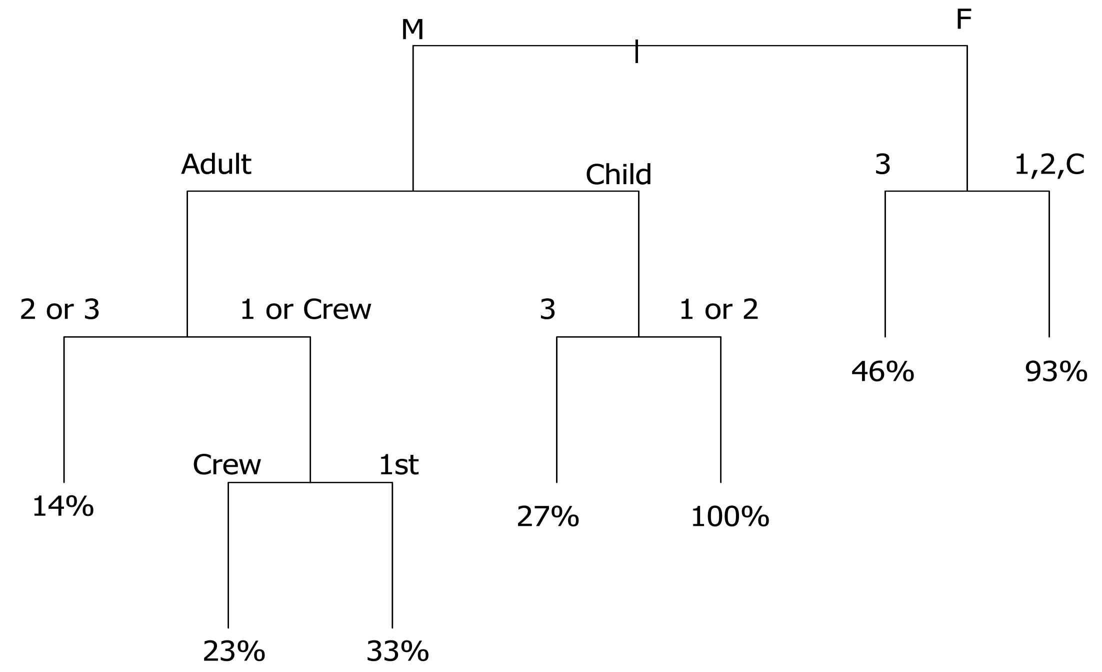
LOGISTIC REGRESSION II

- Points on the line can be interpreted as probability, but don't stay within [0,1]
- Use a sigmoidal function instead of linear function to fit the data

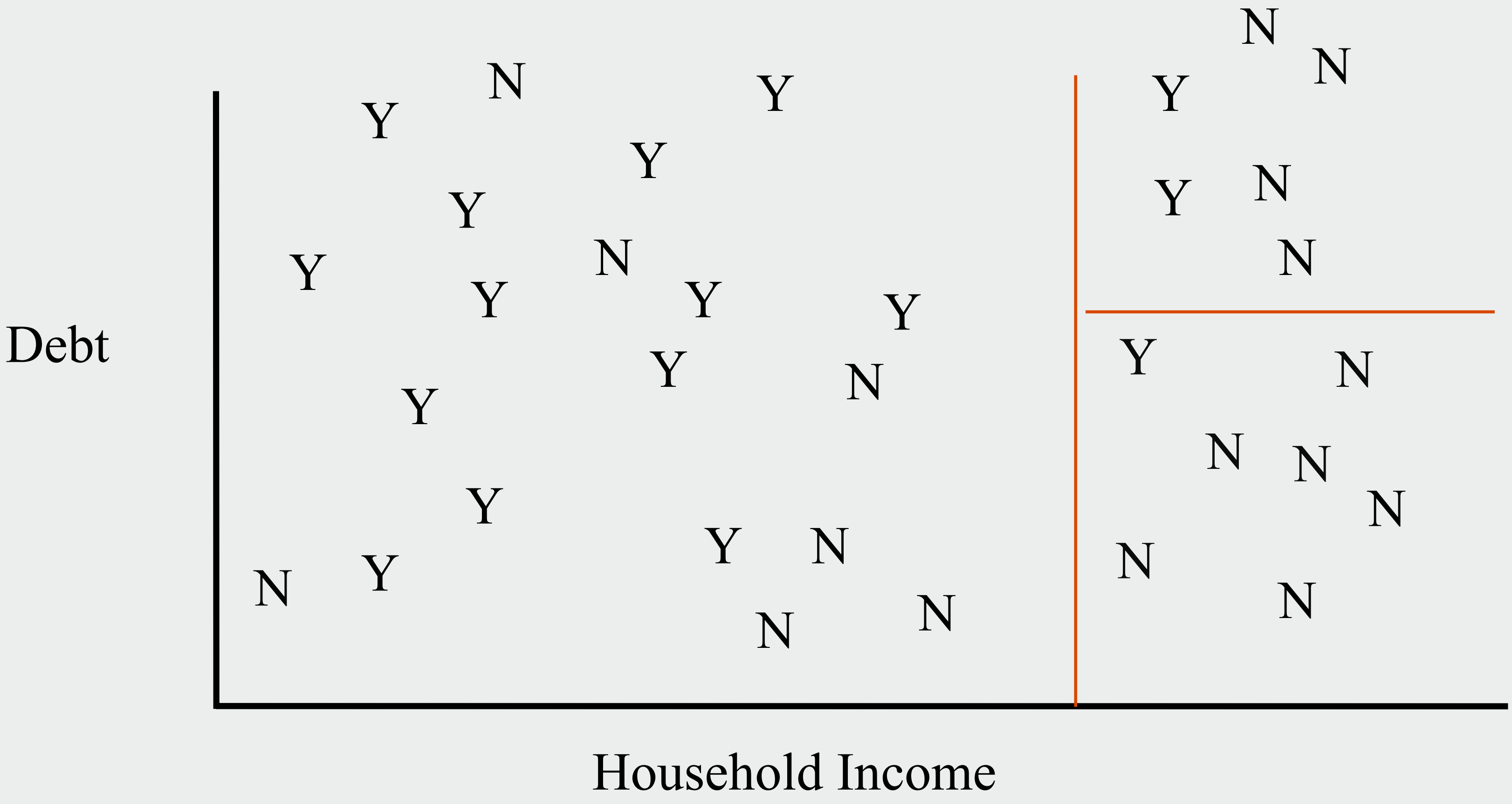
$$f(I) = \frac{1}{1+e^{-I}}$$



BACK TO THE TITANIC



GEOMETRY OF DECISION TREES



CONFUSION MATRIX

Doctors in ER	**Actual Heart Attack**	**No Heart Attack**
Predict Heart Attack	**0.89**	**0.75**
Predict No Heart Attack	**0.11**	**0.25**

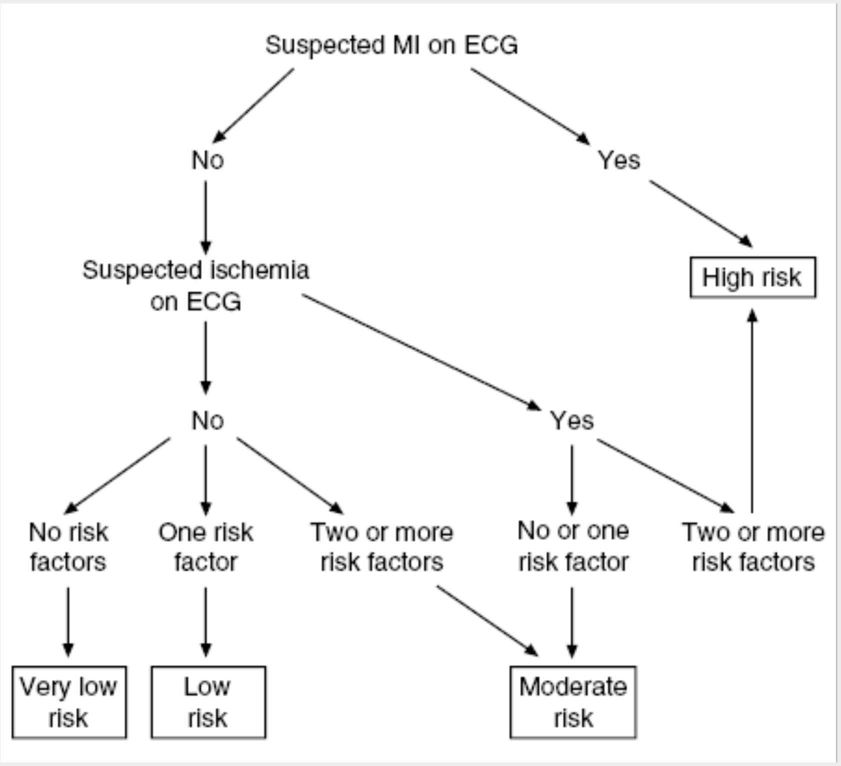
COOK COUNTY ER

The ECG Tests:

- MI: myocardial infarction (heart attack)
- Ischemia – Heart muscle not getting enough blood

The 3 Risk Factors:

1. Is the reported Pain unstable angina?
2. Is there fluid in patient's lungs?
3. Is the patient's systolic BP < 100?

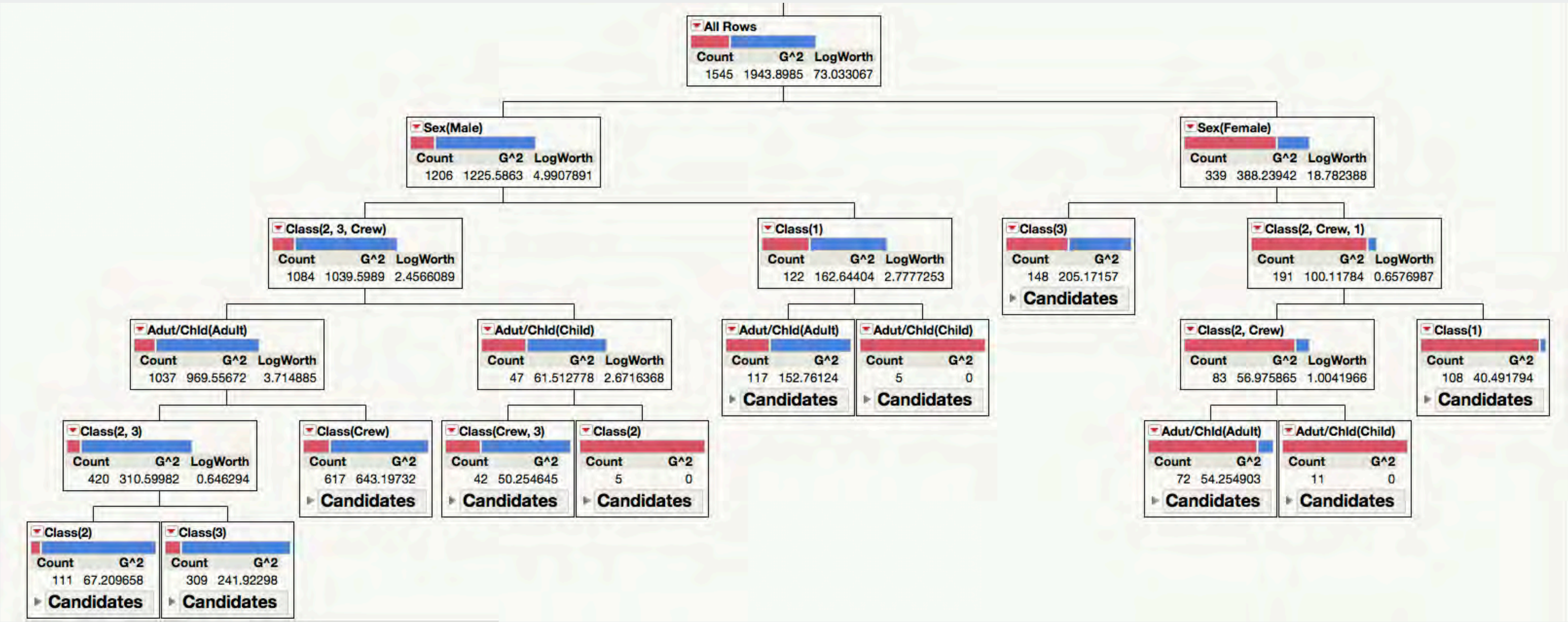


HOW DID IT DO?

Doctors in ER	Actual Heart Attack	No Heart Attack
Predict Heart Attack	0.89	0.75
Predict No Heart Attack	0.11	0.25

Predict Using Tree	Actual Heart Attack	No Heart Attack
Predict Heart Attack	0.96	0.08
Predict No Heart Attack	0.04	0.92

TREE ON TITANIC



All Rows
Count G^2 LogWorth
1545 1943.8985 73.033067

Sex(Male)
Count G^2 LogWorth
1206 1225.5863 4.9907891

Sex(Female)
Count G^2 LogWorth
339 388.23942 18.782388

Class(2, 3, Crew)
Count G^2 LogWorth
1084 1039.5989 2.4566089

Class(1)
Count G^2 LogWorth
122 162.64404 2.7777253

Class(3)
Count G^2
148 205.17157
Candidates

Class(2, Crew, 1)
Count G^2 LogWorth
191 100.11784 0.6576987

Adut/Chld(Adult)
Count G^2 LogWorth
1037 969.55672 3.714885

Adut/Chld(Child)
Count G^2 LogWorth
47 61.512778 2.6716368

Adut/Chld(Adult)
Count G^2
117 152.76124
Candidates

Adut/Chld(Child)
Count G^2
5 0
Candidates

Class(2, Crew)
Count G^2 LogWorth
83 56.975865 1.0041966

Class(1)
Count G^2
108 40.491794
Candidates

Class(2, 3)
Count G^2 LogWorth
420 310.59982 0.646294

Class(Crew)
Count G^2
617 643.19732
Candidates

Class(Crew, 3)
Count G^2
42 50.254645
Candidates

Class(2)
Count G^2
5 0
Candidates

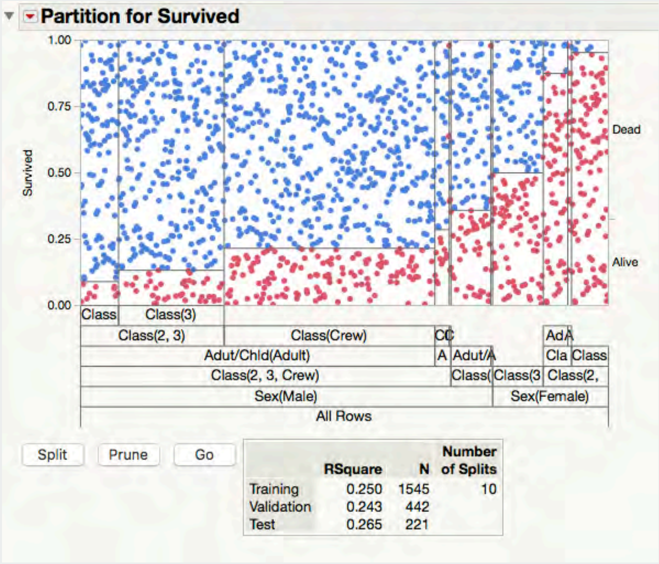
Adut/Chld(Adult)
Count G^2
72 54.254903
Candidates

Adut/Chld(Child)
Count G^2
11 0
Candidates

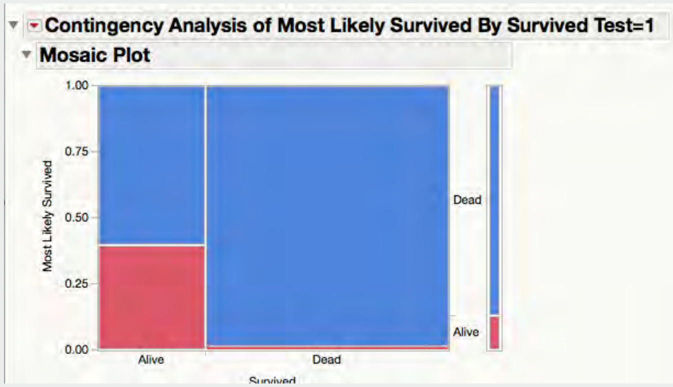
Class(2)
Count G^2
111 67.209658
Candidates

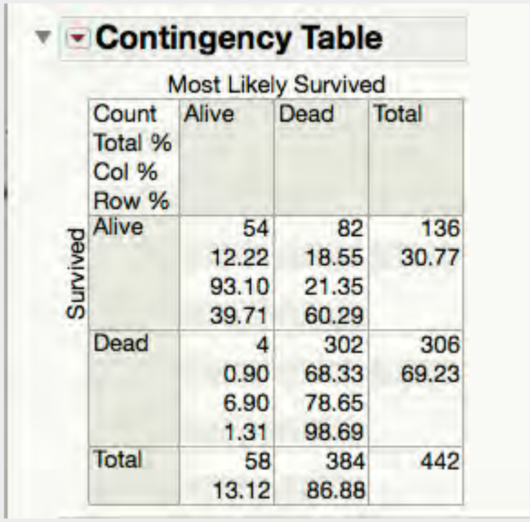
Class(3)
Count G^2
309 241.92298
Candidates

TREE ON TITANIC



	RSquare	N	Number of Splits
Training	0.250	1545	10
Validation	0.243	442	
Test	0.265	221	





Contingency Table

		Most Likely Survived		
	Count Total % Col % Row %	Alive	Dead	Total
Survived	Alive	54 12.22 93.10 39.71	82 18.55 21.35 60.29	136 30.77
	Dead	4 0.90 6.90 1.31	302 68.33 78.65 98.69	306 69.23
	Total	58 13.12	384 86.88	442

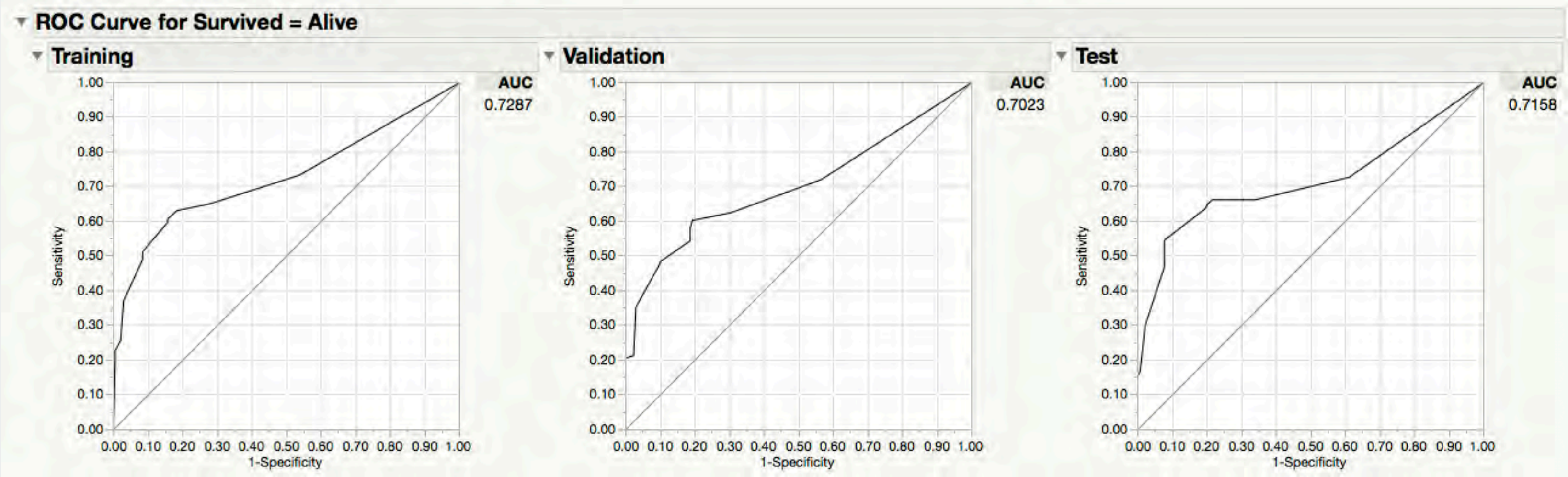
Can (should) optimize levels to trade off
Type I and Type II errors
Here — Type I 60.29% (!)
Type II 1.31%

CHANGING THRESHOLDS

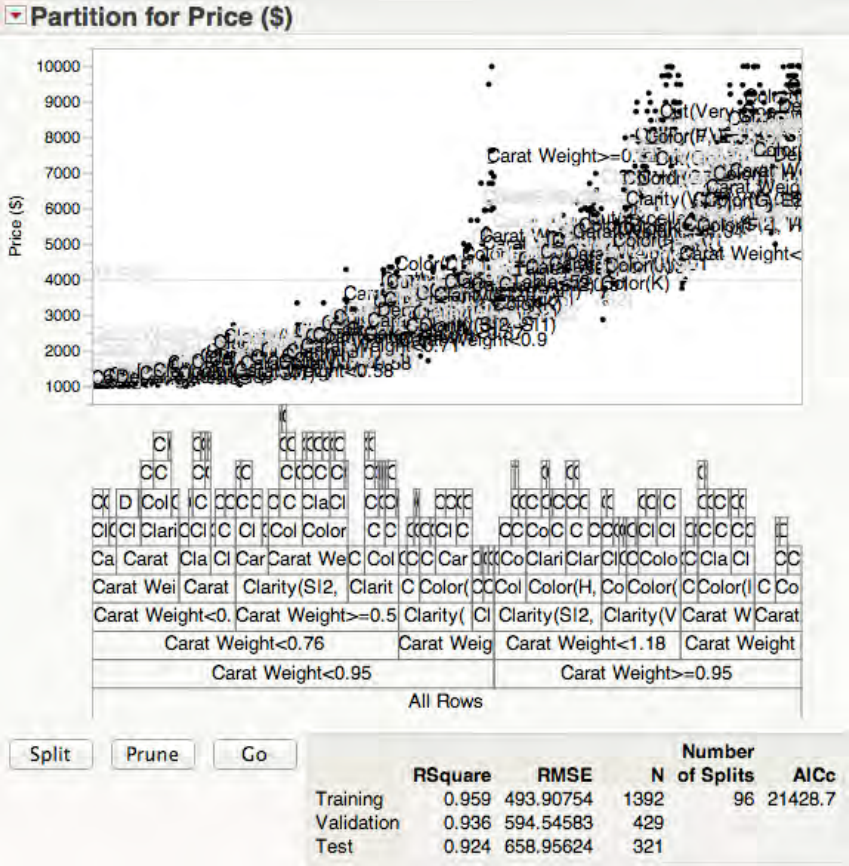
- For most classification algorithms, the default threshold (yes/no) is 0.5
- You do not have to use 0.5 especially when a Type II error is important
- We'll use 0.1 here (but typically you should try several values and see which has best average misclassification rates).

ROC CURVE

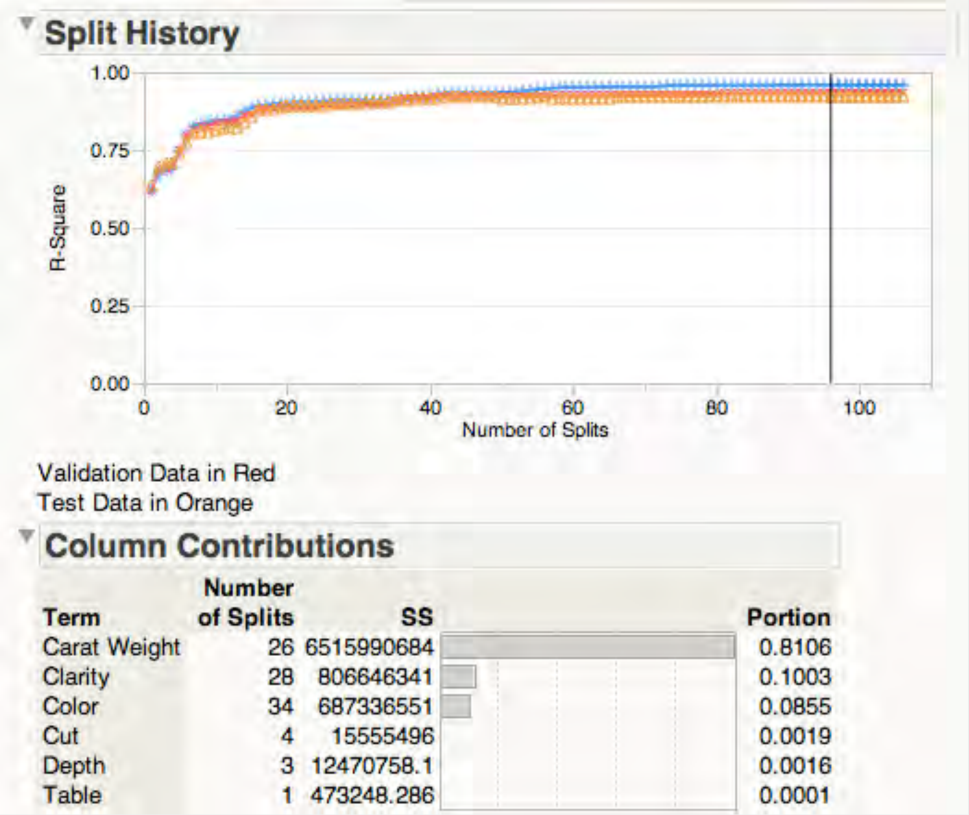
- Trade off and optimize Type I vs. Type II error
- Sensitivity (1- Type II error) vs. 1- Specificity (Type I error)
- Typically used to compare models (and vs. null model)



TREE ON DIAMOND DATA



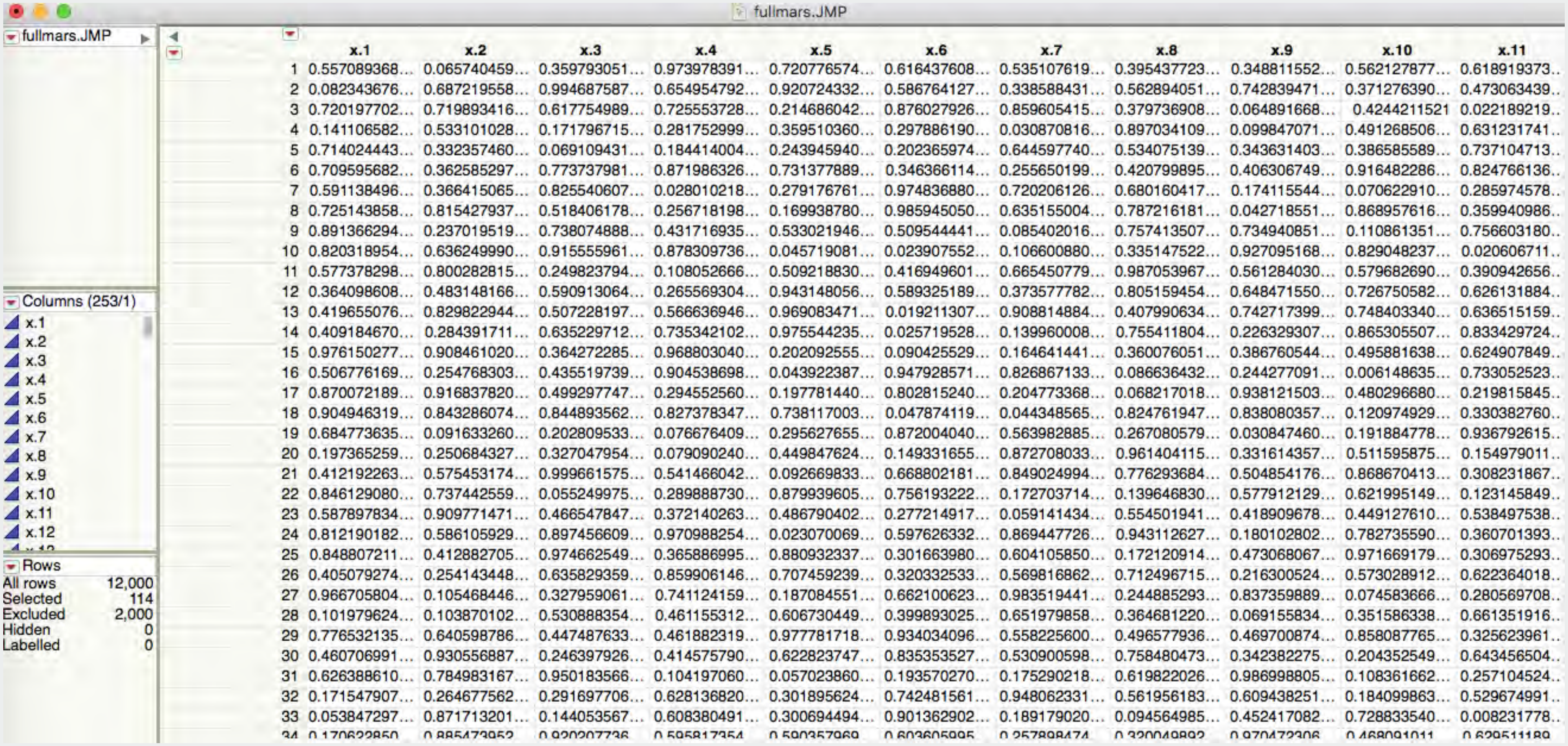
	RSquare	RMSE	N	Number of Splits	AICc
Training	0.959	493.90754	1392	96	21428.7
Validation	0.936	594.54583	429		
Test	0.924	658.95624	321		



Validation Data in Red
Test Data in Orange

Term	Number of Splits	SS	Portion
Carat Weight	26	6515990684	0.8106
Clarity	28	806646341	0.1003
Color	34	687336551	0.0855
Cut	4	15555496	0.0019
Depth	3	12470758.1	0.0016
Table	1	473248.286	0.0001

MORE REALISTIC



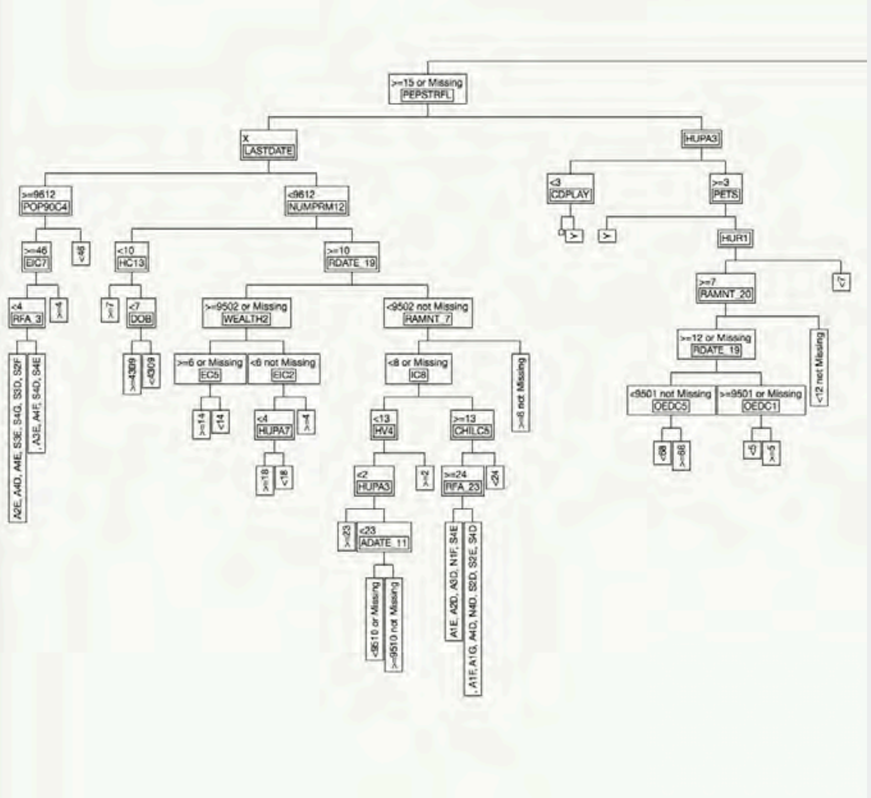
All rows	12,000
Selected	114
Excluded	2,000
Hidden	0
Labelled	0

	x.1	x.2	x.3	x.4	x.5	x.6	x.7	x.8	x.9	x.10	x.11
1	0.557089368...	0.065740459...	0.359793051...	0.973978391...	0.720776574...	0.616437608...	0.535107619...	0.395437723...	0.348811552...	0.562127877...	0.618919373..
2	0.082343676...	0.687219558...	0.994687587...	0.654954792...	0.920724332...	0.586764127...	0.338588431...	0.562894051...	0.742839471...	0.371276390...	0.473063439..
3	0.720197702...	0.719893416...	0.617754989...	0.725553728...	0.214686042...	0.876027926...	0.859605415...	0.379736908...	0.064891668...	0.4244211521	0.022189219..
4	0.141106582...	0.533101028...	0.171796715...	0.281752999...	0.359510360...	0.297886190...	0.030870816...	0.897034109...	0.099847071...	0.491268506...	0.631231741..
5	0.714024443...	0.332357460...	0.069109431...	0.184414004...	0.243945940...	0.202365974...	0.644597740...	0.534075139...	0.343631403...	0.386585589...	0.737104713..
6	0.709595682...	0.362585297...	0.773737981...	0.871986326...	0.731377889...	0.346366114...	0.255650199...	0.420799895...	0.406306749...	0.916482286...	0.824766136..
7	0.591138496...	0.366415065...	0.825540607...	0.028010218...	0.279176761...	0.974836880...	0.720206126...	0.680160417...	0.174115544...	0.070622910...	0.285974578..
8	0.725143858...	0.815427937...	0.518406178...	0.256718198...	0.169938780...	0.985945050...	0.635155004...	0.787216181...	0.042718551...	0.868957616...	0.359940986..
9	0.891366294...	0.237019519...	0.738074888...	0.431716935...	0.533021946...	0.509544441...	0.085402016...	0.757413507...	0.734940851...	0.110861351...	0.756603180..
10	0.820318954...	0.636249990...	0.915555961...	0.878309736...	0.045719081...	0.023907552...	0.106600880...	0.335147522...	0.927095168...	0.829048237...	0.020606711..
11	0.577378298...	0.800282815...	0.249823794...	0.108052666...	0.509218830...	0.416949601...	0.665450779...	0.987053967...	0.561284030...	0.579682690...	0.390942656..
12	0.364098608...	0.483148166...	0.590913064...	0.265569304...	0.943148056...	0.589325189...	0.373577782...	0.805159454...	0.648471550...	0.726750582...	0.626131884..
13	0.419655076...	0.829822944...	0.507228197...	0.566636946...	0.969083471...	0.019211307...	0.908814884...	0.407990634...	0.742717399...	0.748403340...	0.636515159..
14	0.409184670...	0.284391711...	0.635229712...	0.735342102...	0.975544235...	0.025719528...	0.139960008...	0.755411804...	0.226329307...	0.865305507...	0.833429724..
15	0.976150277...	0.908461020...	0.364272285...	0.968803040...	0.202092555...	0.090425529...	0.164641441...	0.360076051...	0.386760544...	0.495881638...	0.624907849..
16	0.506776169...	0.254768303...	0.435519739...	0.904538698...	0.043922387...	0.947928571...	0.826867133...	0.086636432...	0.244277091...	0.006148635...	0.733052523..
17	0.870072189...	0.916837820...	0.499297747...	0.294552560...	0.197781440...	0.802815240...	0.204773368...	0.068217018...	0.938121503...	0.480296680...	0.219815845..
18	0.904946319...	0.843286074...	0.844893562...	0.827378347...	0.738117003...	0.047874119...	0.044348565...	0.824761947...	0.838080357...	0.120974929...	0.330382760..
19	0.684773635...	0.091633260...	0.202809533...	0.076676409...	0.295627655...	0.872004040...	0.563982885...	0.267080579...	0.030847460...	0.191884778...	0.936792615..
20	0.197365259...	0.250684327...	0.327047954...	0.079090240...	0.449847624...	0.149331655...	0.872708033...	0.961404115...	0.331614357...	0.511595875...	0.154979011..
21	0.412192263...	0.575453174...	0.999661575...	0.541466042...	0.092669833...	0.668802181...	0.849024994...	0.776293684...	0.504854176...	0.868670413...	0.308231867..
22	0.846129080...	0.737442559...	0.055249975...	0.289888730...	0.879939605...	0.756193222...	0.172703714...	0.139646830...	0.577912129...	0.621995149...	0.123145849..
23	0.587897834...	0.909771471...	0.466547847...	0.372140263...	0.486790402...	0.277214917...	0.059141434...	0.554501941...	0.418909678...	0.449127610...	0.538497538..
24	0.812190182...	0.586105929...	0.897456609...	0.970988254...	0.023070069...	0.597626332...	0.869447726...	0.943112627...	0.180102802...	0.782735590...	0.360701393..
25	0.848807211...	0.412882705...	0.974662549...	0.365886995...	0.880932337...	0.301663980...	0.604105850...	0.172120914...	0.473068067...	0.971669179...	0.306975293..
26	0.405079274...	0.254143448...	0.635829359...	0.859906146...	0.707459239...	0.320332533...	0.569816862...	0.712496715...	0.216300524...	0.573028912...	0.622364018..
27	0.966705804...	0.105468446...	0.327959061...	0.741124159...	0.187084551...	0.662100623...	0.983519441...	0.244885293...	0.837359889...	0.074583666...	0.280569708..
28	0.101979624...	0.103870102...	0.530888354...	0.461155312...	0.606730449...	0.399893025...	0.651979858...	0.364681220...	0.069155834...	0.351586338...	0.661351916..
29	0.776532135...	0.640598786...	0.447487633...	0.461882319...	0.977781718...	0.934034096...	0.558225600...	0.496577936...	0.469700874...	0.858087765...	0.325623961..
30	0.460706991...	0.930556887...	0.246397926...	0.414575790...	0.622823747...	0.835353527...	0.530900598...	0.758480473...	0.342382275...	0.204352549...	0.643456504..
31	0.626388610...	0.784983167...	0.950183566...	0.104197060...	0.057023860...	0.193570270...	0.175290218...	0.619822026...	0.986998805...	0.108361662...	0.257104524..
32	0.171547907...	0.264677562...	0.291697706...	0.628136820...	0.301895624...	0.742481561...	0.948062331...	0.561956183...	0.609438251...	0.184099863...	0.529674991..
33	0.053847297...	0.871713201...	0.144053567...	0.608380491...	0.300694494...	0.901362902...	0.189179020...	0.094564985...	0.452417082...	0.728833540...	0.008231778..
34	0.170622850	0.885473952	0.920207736	0.595817354	0.590357969	0.603605995	0.257898474	0.320049892	0.970472306	0.468091011	0.629511189

HOW TO START?

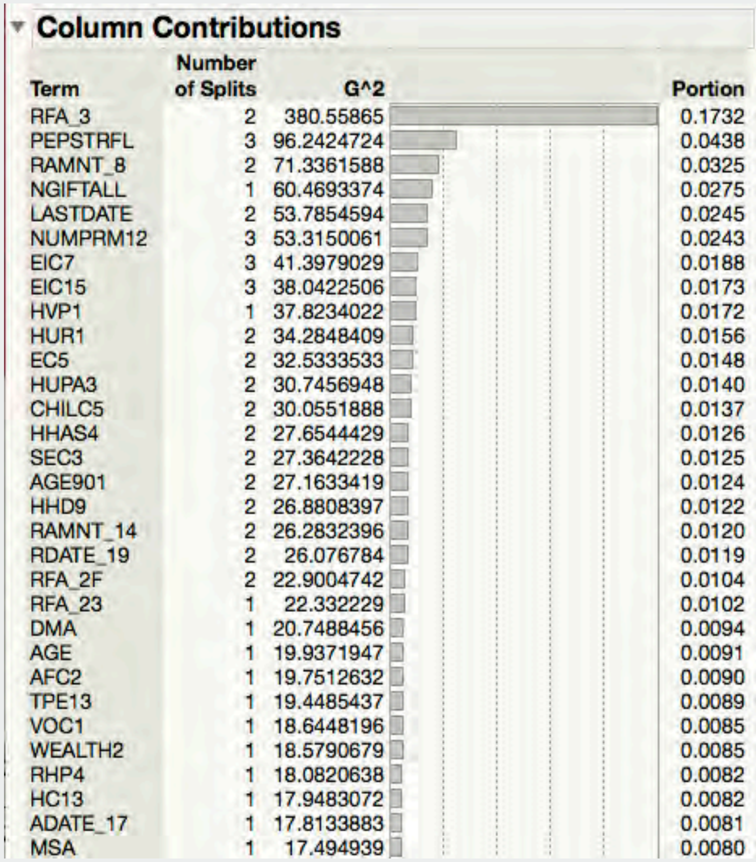
- Explore?
  - How many plots?
- Missing values
- Data transformations
- Feature creation
  - Fancy name for making a new predictor — e.g. Debt/Income
- Data quality

TREE ON PVA



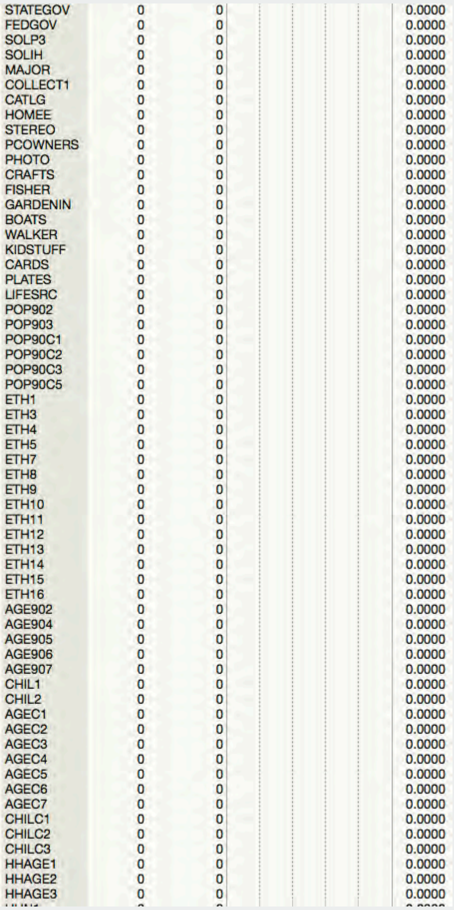
- Don't use tree to model, but for column contributions

COLUMN CONTRIBUTIONS



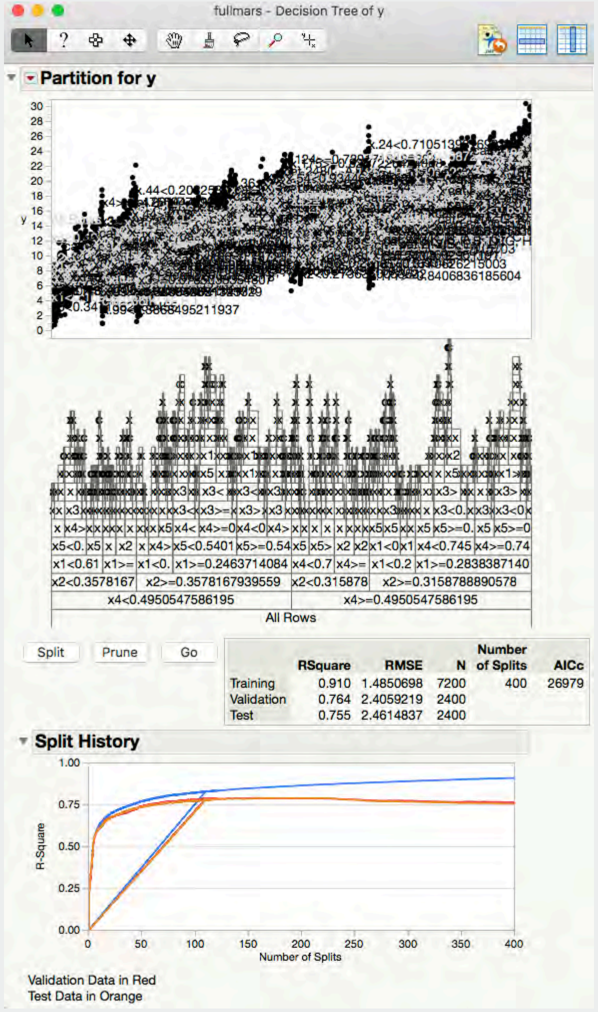
Column Contributions

Term	Number of Splits	G^2	Portion
RFA_3	2	380.55865	0.1732
PEPSTRFL	3	96.2424724	0.0438
RAMNT_8	2	71.3361588	0.0325
NGIFTALL	1	60.4693374	0.0275
LASTDATE	2	53.7854594	0.0245
NUMPRM12	3	53.3150061	0.0243
EIC7	3	41.3979029	0.0188
EIC15	3	38.0422506	0.0173
HVP1	1	37.8234022	0.0172
HUR1	2	34.2848409	0.0156
EC5	2	32.5333533	0.0148
HUPA3	2	30.7456948	0.0140
CHILC5	2	30.0551888	0.0137
HHAS4	2	27.6544429	0.0126
SEC3	2	27.3642228	0.0125
AGE901	2	27.1633419	0.0124
HHD9	2	26.8808397	0.0122
RAMNT_14	2	26.2832396	0.0120
RDATE_19	2	26.076784	0.0119
RFA_2F	2	22.9004742	0.0104
RFA_23	1	22.332229	0.0102
DMA	1	20.7488456	0.0094
AGE	1	19.9371947	0.0091
AFC2	1	19.7512632	0.0090
TPE13	1	19.4485437	0.0089
VOC1	1	18.6448196	0.0085
WEALTH2	1	18.5790679	0.0085
RHP4	1	18.0820638	0.0082
HC13	1	17.9483072	0.0082
ADATE_17	1	17.8133883	0.0081
MSA	1	17.494939	0.0080



STATEGOV	0	0	0.0000
FEDGOV	0	0	0.0000
SOLP3	0	0	0.0000
SOLIH	0	0	0.0000
MAJOR	0	0	0.0000
COLLECT1	0	0	0.0000
CATLG	0	0	0.0000
HOMEE	0	0	0.0000
STEREO	0	0	0.0000
PCOWNERS	0	0	0.0000
PHOTO	0	0	0.0000
CRAFTS	0	0	0.0000
FISHER	0	0	0.0000
GARDENIN	0	0	0.0000
BOATS	0	0	0.0000
WALKER	0	0	0.0000
KIDSTUFF	0	0	0.0000
CARDS	0	0	0.0000
PLATES	0	0	0.0000
LIFESRC	0	0	0.0000
POP902	0	0	0.0000
POP903	0	0	0.0000
POP90C1	0	0	0.0000
POP90C2	0	0	0.0000
POP90C3	0	0	0.0000
POP90C5	0	0	0.0000
ETH1	0	0	0.0000
ETH3	0	0	0.0000
ETH4	0	0	0.0000
ETH5	0	0	0.0000
ETH7	0	0	0.0000
ETH8	0	0	0.0000
ETH9	0	0	0.0000
ETH10	0	0	0.0000
ETH11	0	0	0.0000
ETH12	0	0	0.0000
ETH13	0	0	0.0000
ETH14	0	0	0.0000
ETH15	0	0	0.0000
ETH16	0	0	0.0000
AGE902	0	0	0.0000
AGE904	0	0	0.0000
AGE905	0	0	0.0000
AGE906	0	0	0.0000
AGE907	0	0	0.0000
CHIL1	0	0	0.0000
CHIL2	0	0	0.0000
AGEC1	0	0	0.0000
AGEC2	0	0	0.0000
AGEC3	0	0	0.0000
AGEC4	0	0	0.0000
AGEC5	0	0	0.0000
AGEC6	0	0	0.0000
AGEC7	0	0	0.0000
CHILC1	0	0	0.0000
CHILC2	0	0	0.0000
CHILC3	0	0	0.0000
HHAGE1	0	0	0.0000
HHAGE2	0	0	0.0000
HHAGE3	0	0	0.0000

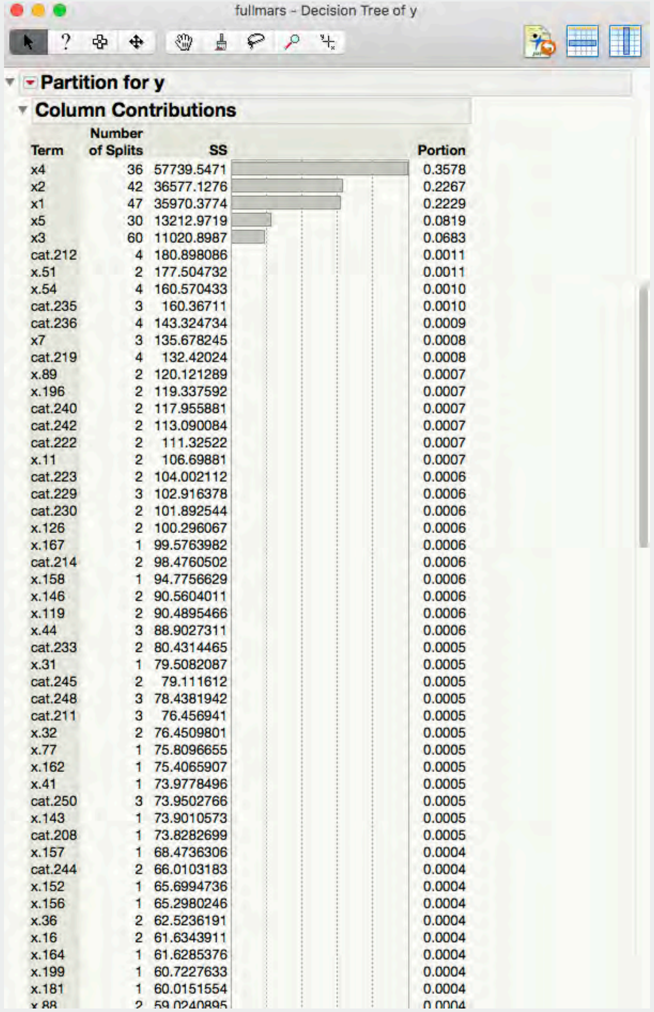
SHAKE THE TREE !!! 250 VARIABLES



	RSquare	RMSE	N	Number of Splits	AICc
Training	0.910	1.4850698	7200	400	26979
Validation	0.764	2.4059219	2400		
Test	0.755	2.4614837	2400		

Split History

Validation Data in Red
Test Data in Orange



Partition for y

Column Contributions

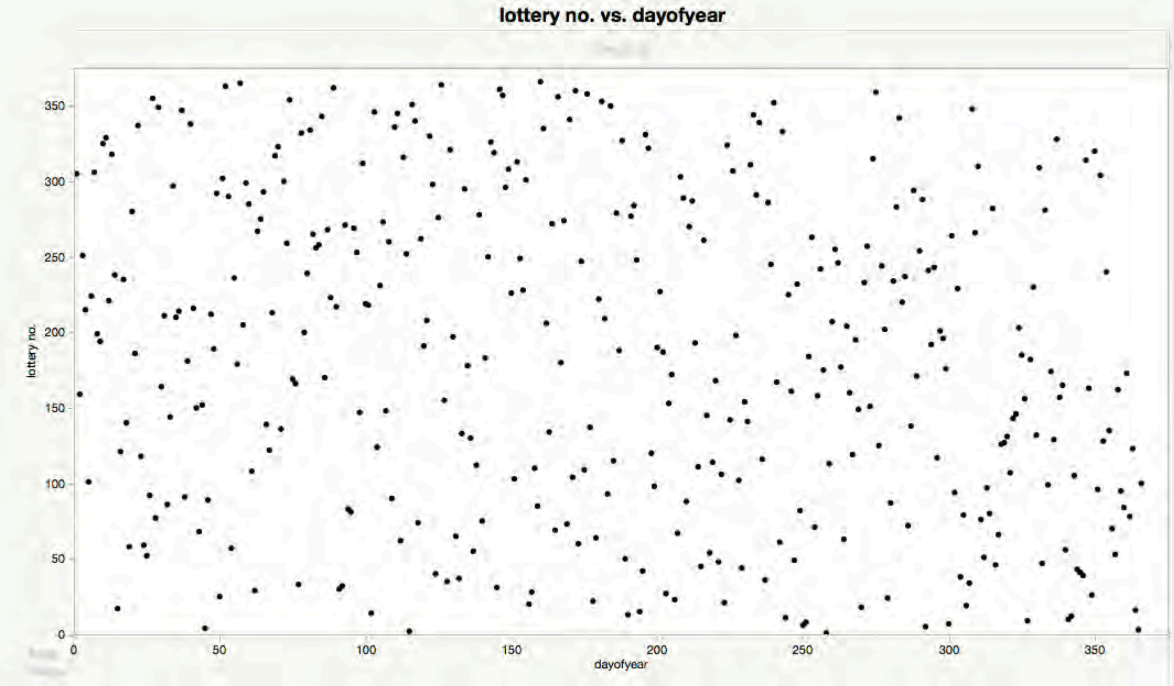
Term	Number of Splits	SS	Portion
x4	36	57739.5471	0.3578
x2	42	36577.1276	0.2267
x1	47	35970.3774	0.2229
x5	30	13212.9719	0.0819
x3	60	11020.8987	0.0683
cat.212	4	180.898086	0.0011
x.51	2	177.504732	0.0011
x.54	4	160.570433	0.0010
cat.235	3	160.36711	0.0010
cat.236	4	143.324734	0.0009
x7	3	135.678245	0.0008
cat.219	4	132.42024	0.0008
x.89	2	120.121289	0.0007
x.196	2	119.337592	0.0007
cat.240	2	117.955881	0.0007
cat.242	2	113.090084	0.0007
cat.222	2	111.32522	0.0007
x.11	2	106.69881	0.0007
cat.223	2	104.002112	0.0006
cat.229	3	102.916378	0.0006
cat.230	2	101.892544	0.0006
x.126	2	100.296067	0.0006
x.167	1	99.5763982	0.0006
cat.214	2	98.4760502	0.0006
x.158	1	94.7756629	0.0006
x.146	2	90.5604011	0.0006
x.119	2	90.4895466	0.0006
x.44	3	88.9027311	0.0006
cat.233	2	80.4314465	0.0005
x.31	1	79.5082087	0.0005
cat.245	2	79.111612	0.0005
cat.248	3	78.4381942	0.0005
cat.211	3	76.456941	0.0005
x.32	2	76.4509801	0.0005
x.77	1	75.8096655	0.0005
x.162	1	75.4065907	0.0005
x.41	1	73.9778496	0.0005
cat.250	3	73.9502766	0.0005
x.143	1	73.9010573	0.0005
cat.208	1	73.8282699	0.0005
x.157	1	68.4736306	0.0004
cat.244	2	66.0103183	0.0004
x.152	1	65.6994736	0.0004
x.156	1	65.2980246	0.0004
x.36	2	62.5236191	0.0004
x.16	2	61.6343911	0.0004
x.164	1	61.6285376	0.0004
x.199	1	60.7227633	0.0004
x.181	1	60.0151554	0.0004
x.88	2	59.0240895	0.0004

TREE ADVANTAGES/DISADVANTAGES

- **Tree are a great first model because**
  - Resistant to outliers
  - Treat missing as a category
  - Can mix continuous/categorical predictors
  - Are fast
  - Can be interpretable
  - Give variable contributions

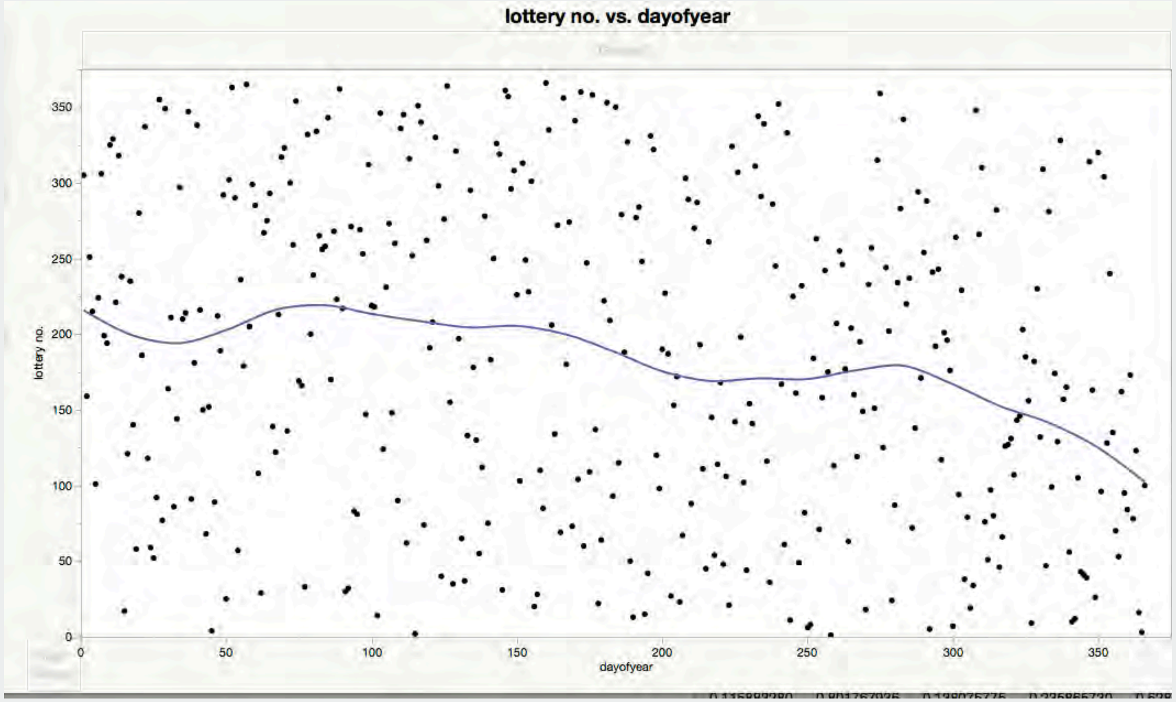
- **But…**
  - Rarely are they a good model
  - Instead, they provide a starting point for more sophisticated models

SMOOTHING – WHAT'S THE TREND?



1970 US Draft Lottery
Lottery Number vs. Day of Year

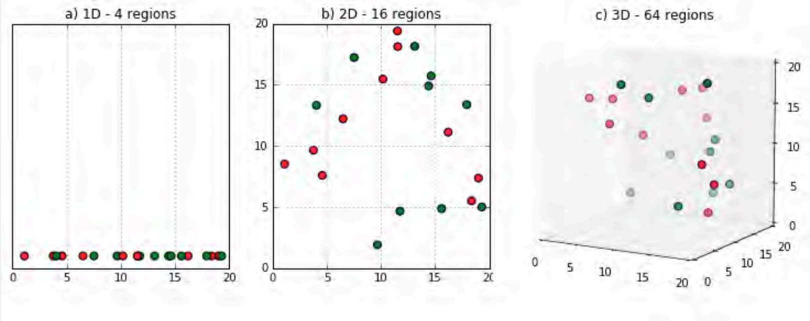
SCATTERPLOT SMOOTHER



1970 US Draft Lottery
Lottery Number vs. Day of Year

MORE DIMENSIONS

- **Why not smooth using 10 predictors?**
  - Curse of dimensionality
  - With 10 predictors, if we use 10% of each as a neighborhood, how many points do we need to get 100 points in cube?
  - Conversely, to get 10% of the points, what percentage do we need to take of each predictor?
  - Need new approach



ADDITIVE MODEL

- Cant get

$$\hat{y} = f\,(x_1,\ldots,x_p)$$

- So, simplify to:

$$\hat{y} = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$$

- Each of the fi are easy to find
  - Scatterplot smoothers

PROBLEMS WITH GAMS

- Interactions are possible, but only order 2 and you have to know which ones

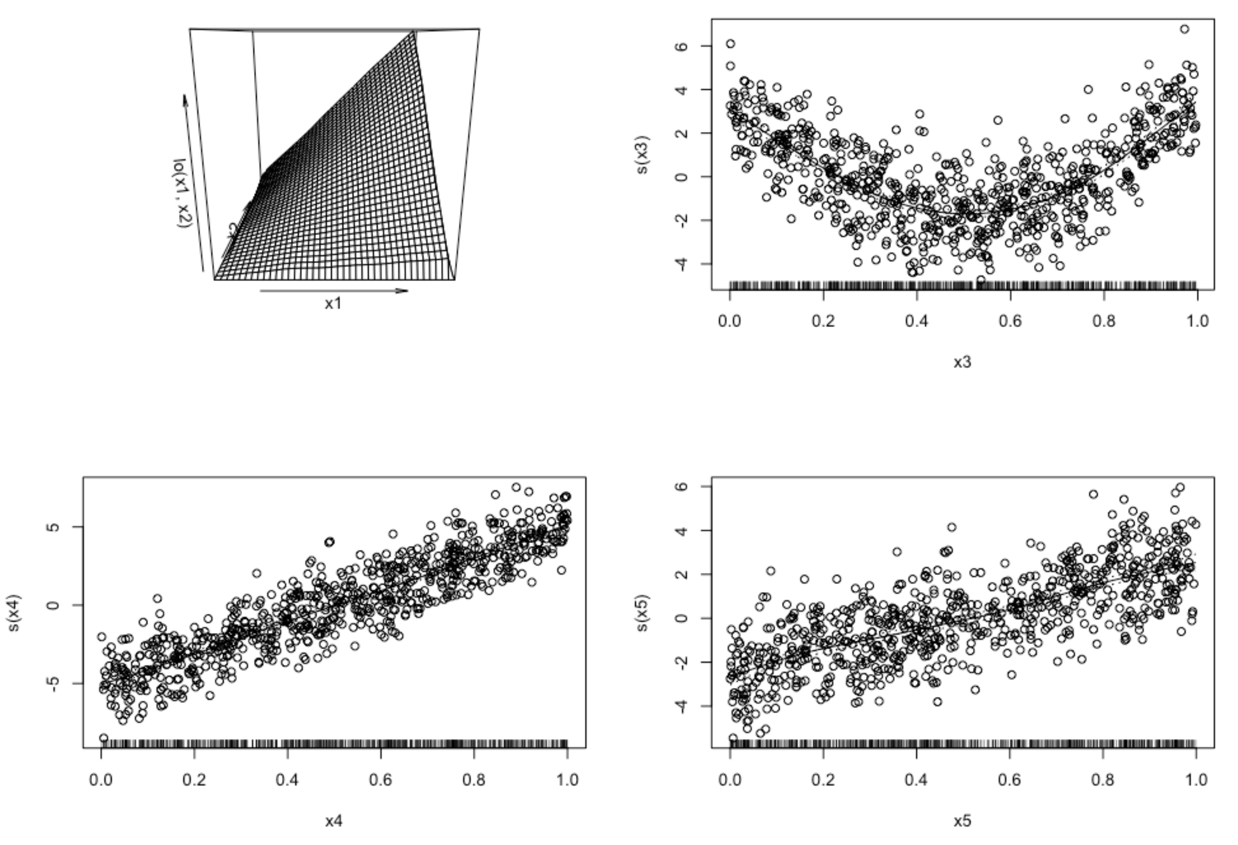
$$\hat{y} = s(x_1) + s(x_2) + s(x_3, x_4) + ...$$

- If I knew it was the first two variables, I could fit:

```
> gam.m1=gam(y~lo(x1,x2)+s(x3)+s(x4)+s(x5),data=subset(marsd,Test==0))
```

```
Anova for Parametric Effects
                 Df Sum Sq Mean Sq  F value Pr(>F)
lo(x1, x2)     2.00 2213.0  1106.5  578.278 <2e-16 ***
s(x3)          1.00    1.0     1.0    0.505 0.4775
s(x4)          1.00 6704.1  6704.1 3503.605 <2e-16 ***
s(x5)          1.00 1684.7  1684.7  880.438 <2e-16 ***
Residuals    729.26 1395.4     1.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

GAM OUTPUT


lo(x1 , x2)
x1
s(x3)
6
4
2
0
-2
-4
0.0
0.2
0.4
0.6
0.8
1.0
x3
s(x4)
5
0
-5
0.0
0.2
0.4
0.6
0.8
1.0
x4
s(x5)
6
4
2
0
-2
-4
0.0
0.2
0.4
0.6
0.8
1.0
x5

NEW STRATEGY

- **Instead of original x's use linear combinations — new "features"**

$$z_j = \alpha_{j0} + \alpha_{j1}x_1 + ... + \alpha_{jp}x_p$$

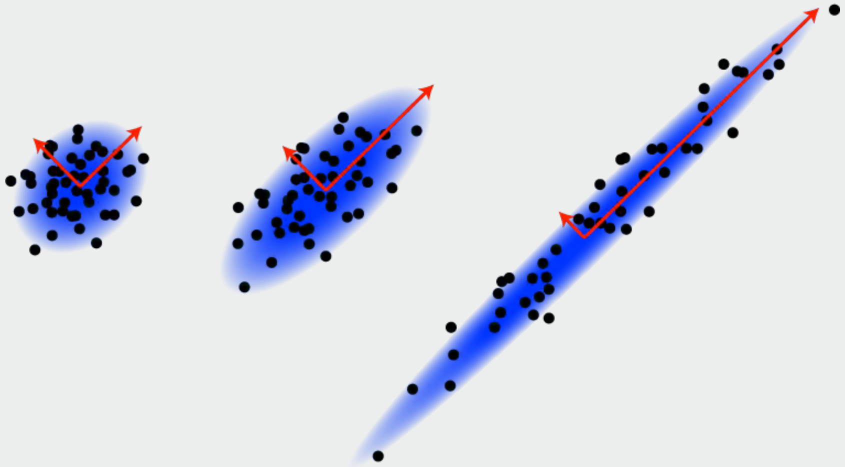
  - Principal components
  - Factor analysis
  - Multidimensional scaling

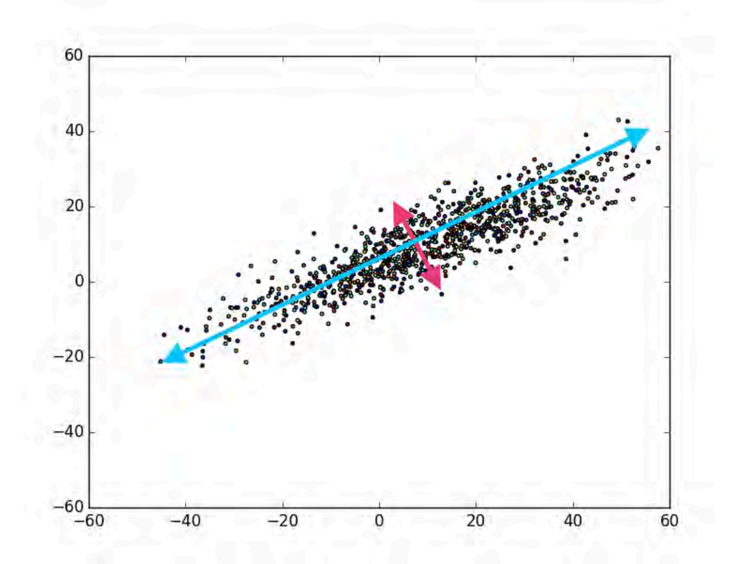
HOW TO FIND THESE FEATURES

- What are interesting directions in the predictors?
  - High variance directions in X - PCA
  - High covariance with Y -- PLS
  - High correlation with Y -- OLS
  - Directions whose smooth is correlated with y - PPR
- Notice which involve y

PRINCIPAL COMPONENTS

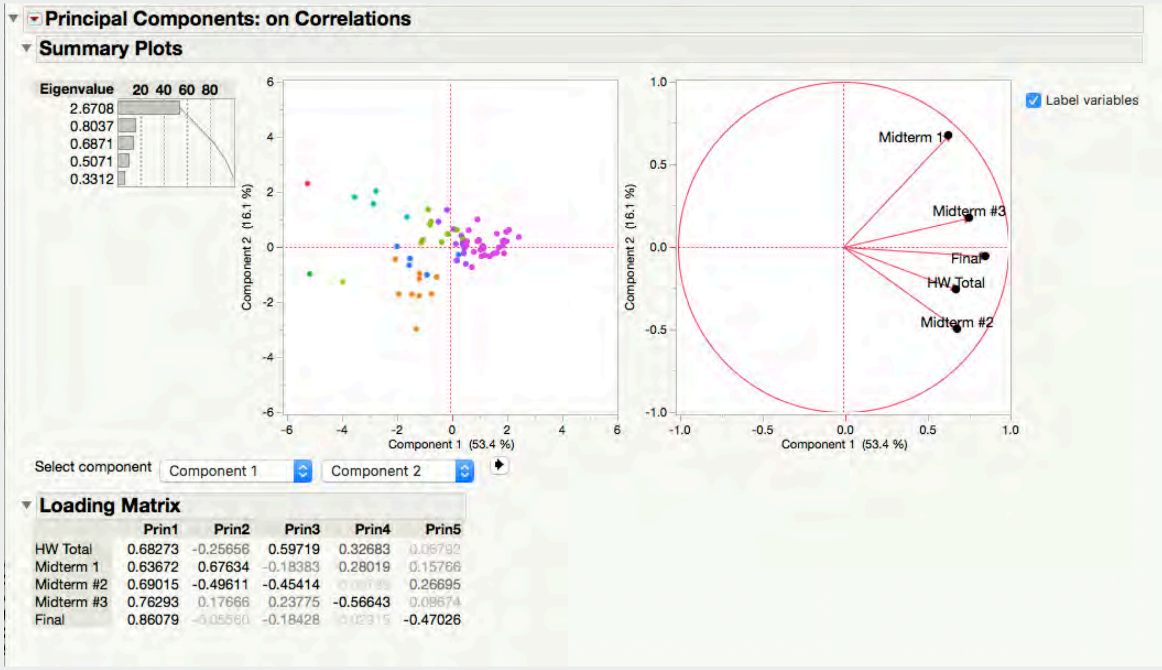
- First direction has maximum variance
- Second direction has maximum variance of all directions perpendicular to first
- Repeat until there are as many directions as original variables
- Reduce to a smaller number
    - Multiple approaches to eliminate directions





WHEN DOES THIS WORK WELL?

- **When you have a group of highly correlated predictor variables**
  - Census information
  - History of past giving
  - 10 temperature sensors
  - Quality control



Principal Components: on Correlations

Summary Plots

Eigenvalue
2.6708
0.8037
0.6871
0.5071
0.3312

Select component: Component 1, Component 2

Loading Matrix

	Prin1	Prin2	Prin3	Prin4	Prin5
HW Total	0.68273	-0.25656	0.59719	0.32683	0.06792
Midterm 1	0.63672	0.67634	-0.18383	0.28019	0.15766
Midterm #2	0.69015	-0.49611	-0.45414	[illegible]	0.26695
Midterm #3	0.76293	0.17666	0.23775	-0.56643	0.09674
Final	0.86079	-0.05560	-0.18428	0.02315	-0.47026

PROJECTION PURSUIT

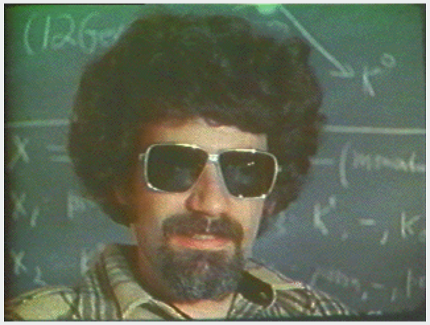
- Instead of original x's use linear combinations — new "features"

$$z_j = \alpha_{j0} + \alpha_{j1}x_1 + ... + \alpha_{jp}x_p$$

  - Now, directions are just "interesting"
  - Minimize sum of squares of residuals
  - Interpretation is a problem

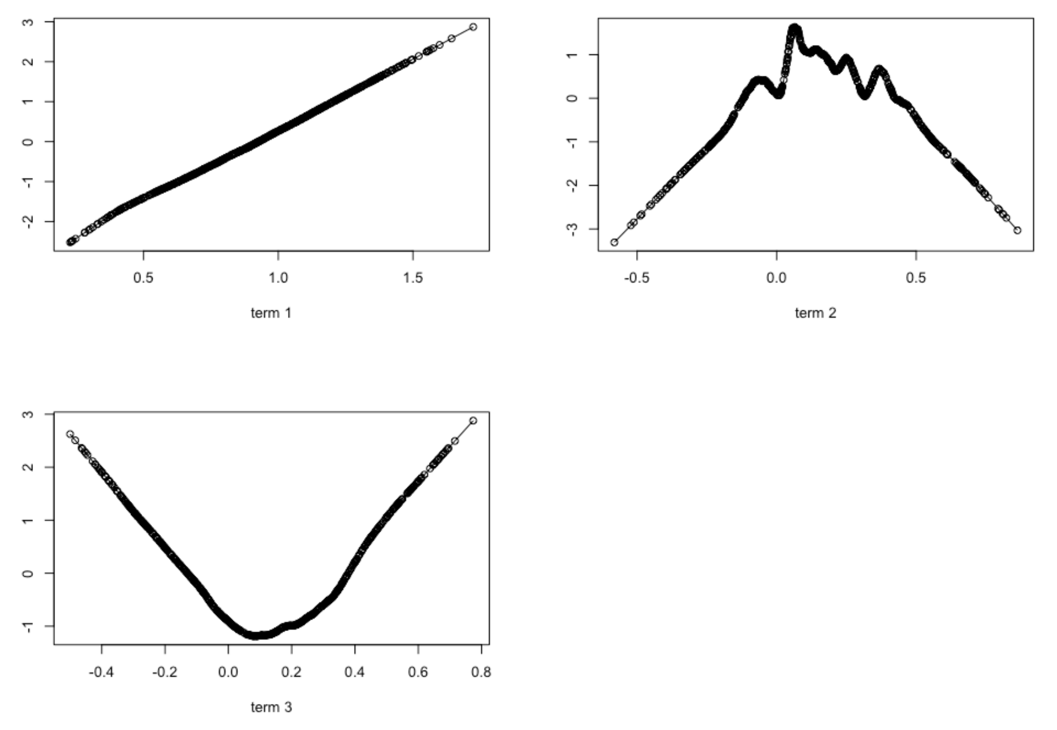
- Once found, fit the following model:

$$\hat{y} = s(z_1) + s(z_2) + s(z_3) + ...$$



PROJECTION PURSUIT OUTPUT

- ppr1=ppr(y~. - Test, nterms=3, data=subset(mars,Test==0))



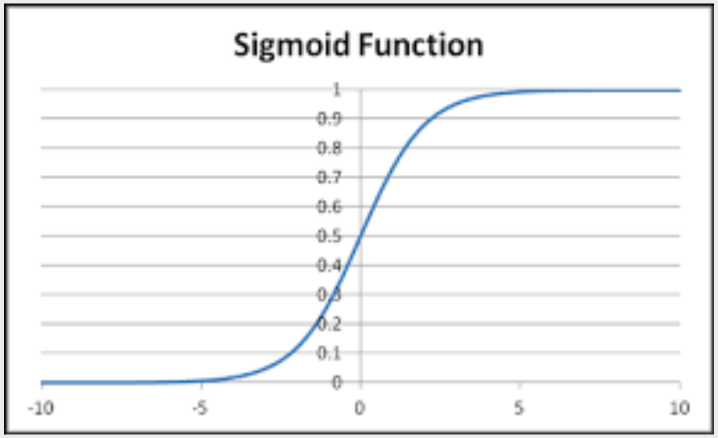
```
Goodness of fit:
3 terms
1615.42

Projection direction vectors:
    term 1       term 2       term 3
x1   0.342907504 -0.659698547 -0.108766485
x2   0.334347163  0.718257325 -0.150105151
x3   0.009388934 -0.058771539  0.964990347
x4   0.788543823  0.099639752 -0.078530289
x5   0.384260285  0.117609297 -0.070077780
x6   0.016657236 -0.038726581 -0.077071263
x7  -0.011492871  0.130117053 -0.069838552
x8  -0.024787784 -0.026914931 -0.092504698
x9  -0.005044944  0.050158835 -0.036769432
x10 -0.005612762 -0.004427670 -0.051259248

Coefficients of ridge terms:
  term 1   term 2   term 3
3.725415 0.668497 1.501681
```

ONE MORE VARIANT

- Instead of arbitrary smooth functions, restrict to sigmoidal
  Any cdf will work



NEURAL NETWORK

- Instead of arbitrary smooth functions, restrict to sigmoidal
- As before,

$$z_j = \alpha_{j0} + \alpha_{j1} x_1 + ... + \alpha_{jp} x_p$$

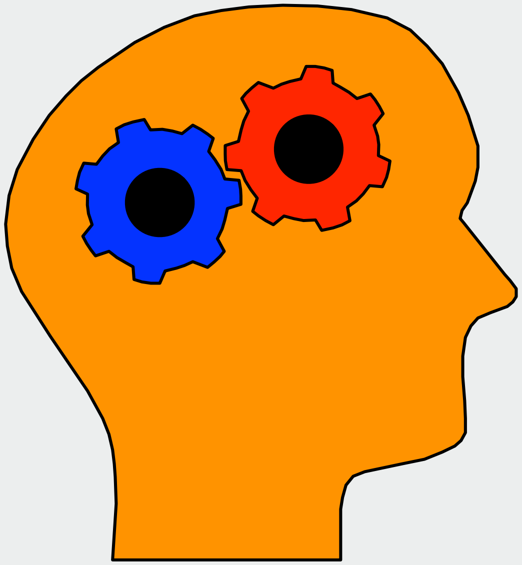
- and

$$\hat{y} = s(z_1) + s(z_2) + s(z_3) + ...$$

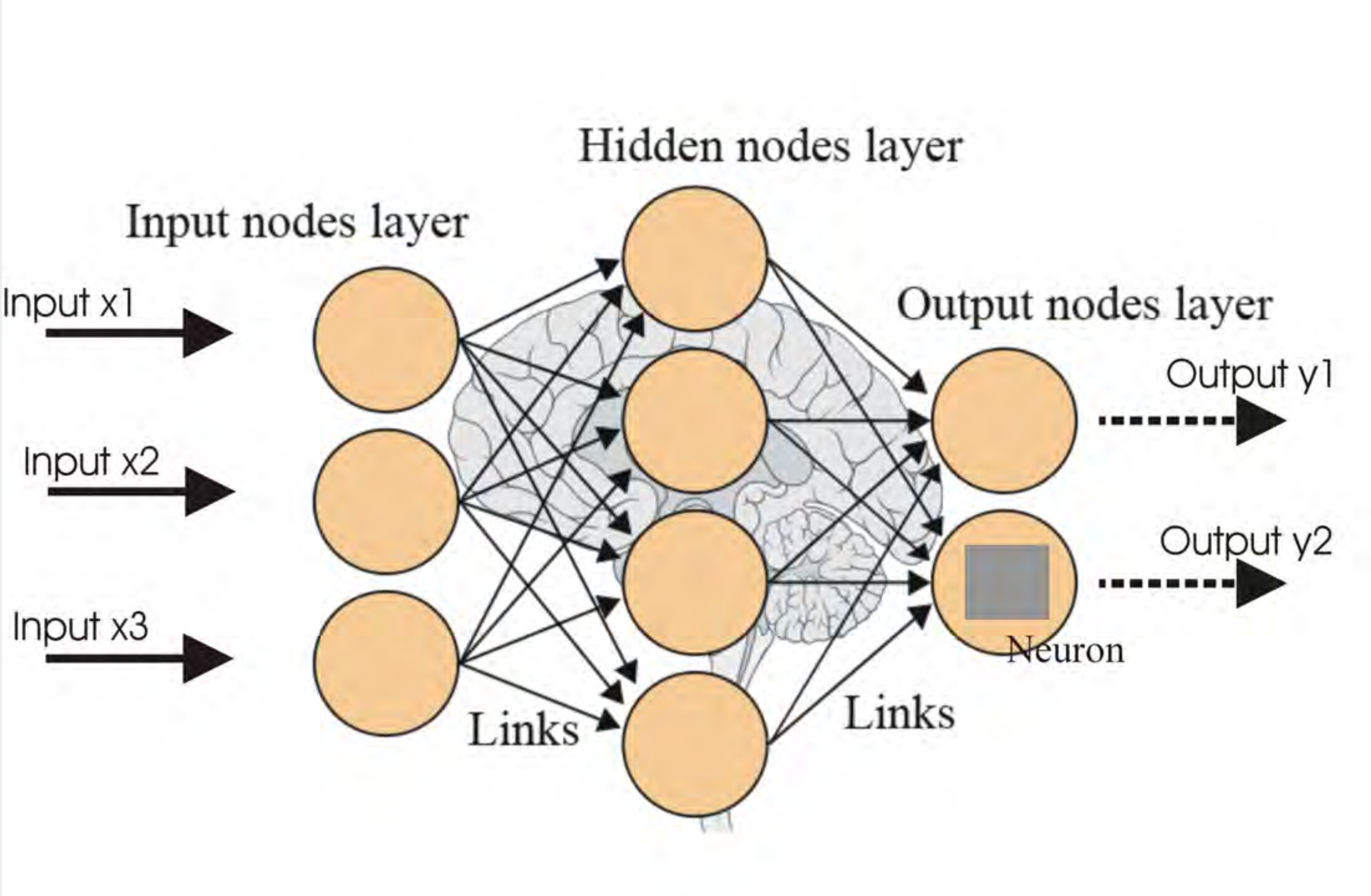
- but now s( ) are sigmoids

NEURAL NETWORKS

- **Don’t resemble the brain**
  - LOTS of hype
  - Are a statistical model
  - Closest relative is project pursuit regression
  - Generally, fit really well (sometimes too well)

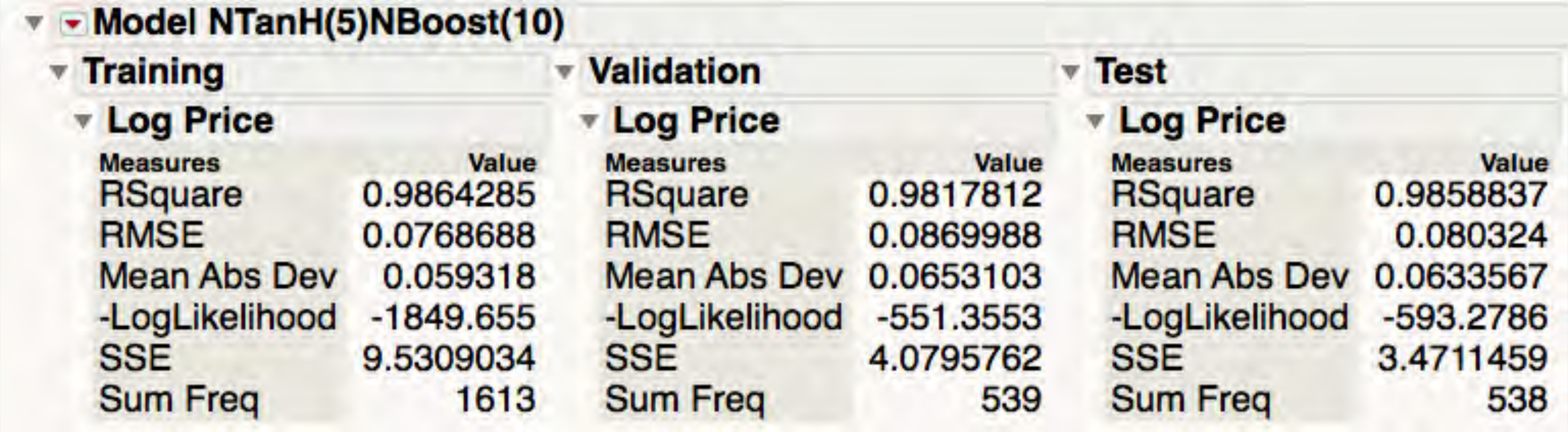


NEURAL NETWORK



NEURAL NETWORK

NEURAL RESULTS



Model NTanH(5)NBoost(10)

Training

Log Price

Measures	Value
RSquare	0.9864285
RMSE	0.0768688
Mean Abs Dev	0.059318
-LogLikelihood	-1849.655
SSE	9.5309034
Sum Freq	1613

Validation

Log Price

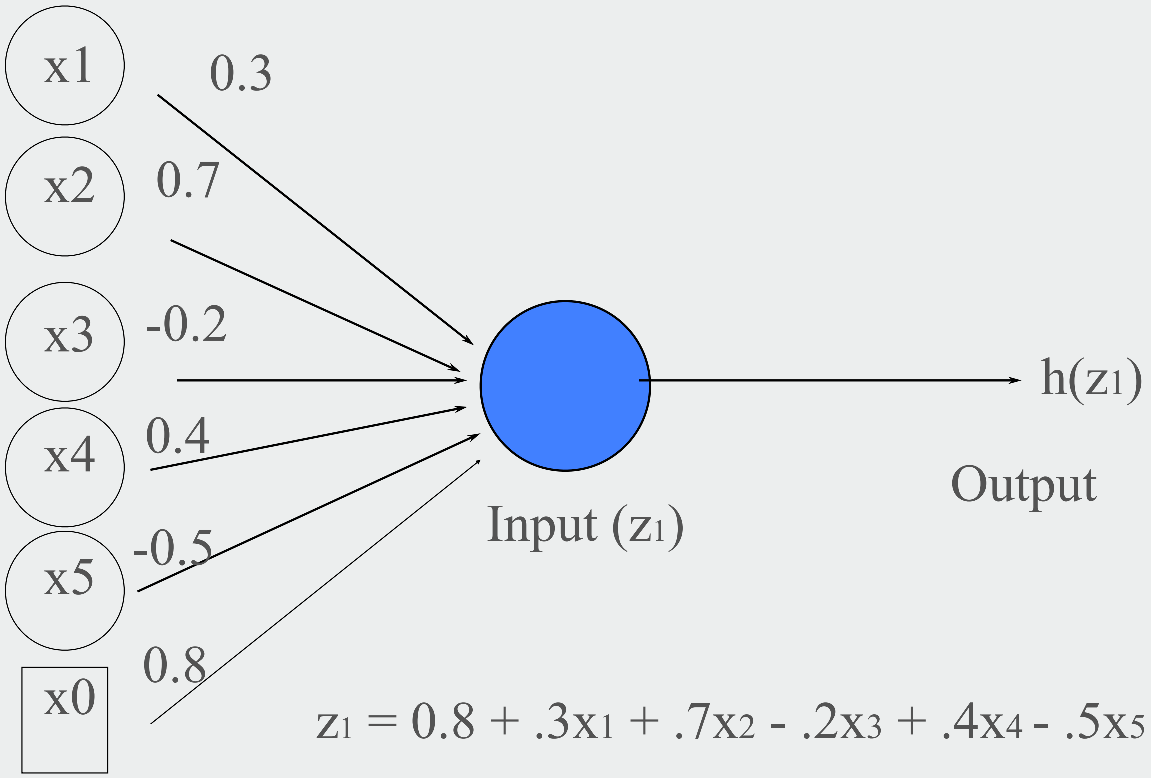
Measures	Value
RSquare	0.9817812
RMSE	0.0869988
Mean Abs Dev	0.0653103
-LogLikelihood	-551.3553
SSE	4.0795762
Sum Freq	539

Test

Log Price

Measures	Value
RSquare	0.9858837
RMSE	0.080324
Mean Abs Dev	0.0633567
-LogLikelihood	-593.2786
SSE	3.4711459
Sum Freq	538

A SINGLE NEURON



$$z_1 = 0.8 + .3x_1 + .7x_2 - .2x_3 + .4x_4 - .5x_5$$

SINGLE NODE

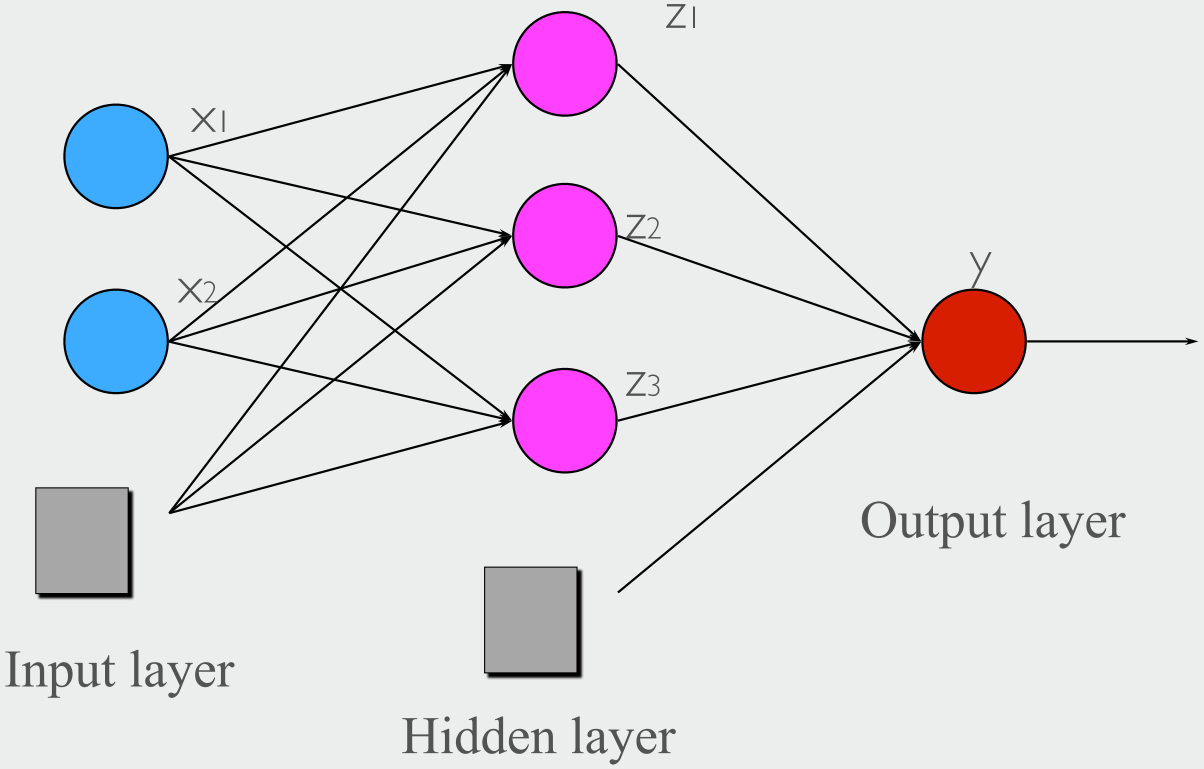
Input to outer layer from "hidden node":

$$I = z_l = \sum_j w_{1jk} x_j + \theta_l$$

Output:

$$\hat{y}_k = h(z_{kl})$$

LAYERED ARCHITECTURE



PUTTING IT TOGETHER

Create lots of features – hidden nodes

$$z_l = \sum_j w_{1jk} x_j + \theta_l$$

Use them in an additive model:

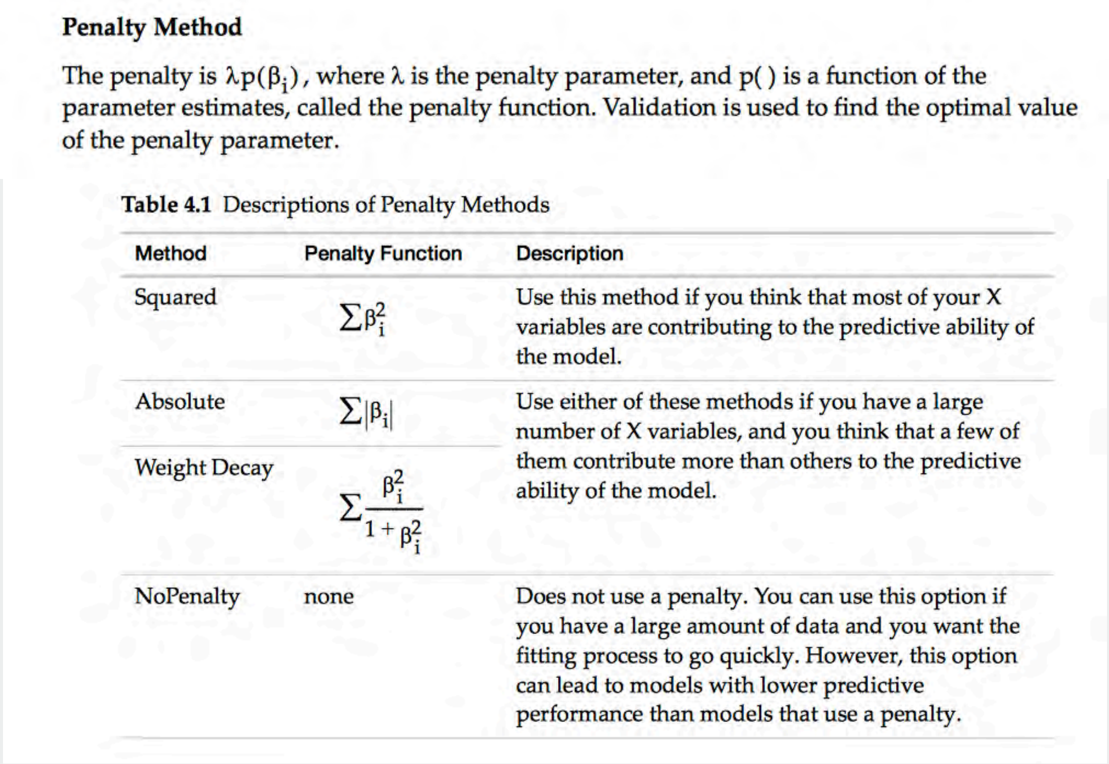
$$\hat{y}_k = w_{21}\, h(z_1) + w_{22}\, h(z_2) + \ldots + \theta_j$$

FINALLY

$$\hat{y}_k = \tilde{h}(\sum_l w_{2kl} \;\; h(\sum_j w_{1jk} x_j + \theta_l) + \theta_j)$$

The resulting model is just a flexible non-linear regression of the response on a set of predictor variables.

NEED TO REGULARIZE



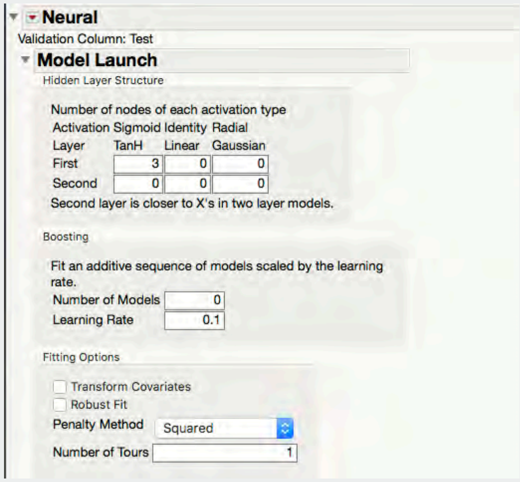
Penalty Method

The penalty is $\lambda p(\beta_i)$, where λ is the penalty parameter, and p( ) is a function of the parameter estimates, called the penalty function. Validation is used to find the optimal value of the penalty parameter.

Table 4.1 Descriptions of Penalty Methods

Method	Penalty Function	Description
Squared	$\sum \beta_i^2$	Use this method if you think that most of your X variables are contributing to the predictive ability of the model.
Absolute	$\sum \lvert\beta_i\rvert$	Use either of these methods if you have a large number of X variables, and you think that a few of them contribute more than others to the predictive ability of the model.
Weight Decay	$\sum \frac{\beta_i^2}{1+\beta_i^2}$	
NoPenalty	none	Does not use a penalty. You can use this option if you have a large amount of data and you want the fitting process to go quickly. However, this option can lead to models with lower predictive performance than models that use a penalty.

ON OUR DATA



Neural

Validation Column: Test

Model Launch

Hidden Layer Structure

Number of nodes of each activation type

Activation	Sigmoid	Identity	Radial
Layer	TanH	Linear	Gaussian
First	3	0	0
Second	0	0	0

Second layer is closer to X's in two layer models.

Boosting

Fit an additive sequence of models scaled by the learning rate.

Number of Models: 0

Learning Rate: 0.1

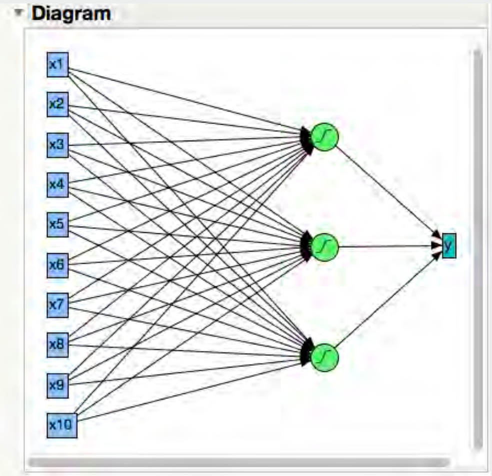
Fitting Options

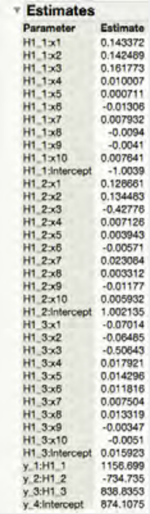
☐ Transform Covariates

☐ Robust Fit

Penalty Method: Squared

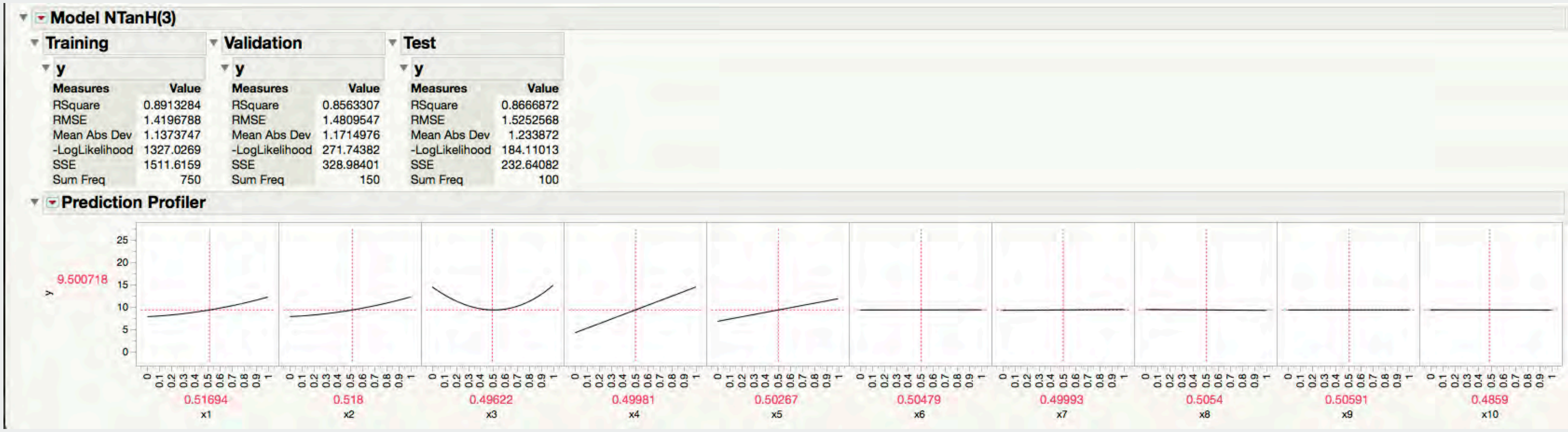
Number of Tours: 1





Estimates

Parameter	Estimate
H1_1:x1	0.143372
H1_1:x2	0.142469
H1_1:x3	0.161773
H1_1:x4	0.010007
H1_1:x5	0.000711
H1_1:x6	-0.01306
H1_1:x7	0.007932
H1_1:x8	-0.0094
H1_1:x9	-0.0041
H1_1:x10	0.007641
H1_1:Intercept	-1.0039
H1_2:x1	0.128661
H1_2:x2	0.134463
H1_2:x3	-0.42776
H1_2:x4	0.007126
H1_2:x5	0.003943
H1_2:x6	-0.00571
H1_2:x7	0.023064
H1_2:x8	0.003312
H1_2:x9	-0.01177
H1_2:x10	0.005932
H1_2:Intercept	1.002135
H1_3:x1	-0.07014
H1_3:x2	-0.06465
H1_3:x3	-0.50643
H1_3:x4	0.017921
H1_3:x5	0.014296
H1_3:x6	0.011816
H1_3:x7	0.007504
H1_3:x8	0.013319
H1_3:x9	-0.00347
H1_3:x10	-0.0051
H1_3:Intercept	0.015923
y_1:H1_1	1156.699
y_2:H1_2	-734.735
y_3:H1_3	838.8353
y_4:Intercept	874.1075



Model NTanH(3)

Training

y

Measures	Value
RSquare	0.8913284
RMSE	1.4196788
Mean Abs Dev	1.1373747
-LogLikelihood	1327.0269
SSE	1511.6159
Sum Freq	750

Validation

y

Measures	Value
RSquare	0.8563307
RMSE	1.4809547
Mean Abs Dev	1.1714976
-LogLikelihood	271.74382
SSE	328.98401
Sum Freq	150

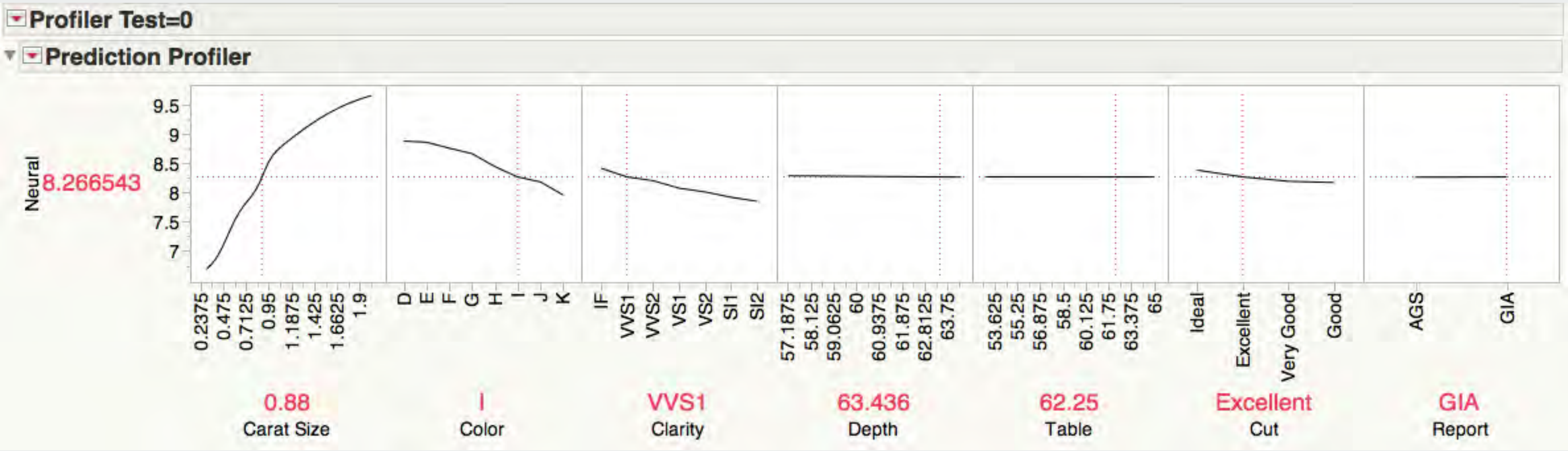
Test

y

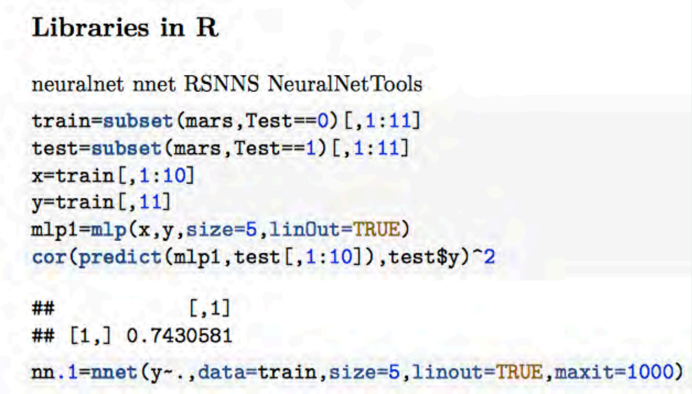
Measures	Value
RSquare	0.8666872
RMSE	1.5252568
Mean Abs Dev	1.233872
-LogLikelihood	184.11013
SSE	232.64082
Sum Freq	100

Prediction Profiler

PROFILER ON DIAMOND DATA



IN **R** — NO PENALTY



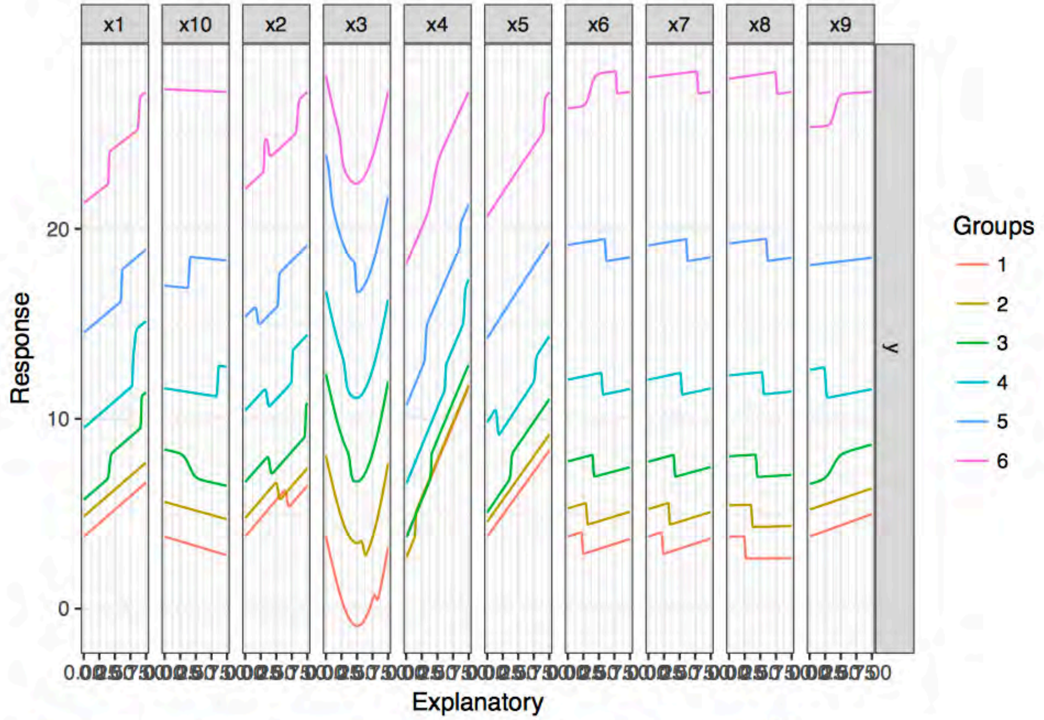
Libraries in R

neuralnet nnet RSNNS NeuralNetTools

```
train=subset(mars,Test==0)[,1:11]
test=subset(mars,Test==1)[,1:11]
x=train[,1:10]
y=train[,11]
mlp1=mlp(x,y,size=5,linOut=TRUE)
cor(predict(mlp1,test[,1:10]),test$y)^2
```

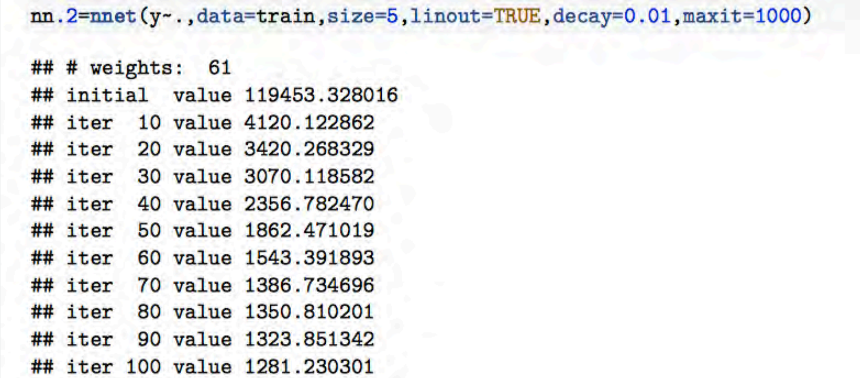
```
##              [,1]
## [1,] 0.7430581
```

```
nn.1=nnet(y~.,data=train,size=5,linout=TRUE,maxit=1000)
```



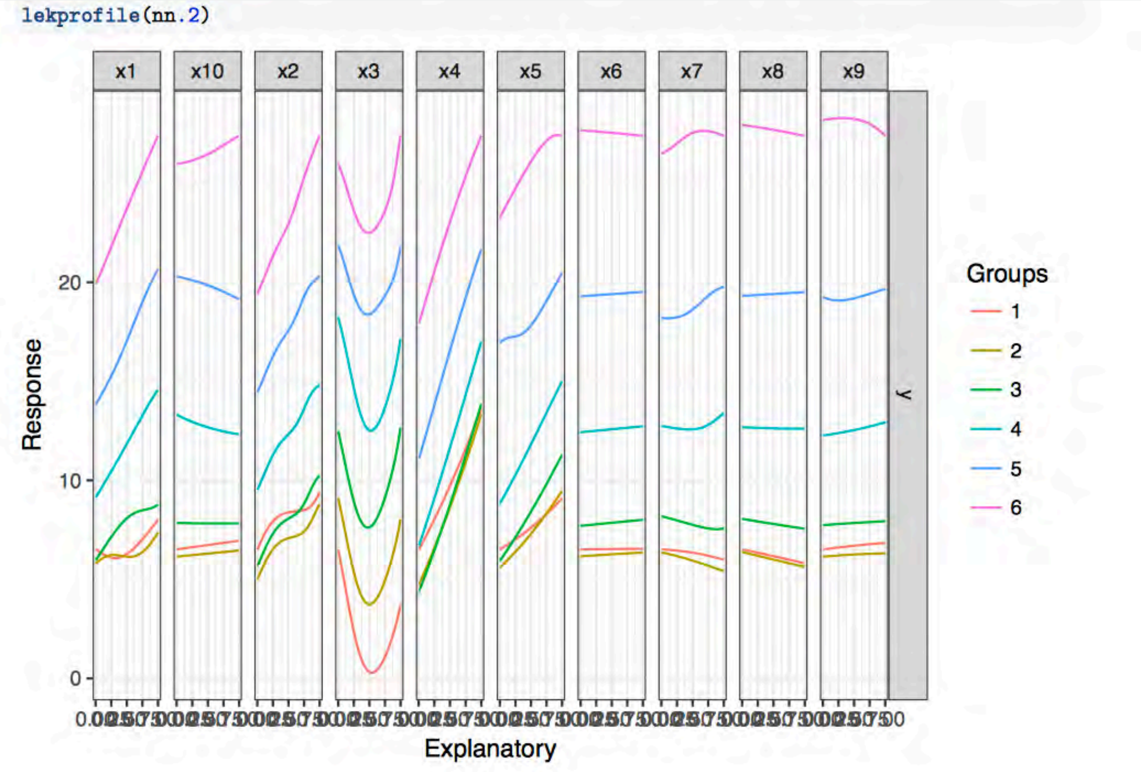
```
nn.2=nnet(y~.,data=train,size=5,linout=TRUE,decay=0.01,maxit=1000)
```

IN **R** — WITH WEIGHT DECAY



```
nn.2=nnet(y~.,data=train,size=5,linout=TRUE,decay=0.01,maxit=1000)

## # weights:  61
## initial  value 119453.328016
## iter  10 value 4120.122862
## iter  20 value 3420.268329
## iter  30 value 3070.118582
## iter  40 value 2356.782470
## iter  50 value 1862.471019
## iter  60 value 1543.391893
## iter  70 value 1386.734696
## iter  80 value 1350.810201
## iter  90 value 1323.851342
## iter 100 value 1281.230301
```



```
lekprofile(nn.2)
```

NEURAL NET PRO

- **Advantages**
    - Handles continuous or discrete values
    - Complex interactions
    - In general, highly accurate for fitting due to flexibility of model
    - Can incorporate known relationships
        - So called grey box models
        - See De Veaux et al, Environmetrics 1999

NEURAL NET CON

- **Disadvantages**
    - Model is not descriptive (black box)
    - Difficult, complex architectures
    - Slow model building
    - Categorical data explosion
    - Sensitive to input variable selection

K-NEAREST NEIGHBORS(KNN)

- To predict y for an x:
  - Find the k most similar x's
  - Average their y's
- Find k by cross validation
- No training (estimation) required
- Works embarrassingly well
  - Friedman, KDDM 1996

COLLABORATIVE FILTERING

- Goal: predict what movies people will like
- Data: list of movies each person has watched

```
Lyle    André, Starwars
Ellen   André, Starwars, Amélie Poulin
Fred    Starwars, Batman
Dean    Starwars, Batman, Rambo
Jason   Amélie Poulin, Caché
```

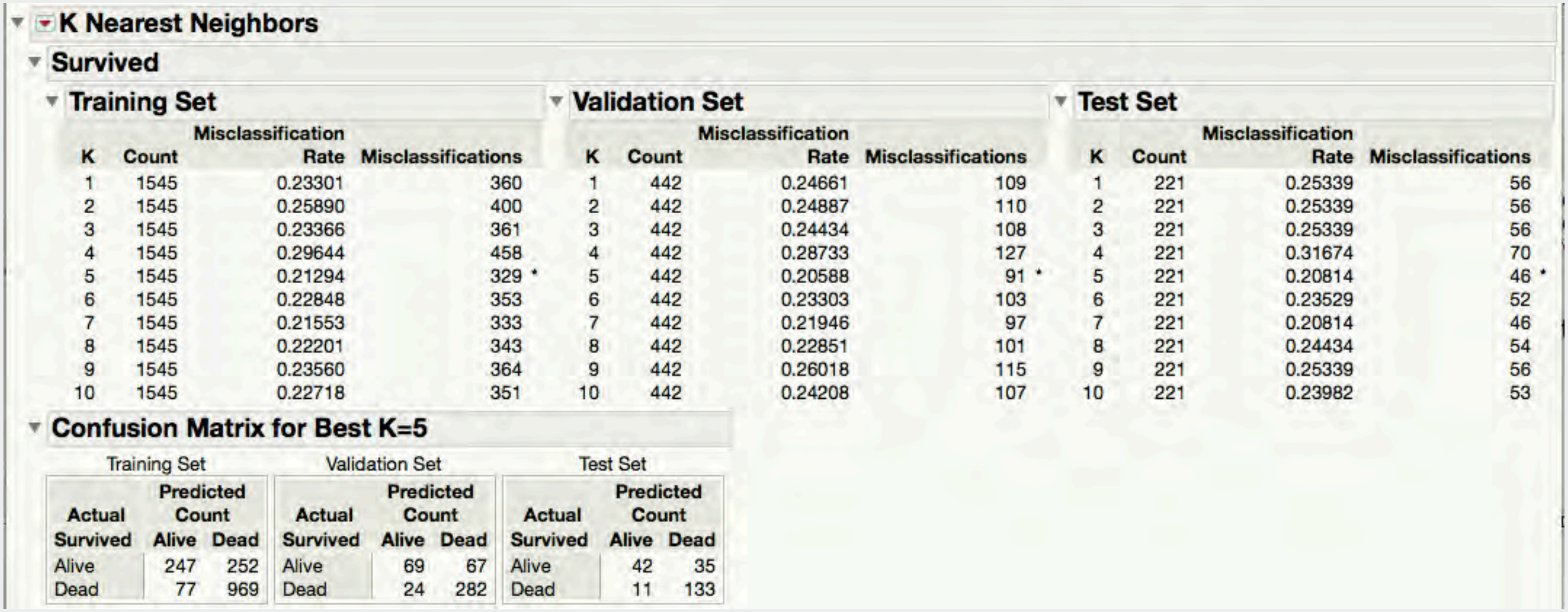
MINI NETFLIX

- Data can be represented as a sparse matrix

	Starwars	**Rambo**	**Batman**	**My Dinner w/André**	**Amélie Poulin**	**Caché**
Lyle	**y**			**y**		
Ellen	**y**			**y**	**y**	
Fred	**y**	**y**				
Dean	**y**	**y**	**y**			
Jason	**y**			**y**		**y**
Karen	**?**	**?**	**?**	**?**	**?**	**?**

- Karen likes André. What else might she like?
- CDNow doubled e-mail responses

KNN ON TITANIC



K Nearest Neighbors

Survived

Training Set

K	Count	Misclassification Rate	Misclassifications
1	1545	0.23301	360
2	1545	0.25890	400
3	1545	0.23366	361
4	1545	0.29644	458
5	1545	0.21294	329 *
6	1545	0.22848	353
7	1545	0.21553	333
8	1545	0.22201	343
9	1545	0.23560	364
10	1545	0.22718	351

Validation Set

K	Count	Misclassification Rate	Misclassifications
1	442	0.24661	109
2	442	0.24887	110
3	442	0.24434	108
4	442	0.28733	127
5	442	0.20588	91 *
6	442	0.23303	103
7	442	0.21946	97
8	442	0.22851	101
9	442	0.26018	115
10	442	0.24208	107

Test Set

K	Count	Misclassification Rate	Misclassifications
1	221	0.25339	56
2	221	0.25339	56
3	221	0.25339	56
4	221	0.31674	70
5	221	0.20814	46 *
6	221	0.23529	52
7	221	0.20814	46
8	221	0.24434	54
9	221	0.25339	56
10	221	0.23982	53

Confusion Matrix for Best K=5

Training Set

Actual Survived	Predicted Count Alive	Dead
Alive	247	252
Dead	77	969

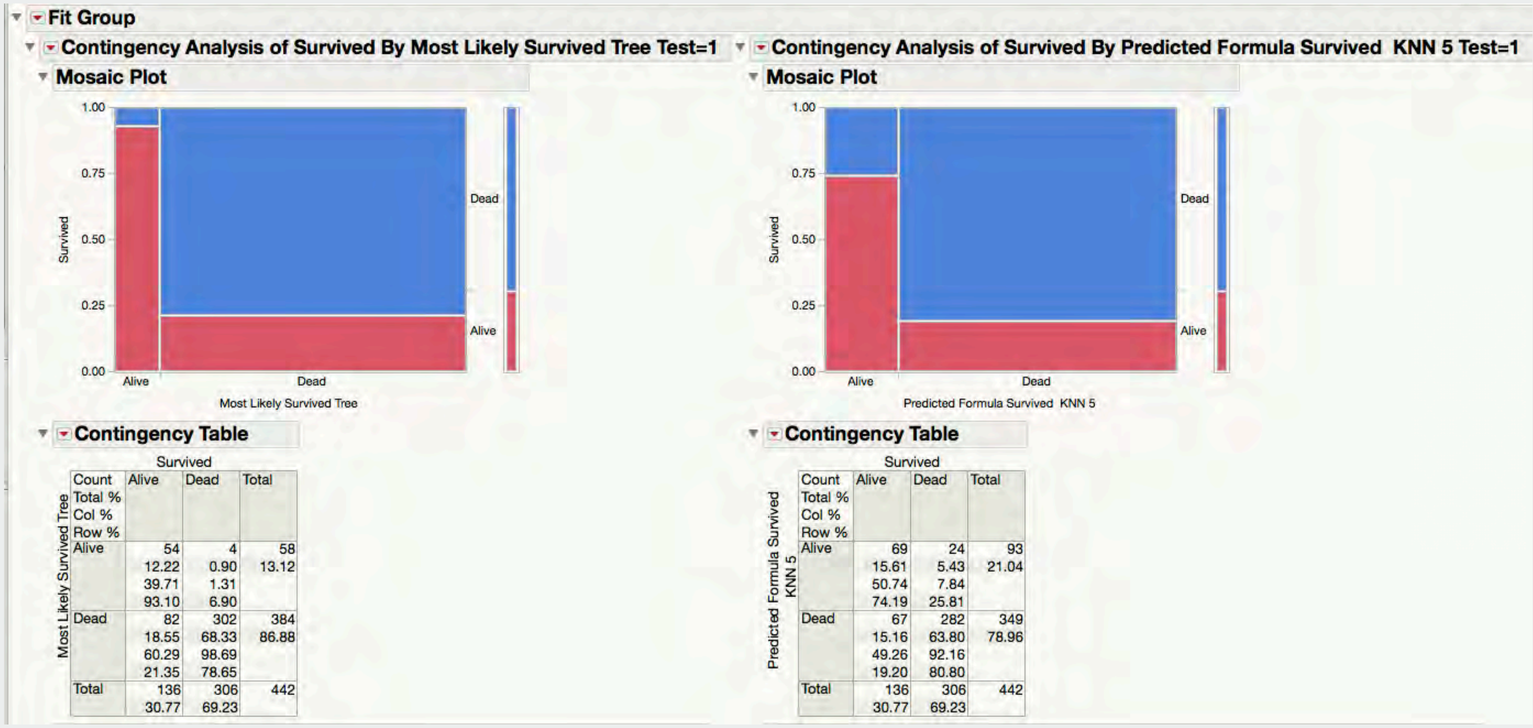
Validation Set

Actual Survived	Predicted Count Alive	Dead
Alive	69	67
Dead	24	282

Test Set

Actual Survived	Predicted Count Alive	Dead
Alive	42	35
Dead	11	133

KNN AND TREE ON TITANIC



Fit Group

Contingency Analysis of Survived By Most Likely Survived Tree Test=1

Mosaic Plot

Contingency Table

Most Likely Survived Tree by Survived

Count Total % Col % Row %	Alive	Dead	Total
Alive	54 12.22 39.71 93.10	4 0.90 1.31 6.90	58 13.12
Dead	82 18.55 60.29 21.35	302 68.33 98.69 78.65	384 86.88
Total	136 30.77	306 69.23	442

Contingency Analysis of Survived By Predicted Formula Survived KNN 5 Test=1

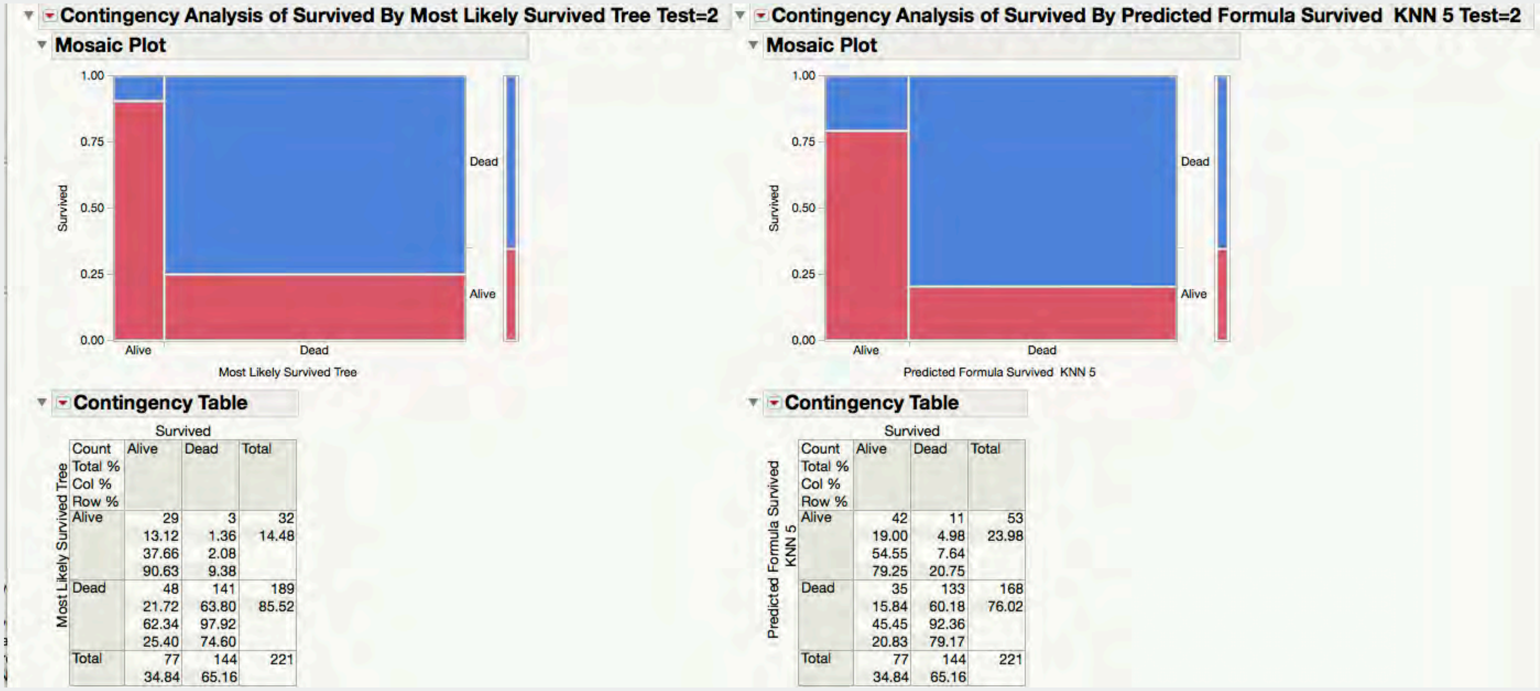
Mosaic Plot

Contingency Table

Predicted Formula Survived KNN 5 by Survived

Count Total % Col % Row %	Alive	Dead	Total
Alive	69 15.61 50.74 74.19	24 5.43 7.84 25.81	93 21.04
Dead	67 15.16 49.26 19.20	282 63.80 92.16 80.80	349 78.96
Total	136 30.77	306 69.23	442

KNN AND TREE ON TITANIC REDUX



Contingency Analysis of Survived By Most Likely Survived Tree Test=2

Contingency Table

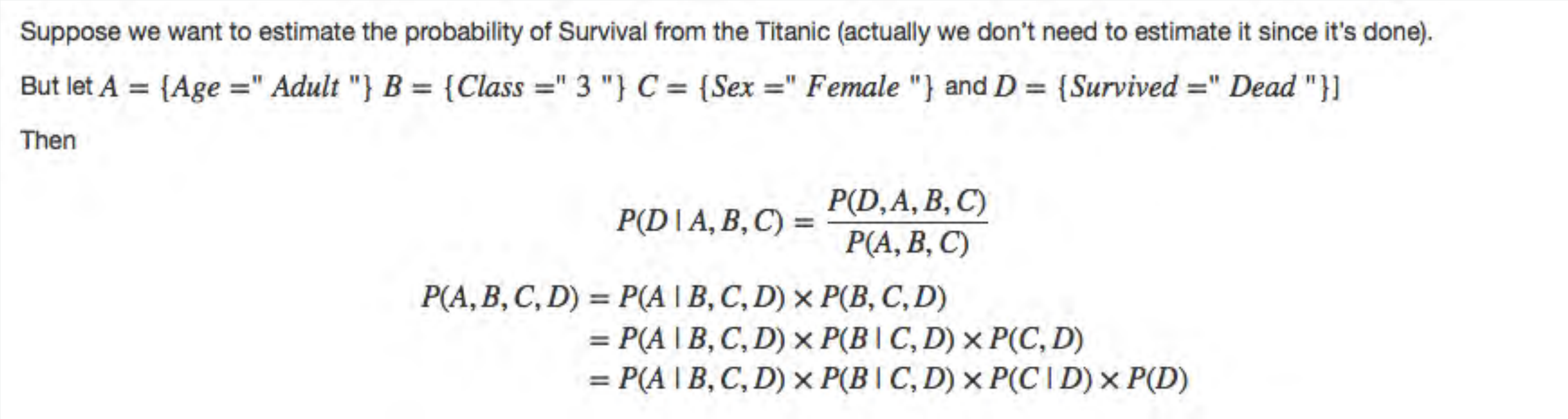
		Survived		
	Count Total % Col % Row %	Alive	Dead	Total
Most Likely Survived Tree	Alive	29 13.12 37.66 90.63	3 1.36 2.08 9.38	32 14.48
	Dead	48 21.72 62.34 25.40	141 63.80 97.92 74.60	189 85.52
	Total	77 34.84	144 65.16	221

Contingency Analysis of Survived By Predicted Formula Survived KNN 5 Test=2

Contingency Table

		Survived		
	Count Total % Col % Row %	Alive	Dead	Total
Predicted Formula Survived KNN 5	Alive	42 19.00 54.55 79.25	11 4.98 7.64 20.75	53 23.98
	Dead	35 15.84 45.45 20.83	133 60.18 92.36 79.17	168 76.02
	Total	77 34.84	144 65.16	221

BAYES



Suppose we want to estimate the probability of Survival from the Titanic (actually we don't need to estimate it since it's done).

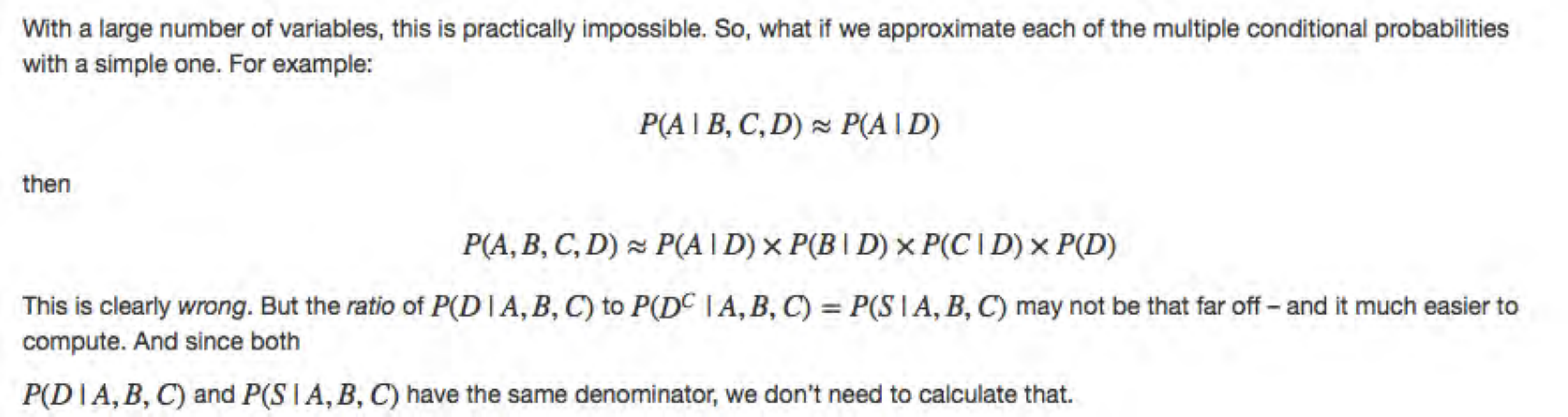
But let $A = \{Age =" Adult "\}$ $B = \{Class =" 3 "\}$ $C = \{Sex =" Female "\}$ and $D = \{Survived =" Dead "\}]$

Then

$$P(D \mid A, B, C) = \frac{P(D, A, B, C)}{P(A, B, C)}$$

$$\begin{aligned} P(A, B, C, D) &= P(A \mid B, C, D) \times P(B, C, D) \\ &= P(A \mid B, C, D) \times P(B \mid C, D) \times P(C, D) \\ &= P(A \mid B, C, D) \times P(B \mid C, D) \times P(C \mid D) \times P(D) \end{aligned}$$

NAÏVE BAYES



With a large number of variables, this is practically impossible. So, what if we approximate each of the multiple conditional probabilities with a simple one. For example:

$$P(A \mid B, C, D) \approx P(A \mid D)$$

then

$$P(A, B, C, D) \approx P(A \mid D) \times P(B \mid D) \times P(C \mid D) \times P(D)$$

This is clearly *wrong*. But the *ratio* of $P(D \mid A, B, C)$ to $P(D^C \mid A, B, C) = P(S \mid A, B, C)$ may not be that far off – and it much easier to compute. And since both

$P(D \mid A, B, C)$ and $P(S \mid A, B, C)$ have the same denominator, we don't need to calculate that.

PERFORMANCE

Naive Bayes

Fit Details

Survived

Training

Count	Misclassification Rate	Misclassifications
1545	0.21424	331

Validation

Count	Misclassification Rate	Misclassifications
442	0.22851	101

Test

Count	Misclassification Rate	Misclassifications
221	0.20814	46

Confusion Matrix

Training

Actual Survived	Predicted Count Alive	Predicted Count Dead
Alive	256	243
Dead	88	958

Validation

Actual Survived	Predicted Count Alive	Predicted Count Dead
Alive	66	70
Dead	31	275

Test

Actual Survived	Predicted Count Alive	Predicted Count Dead
Alive	42	35
Dead	11	133

ROC Curve for Survived = Alive

Training

AUC 0.7287

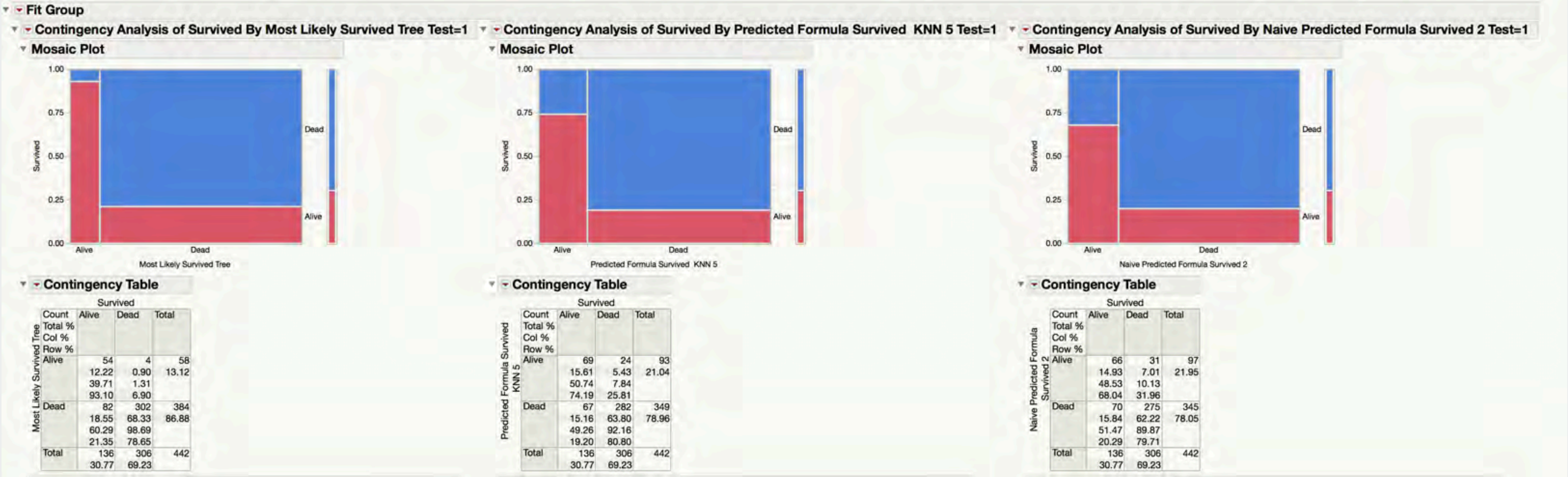
Validation

AUC 0.7023

Test

AUC 0.7158

THREE MODELS



Fit Group

Contingency Analysis of Survived By Most Likely Survived Tree Test=1

Contingency Table

Survived

Most Likely Survived Tree	Count Total % Col % Row %	Alive	Dead	Total
	Alive	54 12.22 39.71 93.10	4 0.90 1.31 6.90	58 13.12
	Dead	82 18.55 60.29 21.35	302 68.33 98.69 78.65	384 86.88
	Total	136 30.77	306 69.23	442

Contingency Analysis of Survived By Predicted Formula Survived KNN 5 Test=1

Contingency Table

Survived

Predicted Formula Survived KNN 5	Count Total % Col % Row %	Alive	Dead	Total
	Alive	69 15.61 50.74 74.19	24 5.43 7.84 25.81	93 21.04
	Dead	67 15.16 49.26 19.20	282 63.80 92.16 80.80	349 78.96
	Total	136 30.77	306 69.23	442

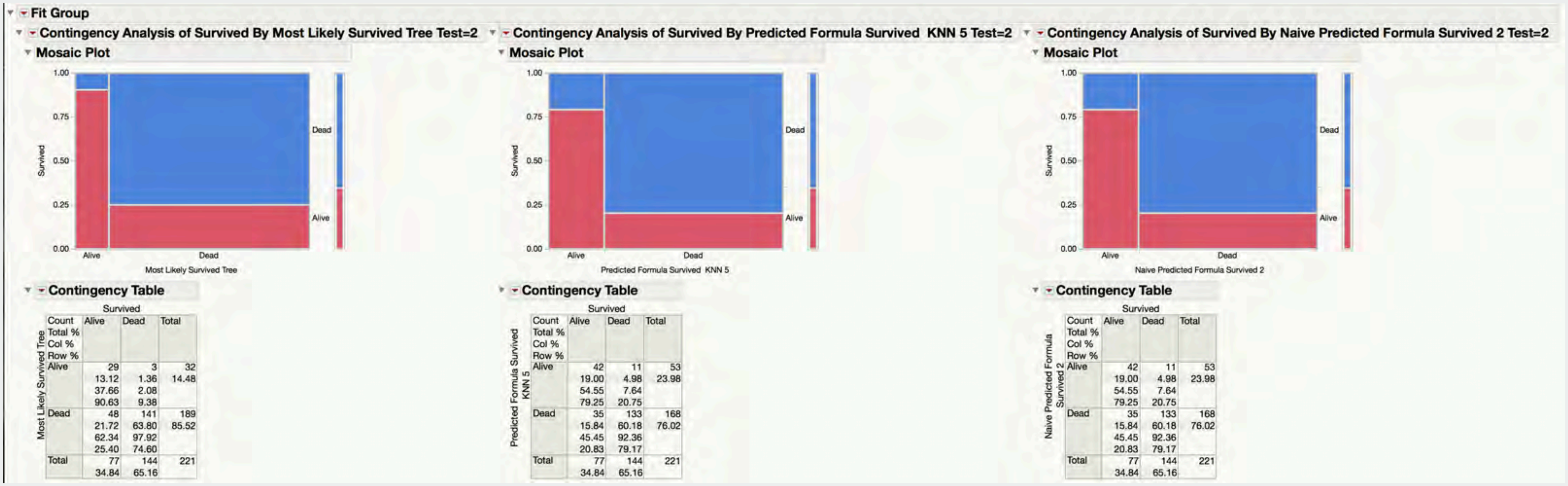
Contingency Analysis of Survived By Naive Predicted Formula Survived 2 Test=1

Contingency Table

Survived

Naive Predicted Formula Survived 2	Count Total % Col % Row %	Alive	Dead	Total
	Alive	66 14.93 48.53 68.04	31 7.01 10.13 31.96	97 21.95
	Dead	70 15.84 51.47 20.29	275 62.22 89.87 79.71	345 78.05
	Total	136 30.77	306 69.23	442

THREE MODELS



Fit Group

Contingency Analysis of Survived By Most Likely Survived Tree Test=2

Mosaic Plot

Contingency Table

Survived

Most Likely Survived Tree

Count Total % Col % Row %	Alive	Dead	Total
Alive	29 13.12 37.66 90.63	3 1.36 2.08 9.38	32 14.48
Dead	48 21.72 62.34 25.40	141 63.80 97.92 74.60	189 85.52
Total	77 34.84	144 65.16	221

Contingency Analysis of Survived By Predicted Formula Survived KNN 5 Test=2

Mosaic Plot

Contingency Table

Survived

Predicted Formula Survived KNN 5

Count Total % Col % Row %	Alive	Dead	Total
Alive	42 19.00 54.55 79.25	11 4.98 7.64 20.75	53 23.98
Dead	35 15.84 45.45 20.83	133 60.18 92.36 79.17	168 76.02
Total	77 34.84	144 65.16	221

Contingency Analysis of Survived By Naive Predicted Formula Survived 2 Test=2

Mosaic Plot

Contingency Table

Survived

Naive Predicted Formula Survived 2

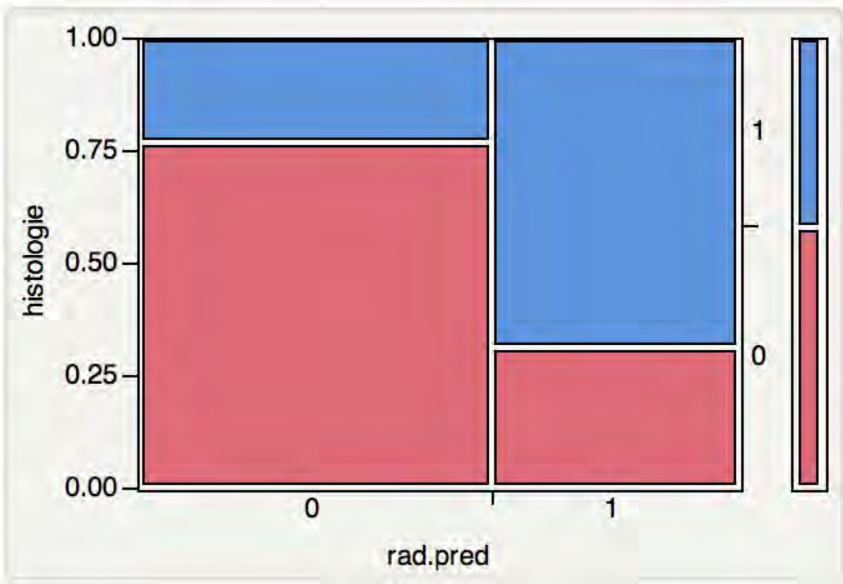
Count Total % Col % Row %	Alive	Dead	Total
Alive	42 19.00 54.55 79.25	11 4.98 7.64 20.75	53 23.98
Dead	35 15.84 45.45 20.83	133 60.18 92.36 79.17	168 76.02
Total	77 34.84	144 65.16	221

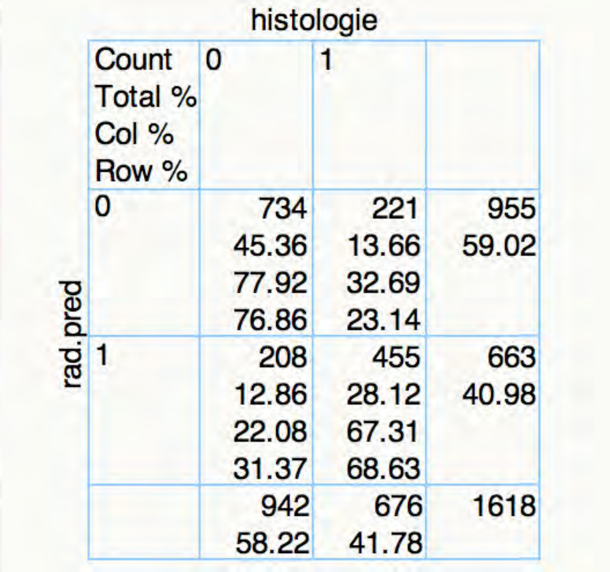
PREDICTING MALIGNANCY

- Very high false positive and false negative rates
  - Approaching 30% and higher — both Type I and Type II errors
  - FDA still says the best way to screen

PREDICTING MALIGNANCY

- ✦ Very high false positive and false negative rates





histologie

rad.pred	Count Total % Col % Row %	0	1	
	0	734 45.36 77.92 76.86	221 13.66 32.69 23.14	955 59.02
	1	208 12.86 22.08 31.37	455 28.12 67.31 68.63	663 40.98
		942 58.22	676 41.78	1618

PREDICTING MALIGNANCY

- Very high false positive and false negative rates
- Radiology very expensive for the developing world

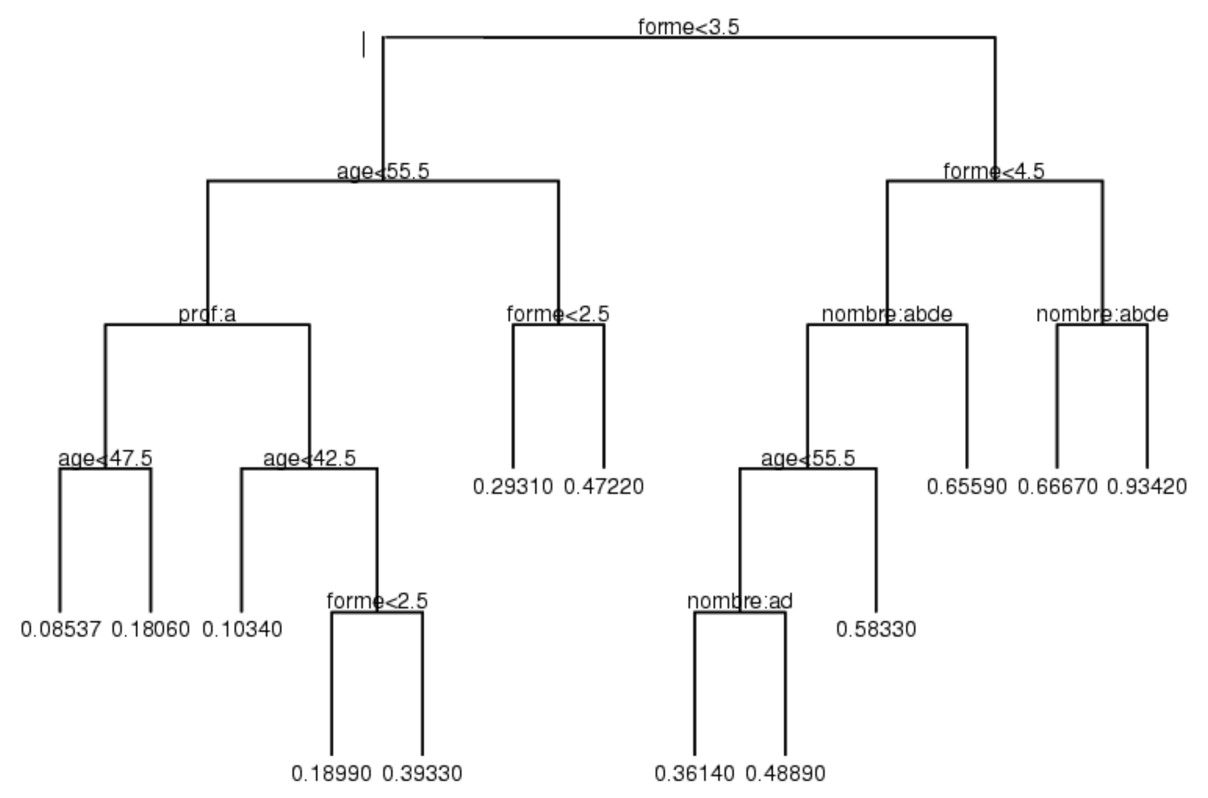
PREDICTING MALIGNANCY

- Very high false positive and false negative rates
- Radiology very expensive for the developing world
- X-rays preprocessed by computer scientists

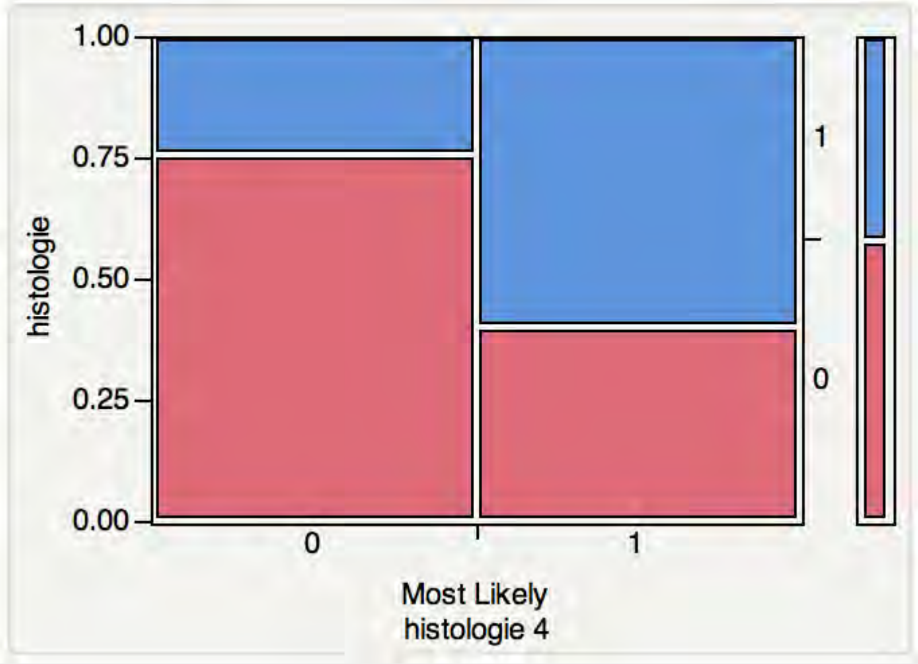
PREDICTING MALIGNANCY

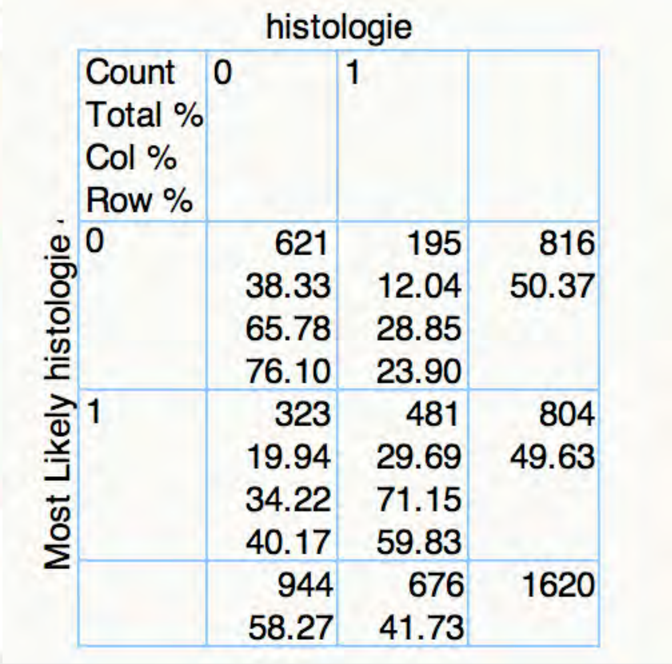
- Very high false positive and false negative rates
- Radiology very expensive for the developing world
- X-rays preprocessed by computer scientists
- 1600 X-rays with known diagnosis

TREE MODEL


forme<3.5
age<55.5
forme<4.5
prof:a
forme<2.5
nombre:abde
nombre:abde
age<47.5
age<42.5
0.29310
0.47220
age<55.5
0.65590
0.66670
0.93420
0.08537
0.18060
0.10340
forme<2.5
nombre:ad
0.58330
0.18990
0.39330
0.36140
0.48890

HOW WELL DOES A TREE DO?





histologie

Most Likely histologie

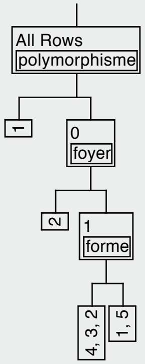
Count Total % Col % Row %	0	1	
0	621 38.33 65.78 76.10	195 12.04 28.85 23.90	816 50.37
1	323 19.94 34.22 40.17	481 29.69 71.15 59.83	804 49.63
	944 58.27	676 41.73	1620

COMBINING MODELS

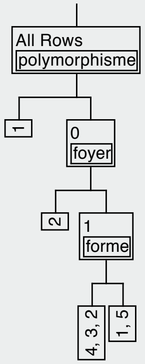
- Bagging (Bootstrap Aggregation)
  - Bootstrap a data set repeatedly
  - Take many versions of same model (e.g. tree)
    - Random Forest Variation
  - Form a committee of models
  - Take majority rule of predictions
- Boosting
  - Create repeated samples of weighted data
  - Weights based on misclassification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

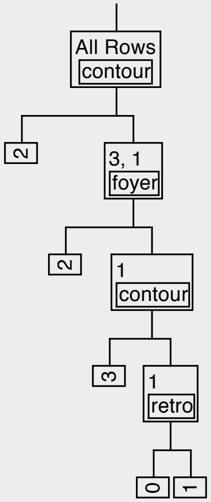
Committee of experts

LOTS OF TREES

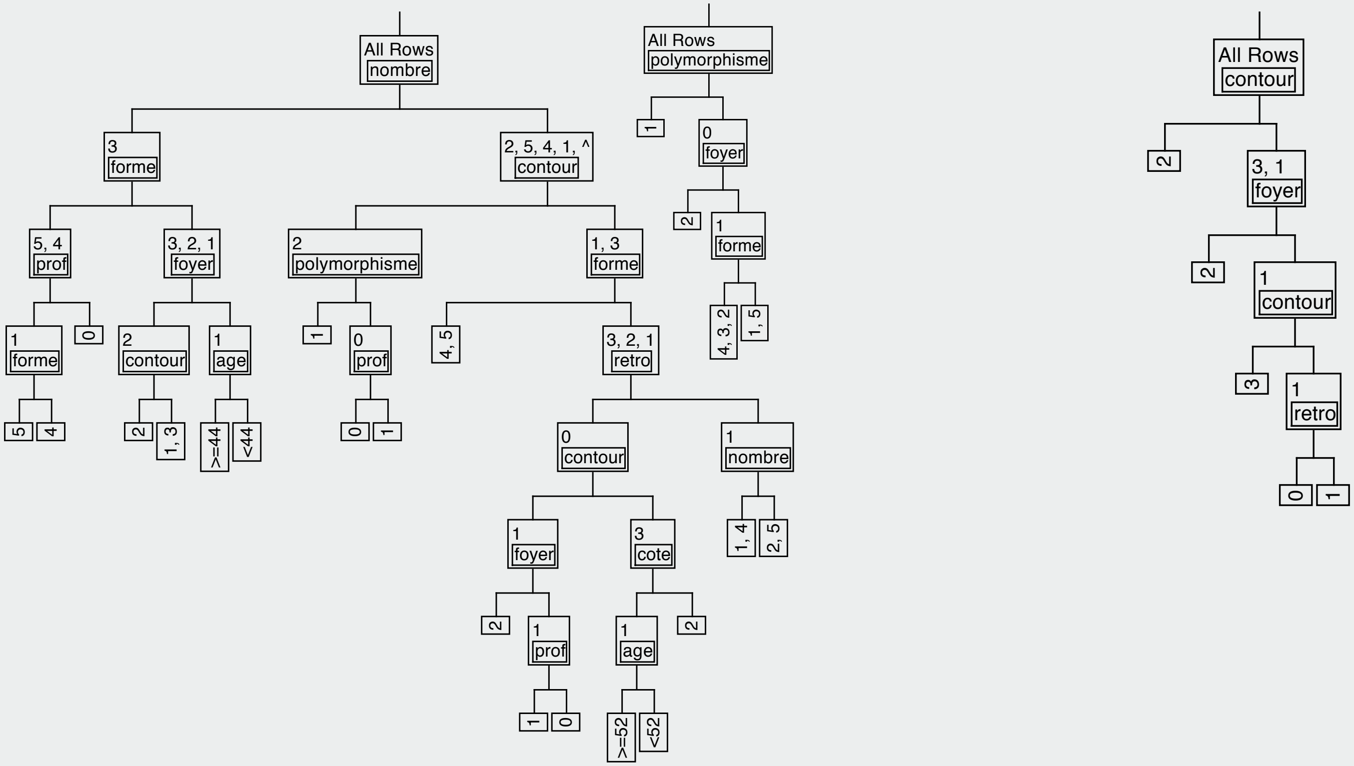


LOTS OF TREES

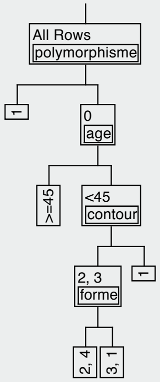


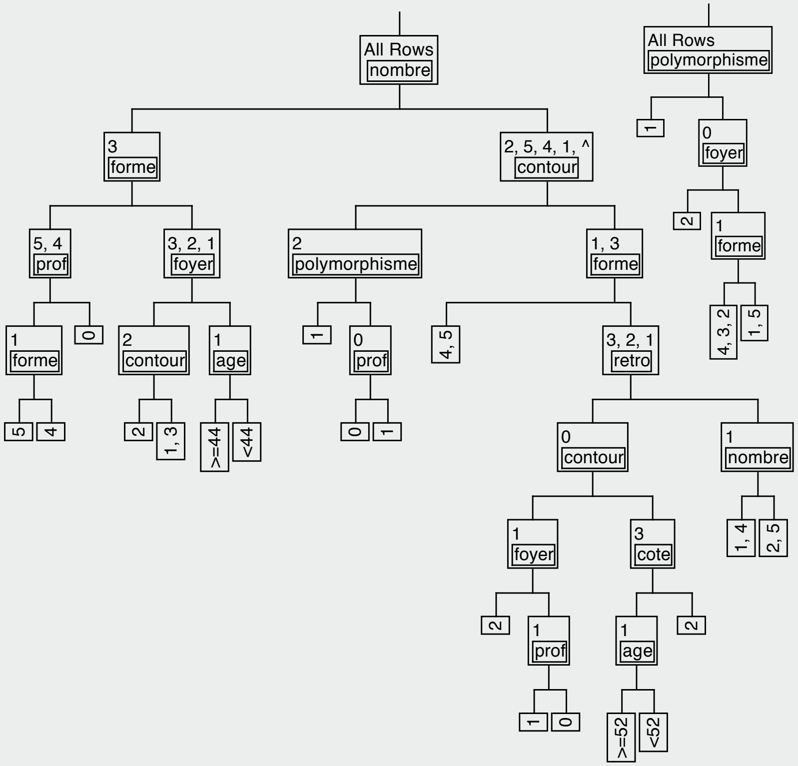


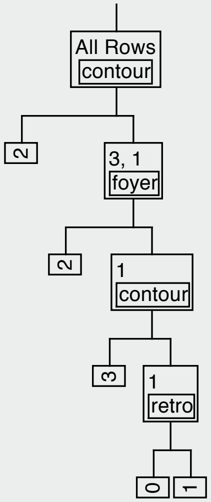
LOTS OF TREES



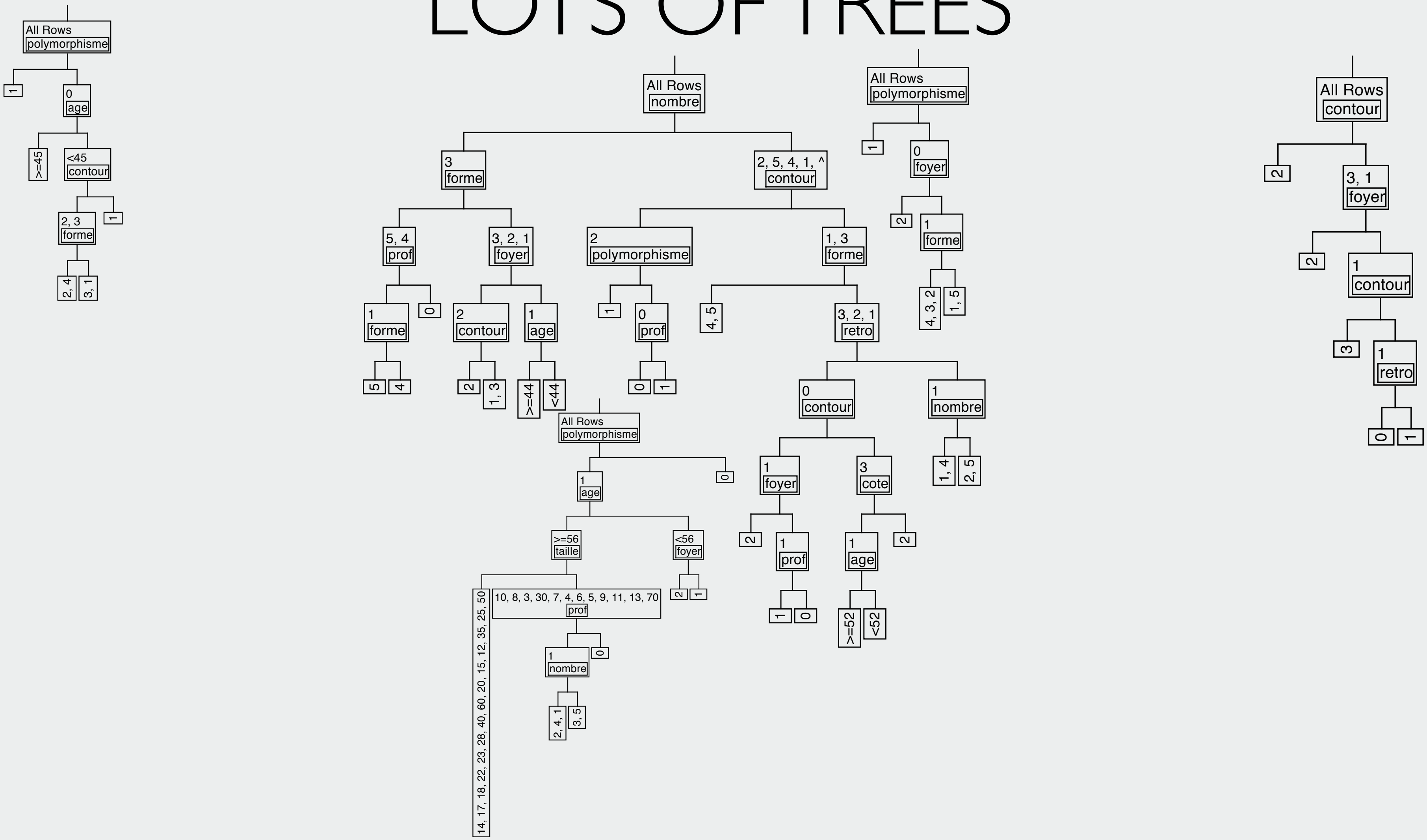
LOTS OF TREES



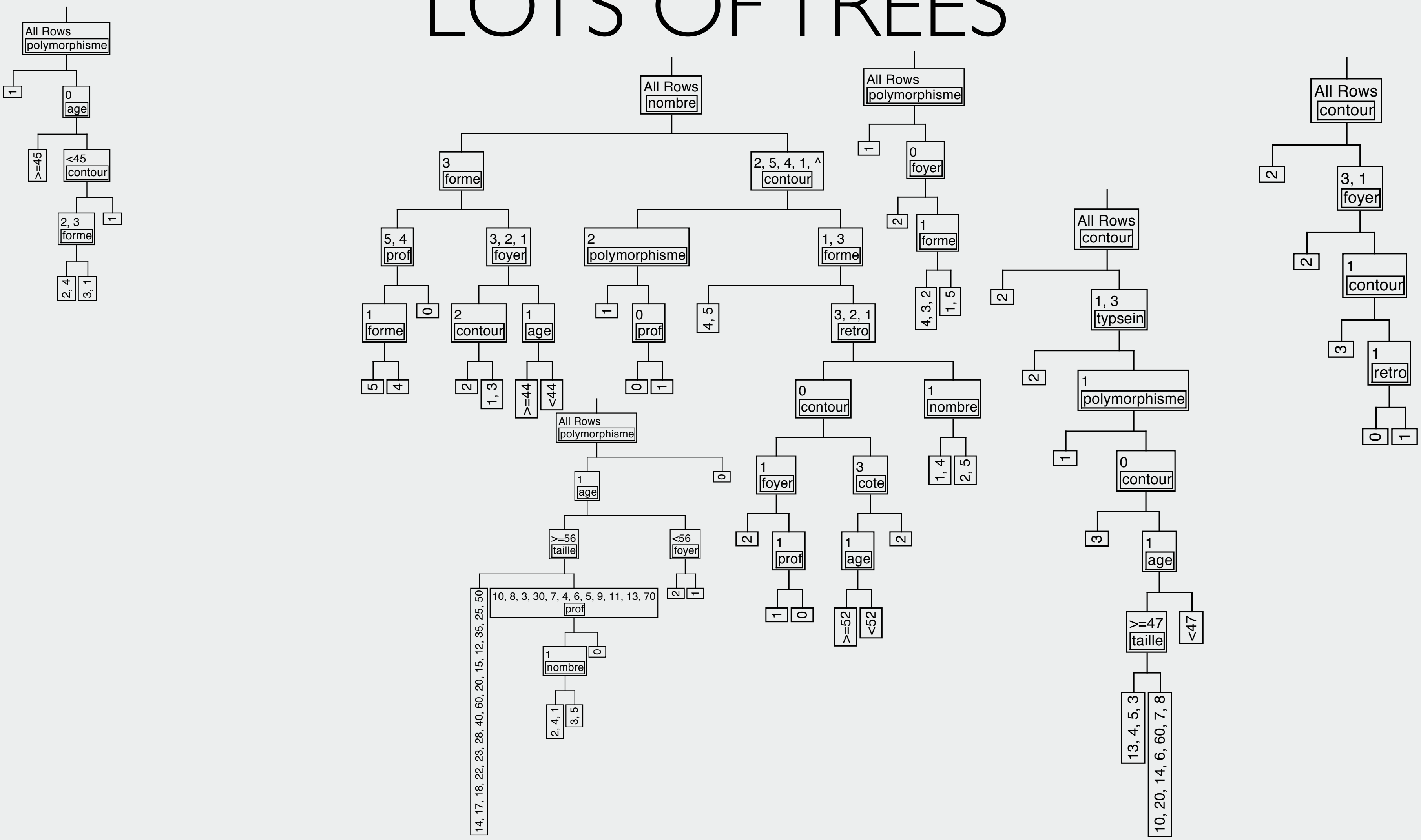




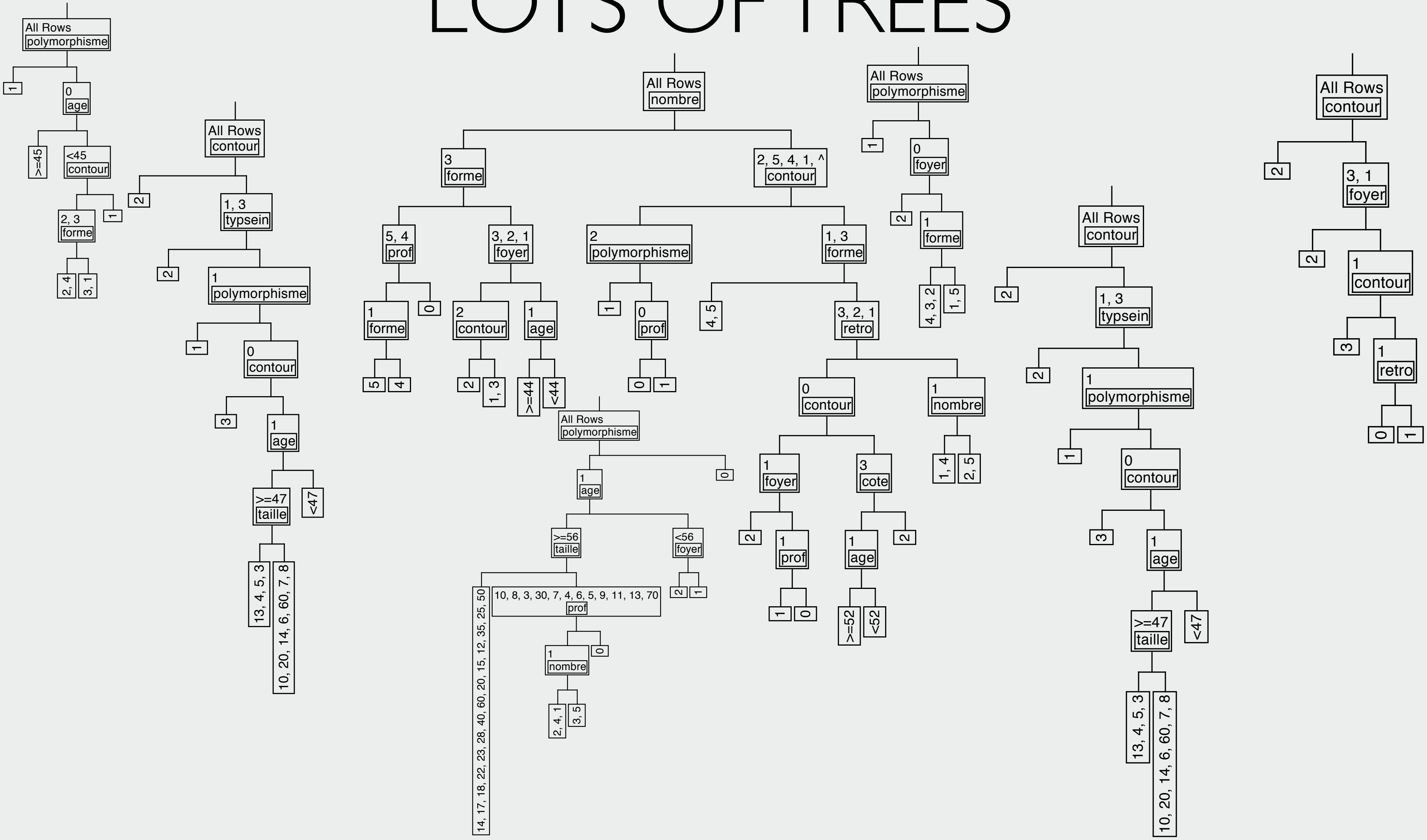
LOTS OF TREES



LOTS OF TREES



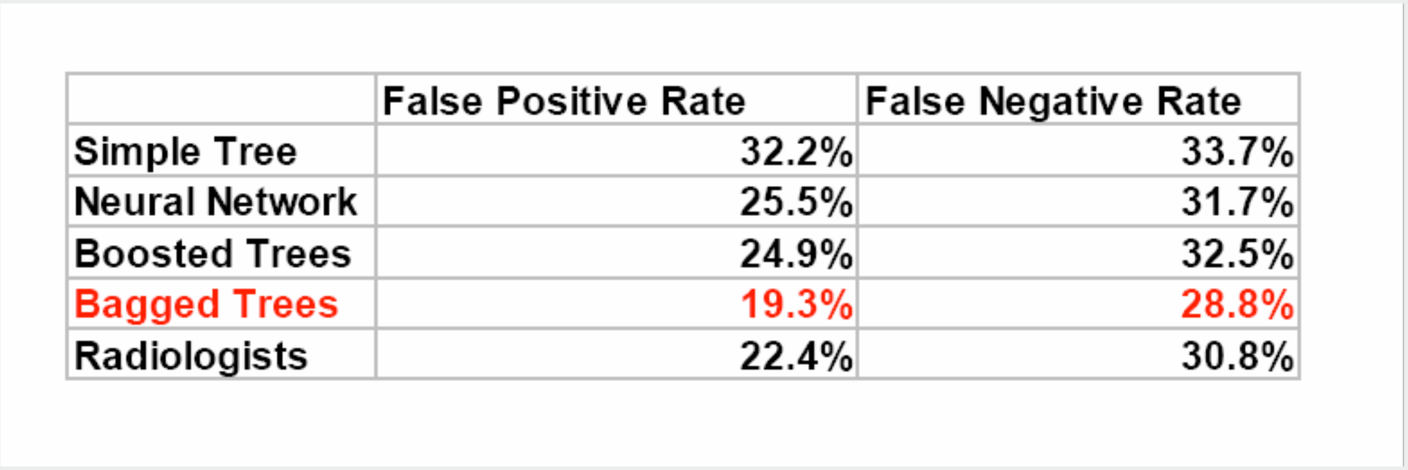
LOTS OF TREES



RANDOM FOREST

RESULTS FOR CANCER DATA SET

- Split data into train and test (62.5% - 37.5%)
- Repeat random splits 1000 times
  - ➢For each iteration, count false positives and false negatives on the 600 test set cases

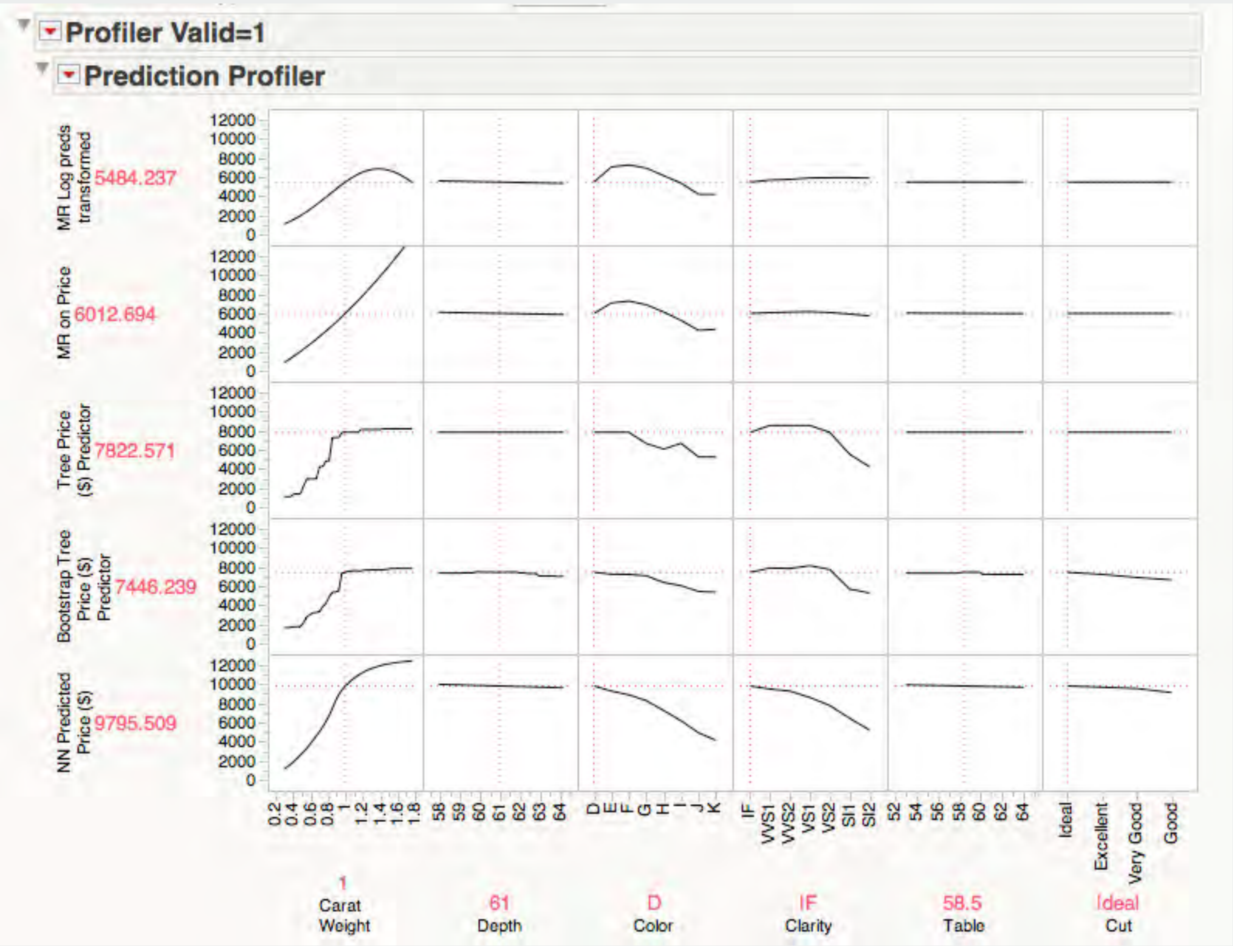


	False Positive Rate	False Negative Rate
Simple Tree	32.2%	33.7%
Neural Network	25.5%	31.7%
Boosted Trees	24.9%	32.5%
Bagged Trees	19.3%	28.8%
Radiologists	22.4%	30.8%

WHERE ARE WE?

- Model comparisons
- At this point you should have a collection of different models, all fit to the training set with results on a tuning set
- Use the tuning set to optimize model choices (thresholds for binary classification), size of model etc.

MODEL COMPARISON DIAMONDS



MODEL COMPARISON

Model Comparison Test=1

Predictors

Measures of Fit for y

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
Neural Net Boosted	Neural		0.8730	1.3925	1.0899	150
Stepwise Linear with Interactions	Fit Least Squares		0.8748	1.3826	1.0981	150
Elastic Net GR	Fit Generalized Adaptive Elastic Net		0.8759	1.3763	1.0812	150
Random Forest	Bootstrap Forest		0.7459	1.9697	1.5389	150
Boosted Tree	Boosted Tree		0.8464	1.5312	1.2063	150

Profiler Test=1

Prediction Profiler

MODEL AVERAGE

At least two choices:

- Simple arithmetic average
- Neural average

MODEL AVERAGE?

Model Comparison Test=0

Predictors

Measures of Fit for y

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
y Avg Predictor By Test	Model Comparison Model Averaged		0.9203	1.2156	0.9742	750
Neural Average	Neural		0.9082	1.3050	1.0464	750

Model Comparison Test=1

Predictors

Measures of Fit for y

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
y Avg Predictor By Test	Model Comparison Model Averaged		0.8718	1.3989	1.0986	150
Neural Average	Neural		0.8787	1.3605	1.0746	150

Profiler Test=1

Prediction Profiler

Model Comparison Test=2

Predictors

Measures of Fit for y

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
y Avg Predictor By Test	Model Comparison Model Averaged		0.8739	1.4836	1.1720	100
Neural Average	Neural		0.8804	1.4448	1.1626	100

HOW DO WE START?

- Life is not always kind
  - Categorical variables — require pre processing
  - Missing data - what to do? Informative missing.

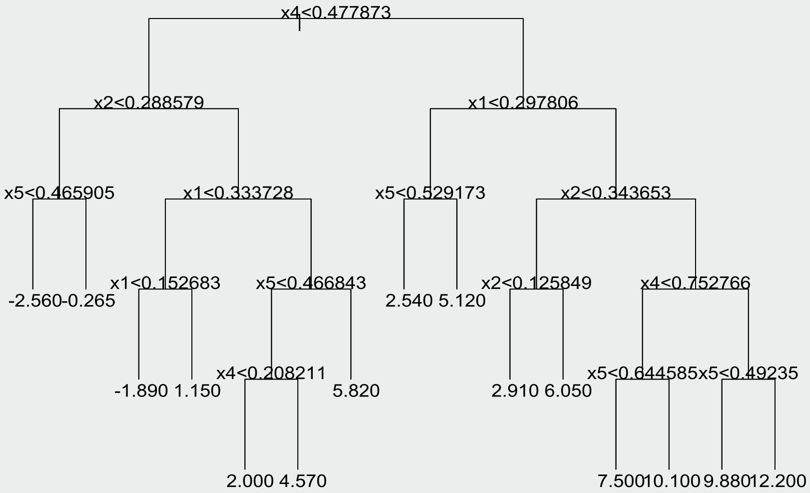
- PVA has 479 variables – where to start?

EXPLORATORY DATA MODELS

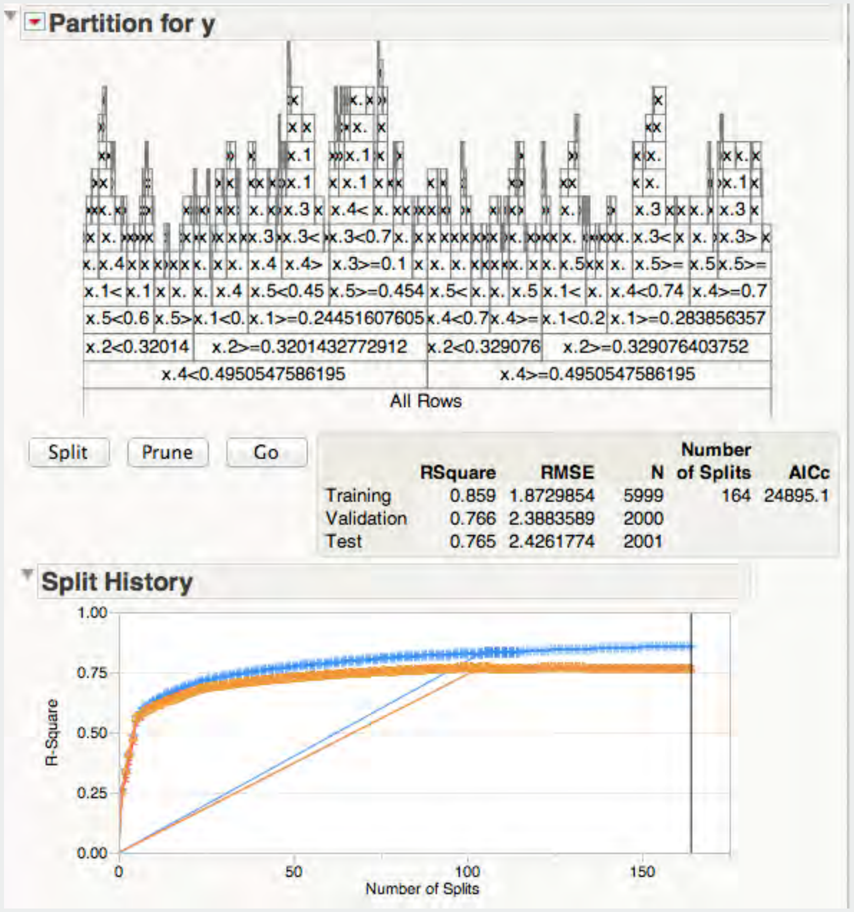
- Use a tree to find a smaller subset of variables to investigate
- Explore this set graphically
  - Start the modeling process over
- Build model
  - Compare model on small subset with full predictive model

START WITH A SIMPLE MODEL

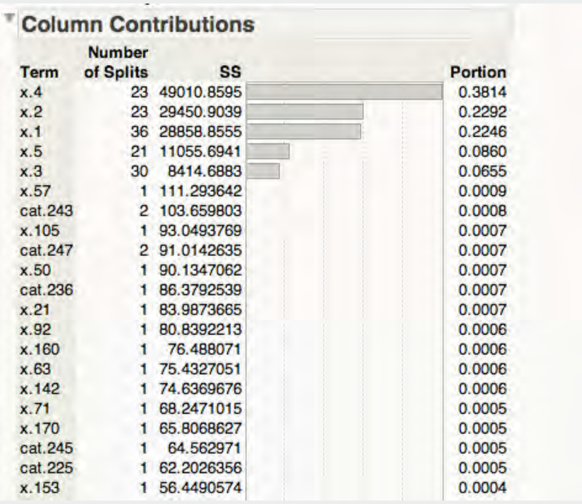
- Tree?



EXPLORATORY TREE

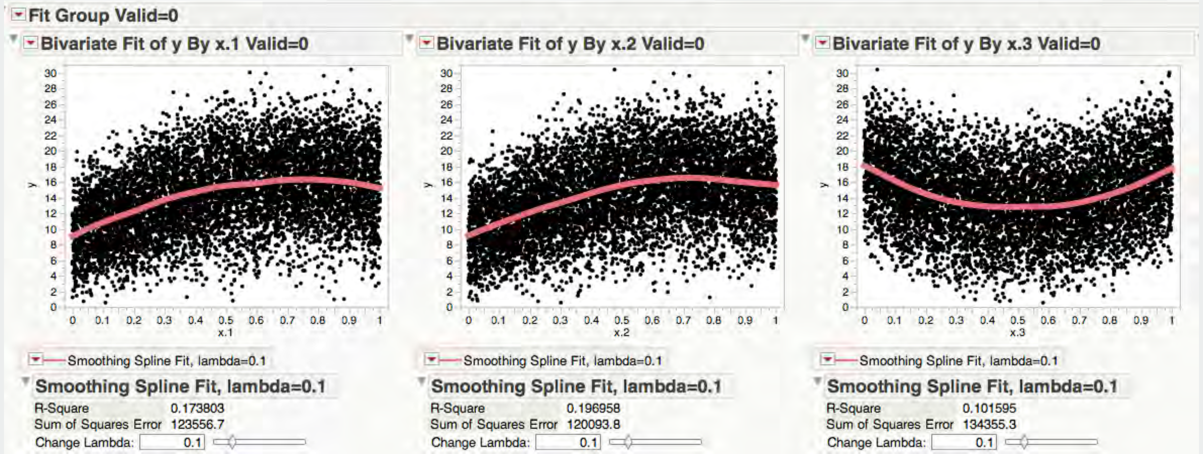


	RSquare	RMSE	N	Number of Splits	AICc
Training	0.859	1.8729854	5999	164	24895.1
Validation	0.766	2.3883589	2000		
Test	0.765	2.4261774	2001		


Column Contributions

Term	Number of Splits	SS	Portion
x.4	23	49010.8595	0.3814
x.2	23	29450.9039	0.2292
x.1	36	28858.8555	0.2246
x.5	21	11055.6941	0.0860
x.3	30	8414.6883	0.0655
x.57	1	111.293642	0.0009
cat.243	2	103.659803	0.0008
x.105	1	93.0493769	0.0007
cat.247	2	91.0142635	0.0007
x.50	1	90.1347062	0.0007
cat.236	1	86.3792539	0.0007
x.21	1	83.9873665	0.0007
x.92	1	80.8392213	0.0006
x.160	1	76.488071	0.0006
x.63	1	75.4327051	0.0006
x.142	1	74.6369676	0.0006
x.71	1	68.2471015	0.0005
x.170	1	65.8068627	0.0005
cat.245	1	64.562971	0.0005
cat.225	1	62.2026356	0.0005
x.153	1	56.4490574	0.0004

NOW BACK TO THE START



Fit Group Valid=0

Bivariate Fit of y By x.1 Valid=0

Smoothing Spline Fit, lambda=0.1

R-Square 0.173803
Sum of Squares Error 123556.7
Change Lambda: 0.1

Bivariate Fit of y By x.2 Valid=0

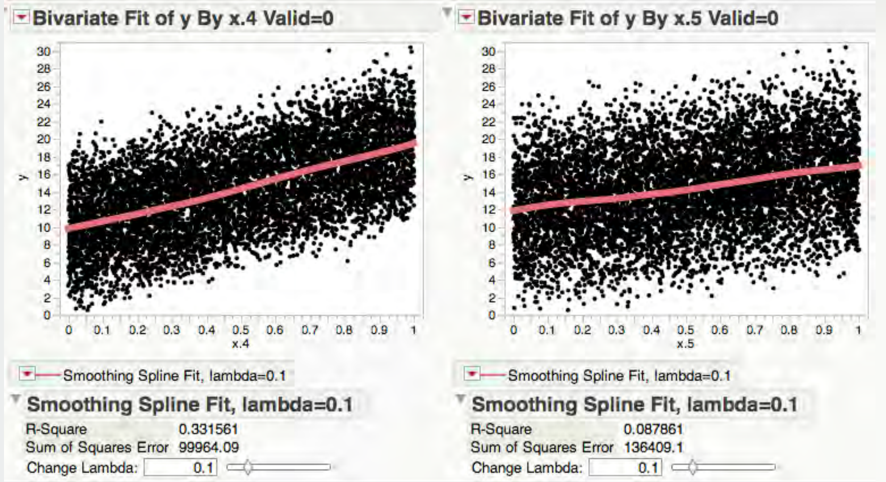
Smoothing Spline Fit, lambda=0.1

R-Square 0.196958
Sum of Squares Error 120093.8
Change Lambda: 0.1

Bivariate Fit of y By x.3 Valid=0

Smoothing Spline Fit, lambda=0.1

R-Square 0.101595
Sum of Squares Error 134355.3
Change Lambda: 0.1



Bivariate Fit of y By x.4 Valid=0

Smoothing Spline Fit, lambda=0.1

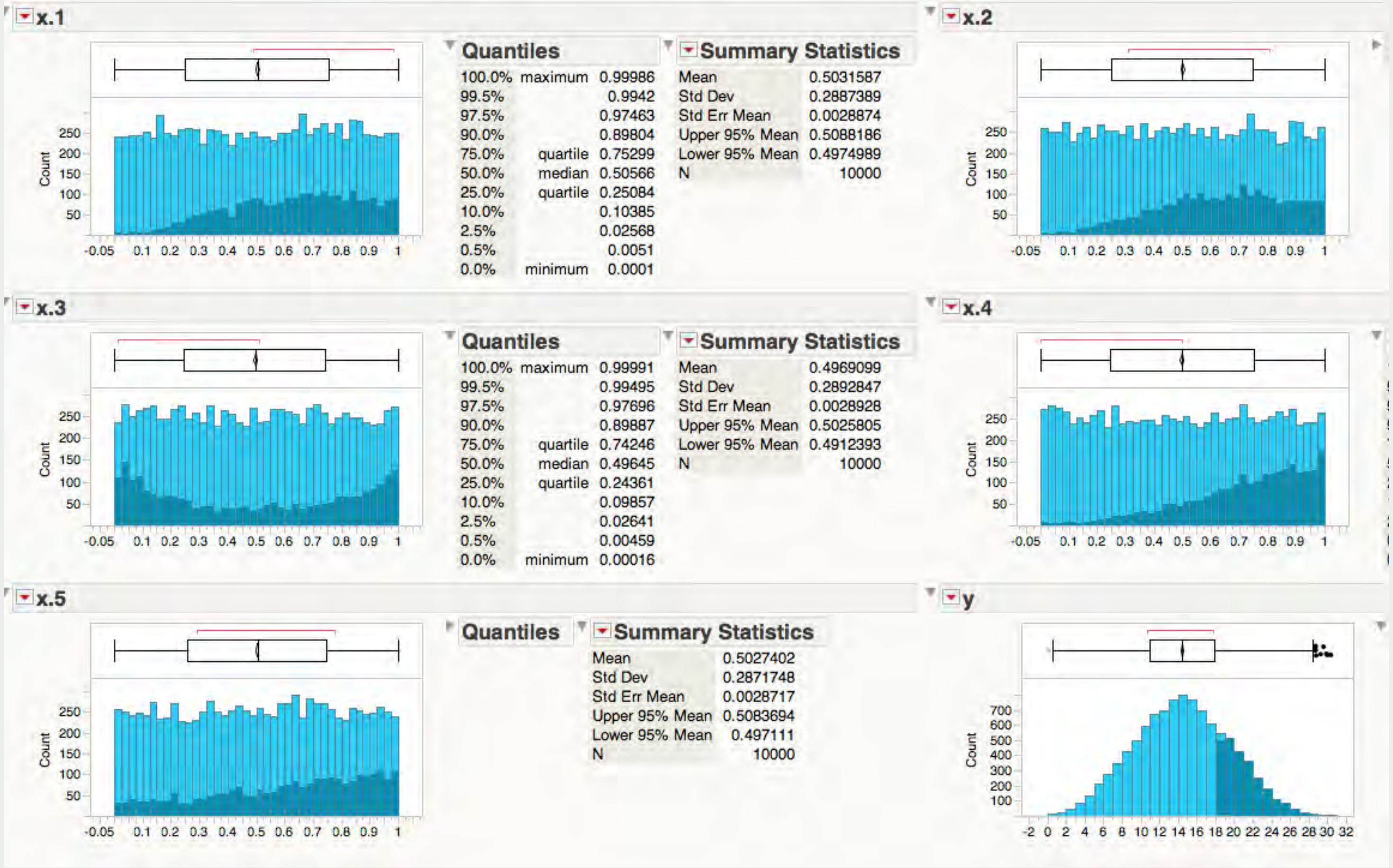
R-Square 0.331561
Sum of Squares Error 99964.09
Change Lambda: 0.1

Bivariate Fit of y By x.5 Valid=0

Smoothing Spline Fit, lambda=0.1

R-Square 0.087861
Sum of Squares Error 136409.1
Change Lambda: 0.1

BRUSHING



x.1

Quantiles

100.0%	maximum	0.99986
99.5%		0.9942
97.5%		0.97463
90.0%		0.89804
75.0%	quartile	0.75299
50.0%	median	0.50566
25.0%	quartile	0.25084
10.0%		0.10385
2.5%		0.02568
0.5%		0.0051
0.0%	minimum	0.0001

Summary Statistics

Mean	0.5031587
Std Dev	0.2887389
Std Err Mean	0.0028874
Upper 95% Mean	0.5088186
Lower 95% Mean	0.4974989
N	10000

x.3

Quantiles

100.0%	maximum	0.99991
99.5%		0.99495
97.5%		0.97696
90.0%		0.89887
75.0%	quartile	0.74246
50.0%	median	0.49645
25.0%	quartile	0.24361
10.0%		0.09857
2.5%		0.02641
0.5%		0.00459
0.0%	minimum	0.00016

Summary Statistics

Mean	0.4969099
Std Dev	0.2892847
Std Err Mean	0.0028928
Upper 95% Mean	0.5025805
Lower 95% Mean	0.4912393
N	10000

x.5

Quantiles

Summary Statistics

Mean	0.5027402
Std Dev	0.2871748
Std Err Mean	0.0028717
Upper 95% Mean	0.5083694
Lower 95% Mean	0.497111
N	10000

CLUSTERING

- Turn the problem around
- Instead of predicting something about a variable, use the variables to group the observations
  - K-means
  - Hierarchical clustering

HIERARCHICAL CLUSTERING

- Define a distance (or similarity) and put in an N*N matrix

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances between the clusters be the same as the distances between the items they contain. How – later
2. Find the closest pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

SINGLE-LINKAGE

- Step 3 can be done in different ways, which is what distinguishes *single-linkage* from *complete-linkage* and *average-linkage* clustering.
  - In single-linkage clustering the distance between one cluster and another cluster is the shortest distance from any member of one cluster to any member of the other cluster.
  - In *complete-linkage* clustering (also called the *diameter* or *maximum* method), we consider the distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster.
  - In *average-linkage* clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.

DISTANCE MATRIX

	Bari	**Firenze**	**Milano**	**Napoli**	**Roma**	**Torino**
Bari	0	662	877	255	412	996
Firenze	662	0	295	468	268	400
Milano	877	295	0	754	564	138
Napoli	255	468	754	0	219	869
Roma	412	268	564	219	0	669
Torino	996	400	138	869	669	0

MERGE MILANO WITH TORINO

RECOMPUTE DISTANCES

	Bari	Firenze	Milano/Torino	Napoli	Roma
Bari	0	662	877	255	412
Firenze	662	0	295	468	268
Milano/Torino	877	295	0	754	564
Napoli	255	468	754	0	219
Roma	412	268	564	219	0

Merge Napoli with Rome

3RD STEP

	Bari	Firenze	Milano/Torino	Napoli/Roma
Bari	0	662	877	255
Firenze	662	0	295	268
Milano/Torino	877	295	0	564
Napoli/Roma	255	268	564	0

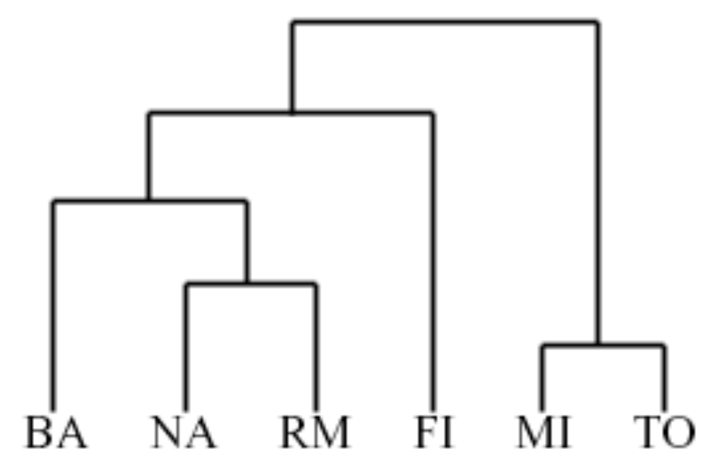
NEXT

	Bari/Napoli/Roma	Firenze	Milano/Torino
Bari/Napoli/Roma	**0**	**268**	**564**
Firenze	**268**	**0**	**295**
Milano/Torino	**564**	**295**	**0**

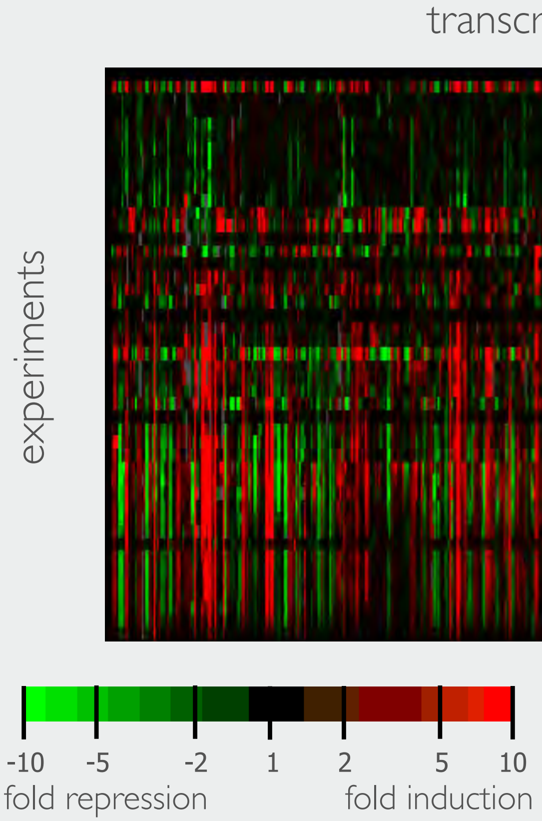
LAST STEP

	Bari/Napoli/Roma/Firenze	Milano/Torino
Bari/Napoli/Roma/ Firenze	0	295
Milano/Torino	295	0

REPRESENTATION AS A TREE

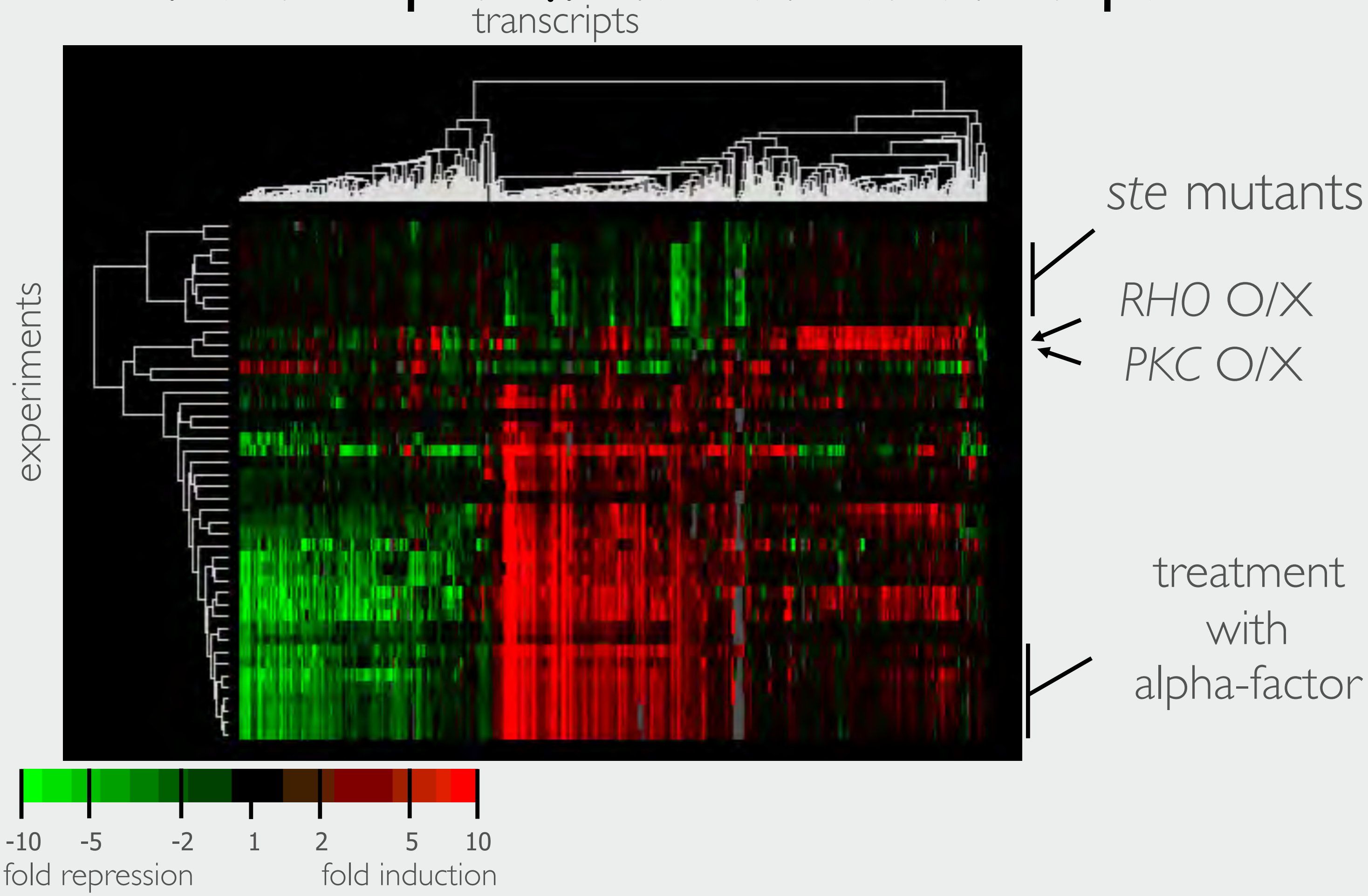


Step 1: cut experiments and transcripts falling below P-value and ratio thresholds

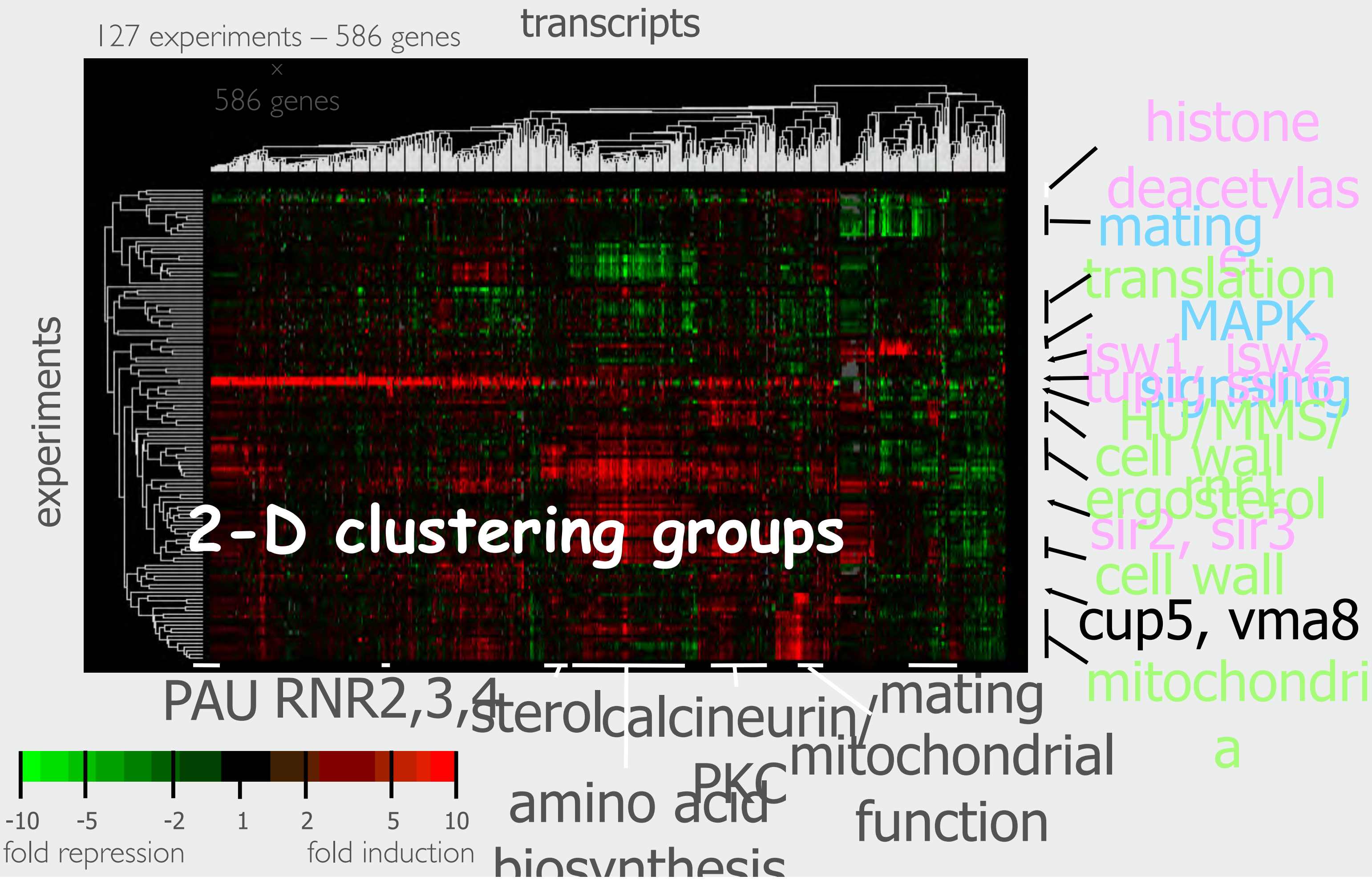


44 experiments
×
407 genes

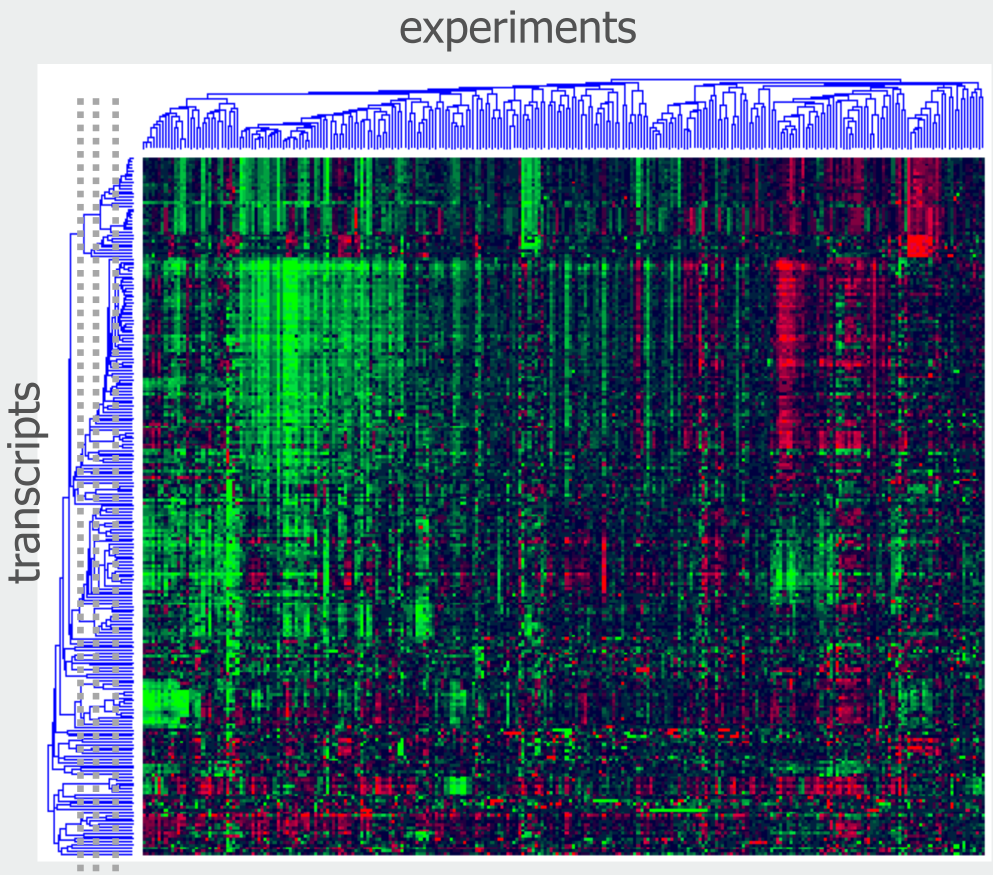
Step 2:
cluster experiments and transcripts



2-D clustering groups



Automatic “labeling” of cluster



amino acid biosynthesis
($p<10^{-14}$)
amino acid metabolism
($p<10^{-14}$)

methionine metabolism
($p=1.07\times10^{-7}$)

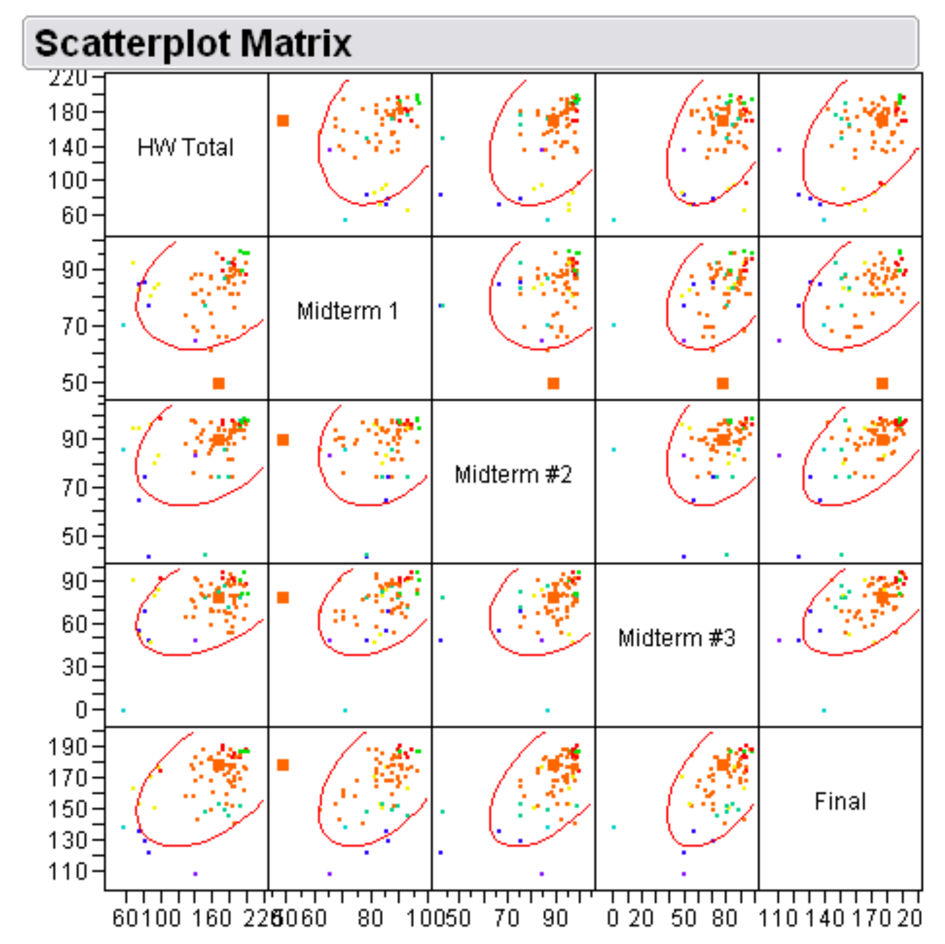
EXAMPLE

- Final Grades
  - Homework
  - 3 Midterms
  - Final
- Principal Components
  - First is weighted average
  - Second is difference between 1 and 3rd midterms and 2nd

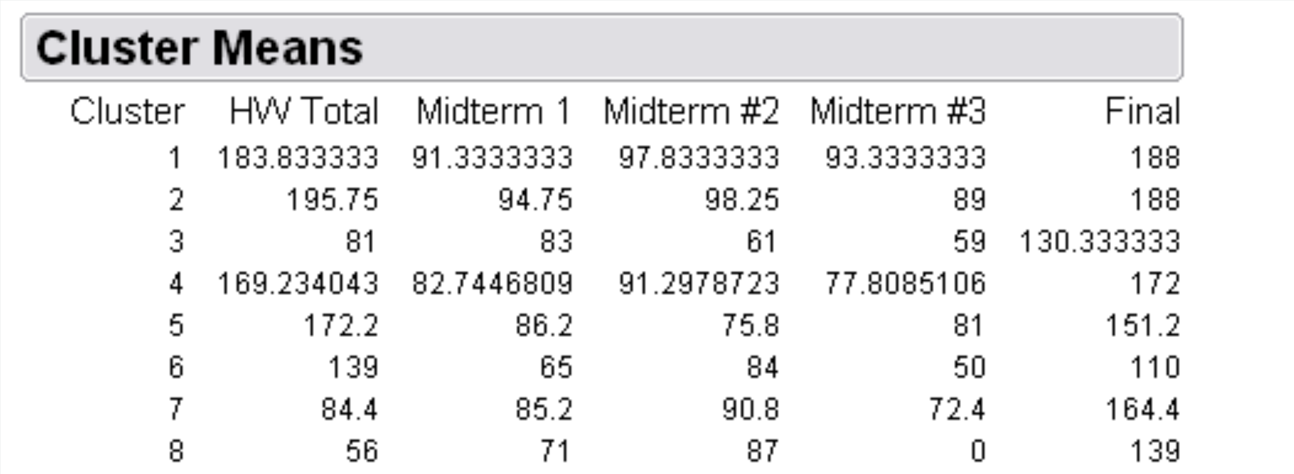
K-MEANS

- Rather than find the K nearest neighbors, find K clusters
- Problem is now to group observations into clusters rather than predict
- Not a predictive model, but a segmentation model

SCATTERPLOT MATRIX



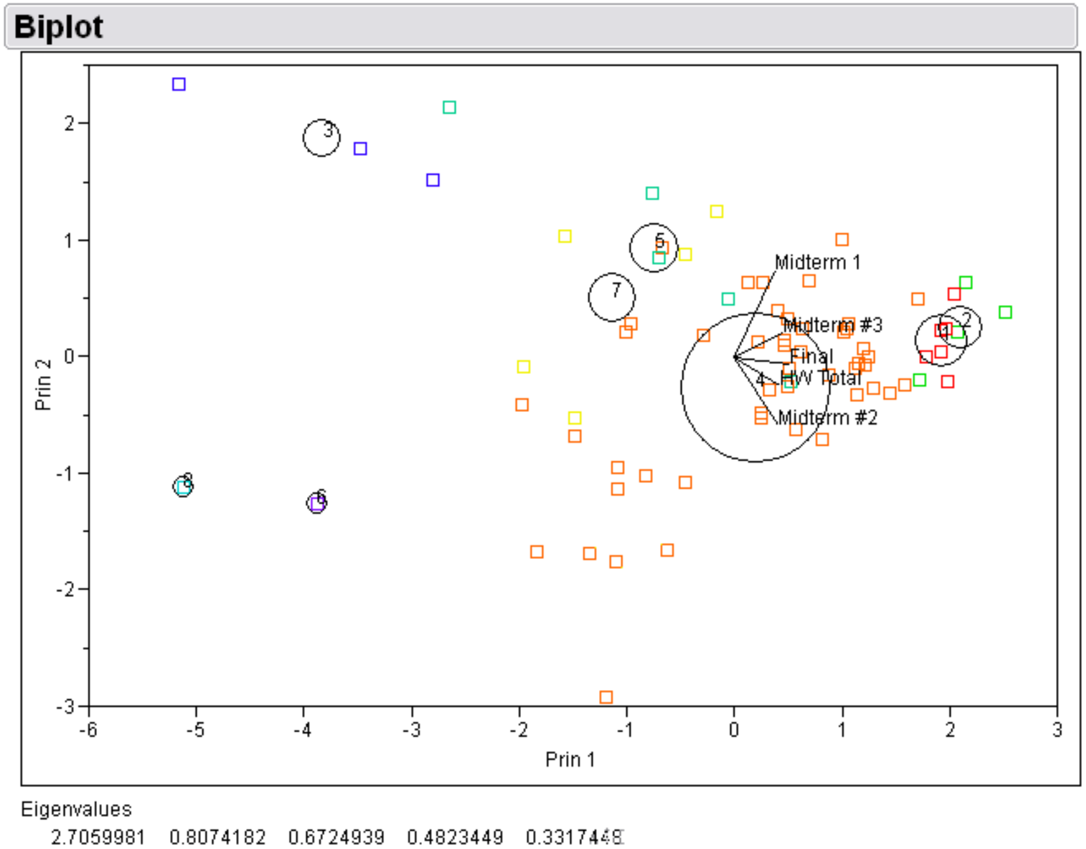
CLUSTER MEANS



Cluster Means

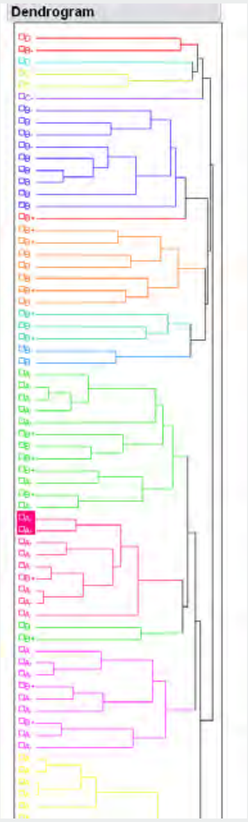
Cluster	HW Total	Midterm 1	Midterm #2	Midterm #3	Final
1	183.833333	91.3333333	97.8333333	93.3333333	188
2	195.75	94.75	98.25	89	188
3	81	83	61	59	130.333333
4	169.234043	82.7446809	91.2978723	77.8085106	172
5	172.2	86.2	75.8	81	151.2
6	139	65	84	50	110
7	84.4	85.2	90.8	72.4	164.4
8	56	71	87	0	139

BIPLOT



```
Eigenvalues
    2.7059981   0.8074182   0.6724939   0.4823449   0.3317448
```

GRADE EXAMPLE



KNOW WHEN TO FOLD 'EM

- 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)

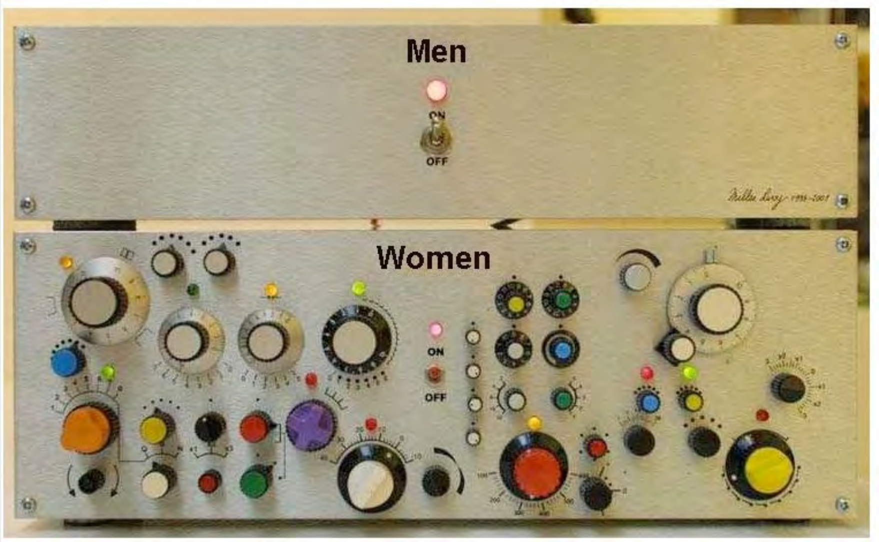
FAST FAIL

- Not every modeling effort is a success
  - A model search can save lots of queries
- Data took 8 months to get ready

- Analyst spent 2 months exploring it
- Tree models, stepwise regression (and a neural network running for several hours) found no out of sample predictive ability

MACHINES ARE SMART

- Why do statisticians like interpretability?
- Black boxes are not interpretable, but there may be important information



BUT YOU ARE SMARTER

- A new backpack inkjet printer is showing higher than expected warranty claims
  - What are the important variables?
  - What's going on?

- A neural networks shows that Zip code is the most important predictor

ZIP CODE?

DATA MINING – DOE

- Data Mining is exploratory
- Efforts can go on simultaneously
- Learning cycle oscillates naturally between the two

WHAT DID WE LEARN?

- Toy problem
  - Functional form of model
- PVA data
  - Useful predictor – increased sales 40%
- Ingots
  - Gave clues as to where to look
  - Experimental design followed
- Churches
  - When to quit
- Printers
  - When to experiment – what factors

CHALLENGES FOR DATA

- Not algorithms
- Overfitting
- Finding an interpretable model that fits reasonably well

RECAP – SUCCESS IN DATA

- Problem formulation
- Data preparation
  - Data definitions
  - Data cleaning
  - Feature creation, transformations
- EDM – exploratory modeling
  - Reduce dimensions

SUCCESS IN DATA MINING II

- Don’t forget Graphics
- Second phase modeling
- Testing, validation, implementation
- Constant re-evaluation of models

WHICH METHOD(S) TO USE?

- No method is best
- Which methods work best when?
- Which method to use?

YES!

FOR MORE INFORMATION

- Two Crows
  - http//www.twocrows.com
- KDNuggets
  - http://www.kdnuggets.com
- deveaux@williams.edu

M. Berry and G. Linoff, Data Mining Techniques, Wiley, 1997

Dorian Pyle, Data Preparation for Data Mining, Morgan Kaufmann, 1999

Hand, D.J., Mannila, H. and Smyth, P., Principles of Data Mining, MIT Press 2001

Tan, P.N, Steinbach, and Kumar: Introduction to Data Mining, Addison-Wesley, 2006

Nisbet, Elder and Miner, Handbook of Statistical Analysis and Data Mining Applications, Academic Press 2009

Bishop, Christopher Pattern Recognition and Machine Learning, Springer

Hastie, Tibshirani and Friedman, Statistical Learning 2nd edition, Springer