Data Mining: An Introduction

February 25, 2014

Susan Watson, FSA
Paul Anderson, FCAS
Rahul Parsa, PhD

Iowa Actuaries Club

Recent Headlines

How Big Data Created a Cruel Result

Amazon says it can send you items before you've ordered

Report: Spies use Angry Birds, apps to track people

You don't want your privacy: Disney and the meat space data race

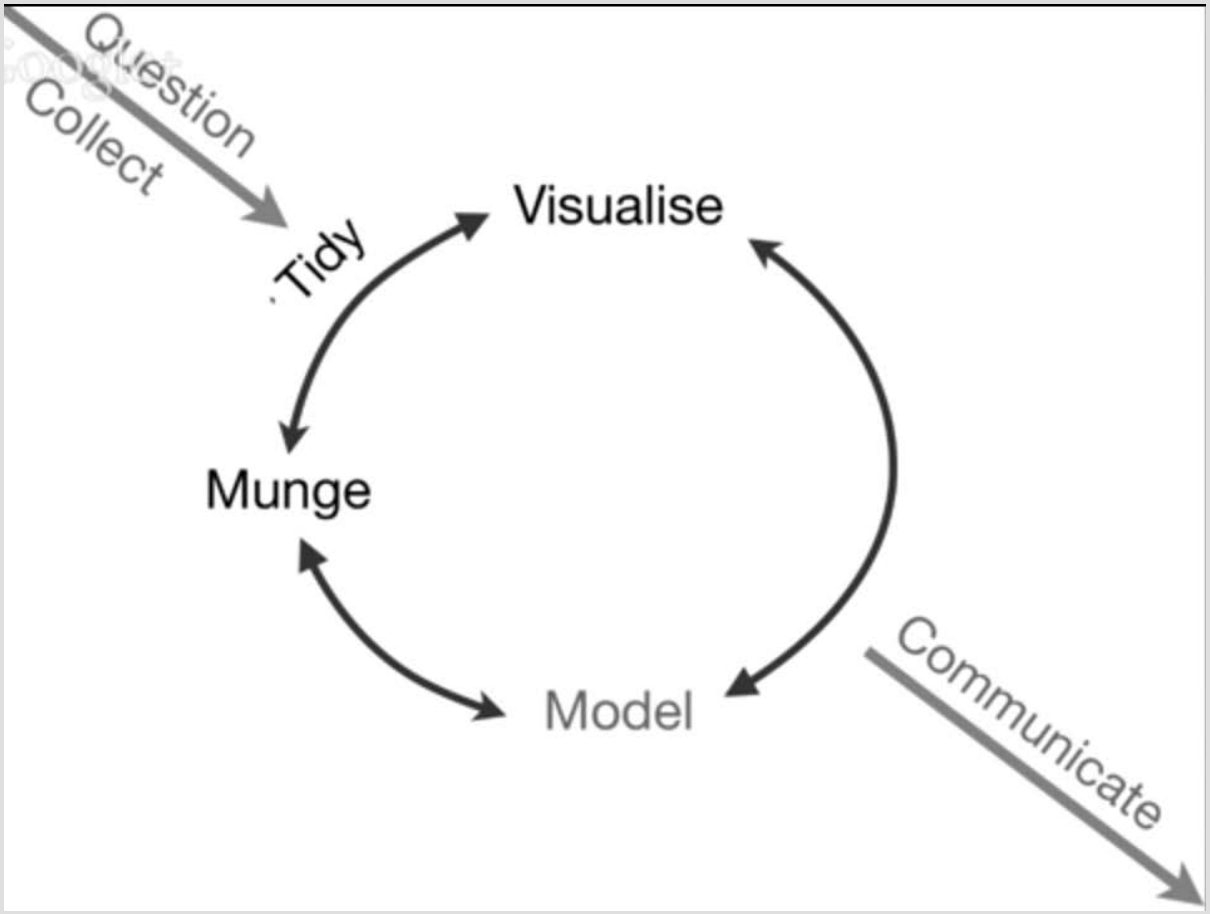
Data Challenges

- Volume
- Velocity
- Variety

Data Mining Process

- Collect Data
- Clean Data
- Exploratory Data Analysis
- Model Building
- Model Assessment
- Model Implementation

Process: Another View



Source: Hadley Wickham

Collect Data

- Internal / External
- Structured / Unstructured
- Observed / Simulated

Clean Data

- Format Data
- Merge Data
- Missing Values
- Data Problems

Exploratory Data Analysis

- Visual Exploration
- Transformations
- Variable Creation
- Variable Interactions
- Variable Reduction

Model Building

Supervised Methods

- Regression
- Classification And Regression Trees (CART)
- Neural Networks
- Support Vector Machines (SVM)

Unsupervised Methods

- Hierarchical Clustering
- K-Means
- Self-Organizing Maps

Assumptions

- Assumptions are the termites of relationships.

  Henry Winkler

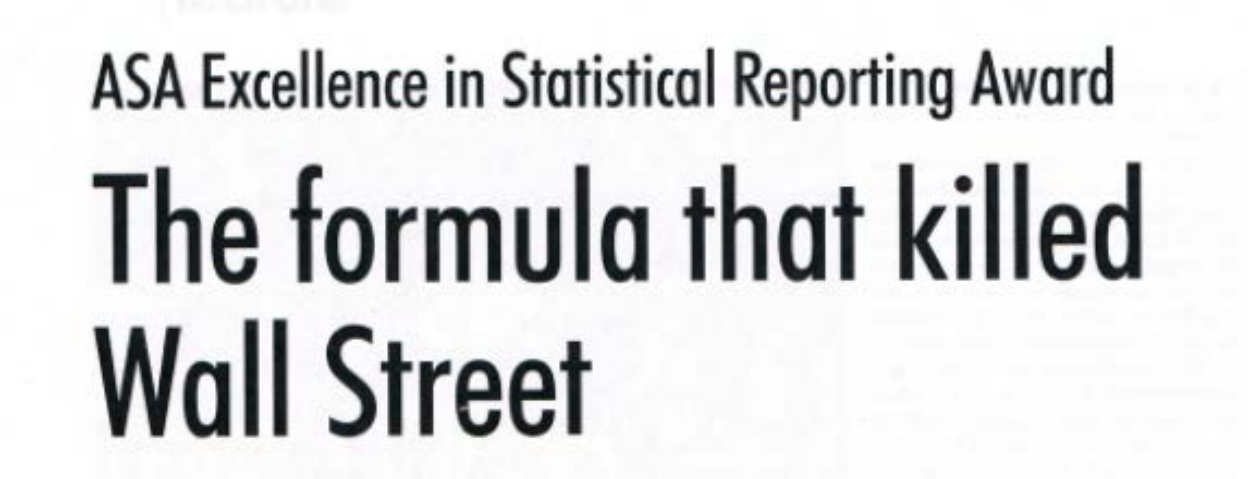
- Begin challenging your assumptions. Your assumptions are the windows on the world. Scrub them off every once in a while or the light won't come in.

  Alan Alda

- Assumptions are the foundation for modeling.

  Susan Watson, Rahul Parsa, Paul Anderson

Assumptions (cont'd)



Assumptions (cont'd) Univariate Analysis

- **OLS (Ordinary Least Squares)**
  - Model: $E(Y) = X\beta$
  - Assumption: Constant Variance

- **GLM (General Linear Models)**
  - Model: $E(Y) = \mu$, $\eta = X\beta$ and $\mu = \eta$
  - Assumption: Y is normally distributed

Assumptions (cont'd)

- **GenLM (Generalized Linear Models)**
  - Model: $E(Y) = \mu$, $\eta = X\beta$ and $g(\mu) = \eta$, where g is any monotonic differentiable function
  - Assumption: Y is from an exponential family
- **Interpretation:**
  - While the assumptions are important, equally important is the interpretation of the results.

Model Assessment

- Differs Based on Approach
- Training / Test Data
- Reduces Likelihood of Overfitting the Model

Example 1:
Australian Open Tennis Tournament

Goal: Predict the winner of the men's and women's singles events

Data Provided:

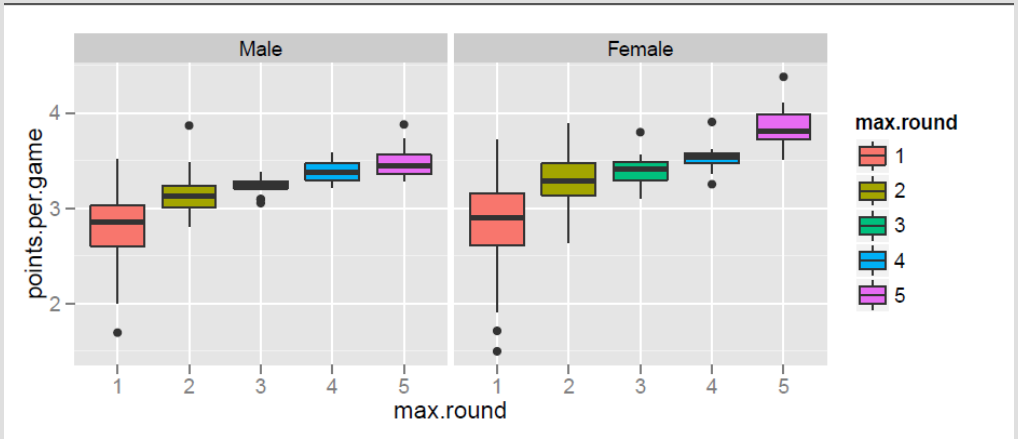
- Match statistics for each player for the first 3 rounds
- 128 men; 128 women
- 27 variables (aces, double faults, % 1st serves in, …)

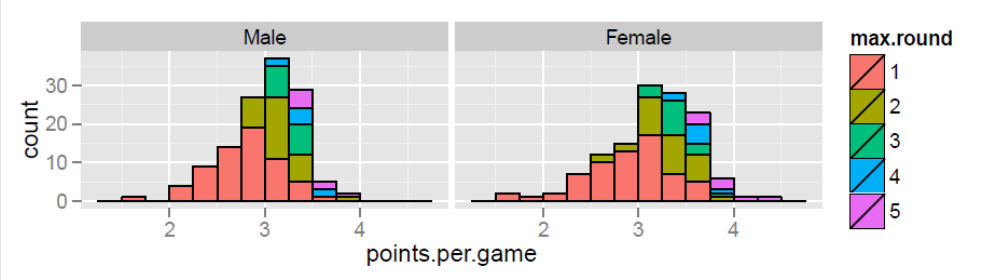
Data Issues:

- Missing values
- Response variable discrete and skewed
- Predictor variables skewed

Response Variable

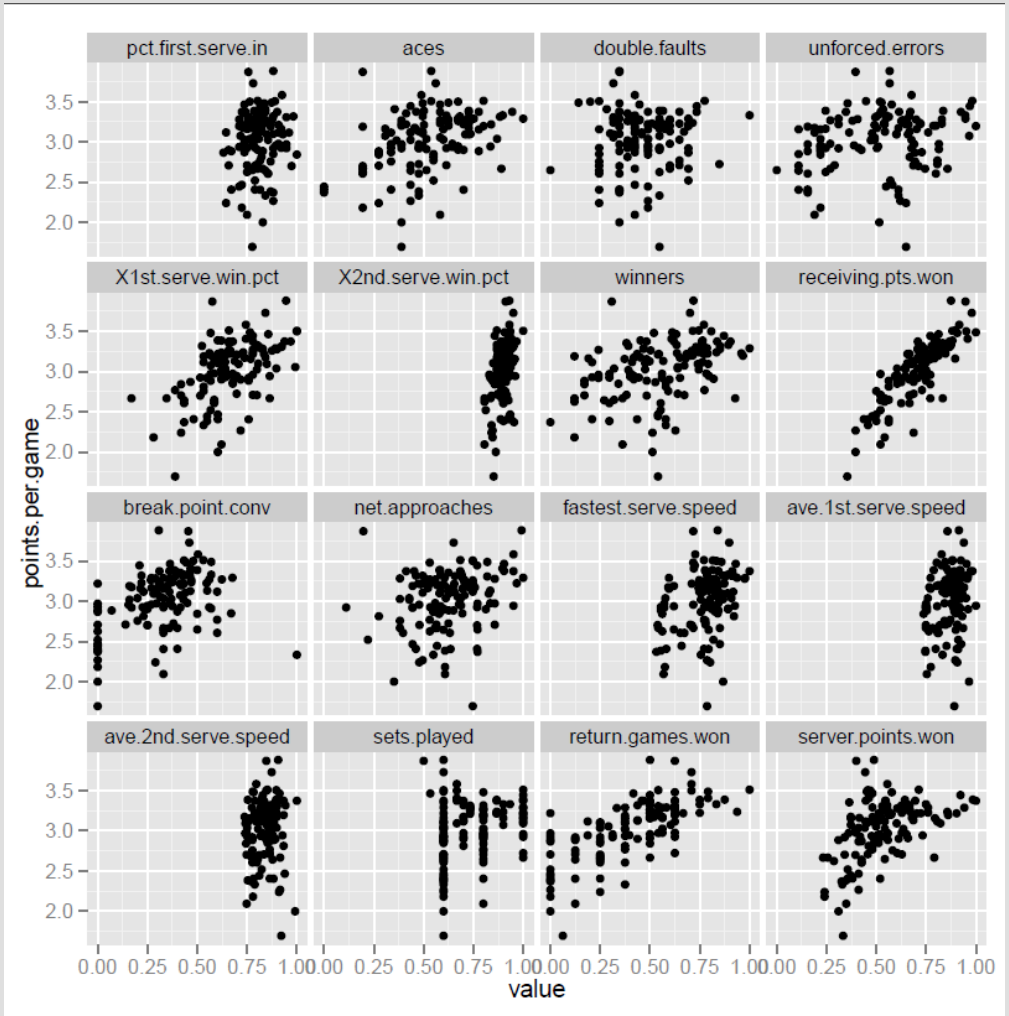
Created new variable: points.per.game





Iowa Actuaries Club

Predictor Variables



Iowa Actuaries Club

Final Model & Results

Multiple Linear Regression Model

- Men: points.per.game = receiving.pts.won + 1st.serve.win.pct + server.points.won + sets.played + ave.2nd.serve speed + net.approaches + winners
- Women: points.per.game = receiving.pts.won + 1st.serve.win.pct + server.points.won + return.games.won + break.point.conv + 2nd.serve.win.pct + sets.played + unforced.errors + pct.first.serve.in + double faults

Prediction

- Men: Andy Murray ☹
- Women: Victoria Azarenka ☺

Insurance Company Applications

- Pricing
- Underwriting
- Claims
- Marketing

Example 2:
Marketing Analysis for Niche Insurer

Background: 2011 results were 20 percentage points worse than 2010 and among the worst in company history

Goals:

- Identify 2 target markets to aggressively (& profitably) grow in
- Implement clear & objective Underwriting & Sales strategies

Data Provided:

- 5 years of policy data = 67,000 policy records; 6,000 claims
- Low-frequency of claims, but high-severity
- Short development period (i.e. claims settle/close quickly)

Final Model & Results

Segmentation Analysis based on Stepwise Regressions

- Identified most significant variables
- Tested interactions
- Identified 2 target markets
  - ✓ 1/3 of book
  - ✓ 15% Better than Average

Insured Amount	Hazard Group		Segment A	Segment B	Segment C	Segment D	Total
> $1 Mil	A, B	Writ Prem	14,646,335	28,851,513	8,323,740	16,368,733	68,190,320
		CR	93.9%	85.5%	96.3%	93.5%	90.5%
		CR Rel	0.90	0.82	0.92	0.90	0.87
> $1 Mil	C, D, E	Writ Prem	17,163,636	23,387,559	9,980,440	16,206,110	66,737,745
		CR	87.2%	100.7%	99.3%	109.0%	99.0%
		CR Rel	0.83	0.96	0.95	1.04	0.95
< $1 Mil	A, B	Writ Prem	13,805,940	38,918,068	7,616,443	27,664,606	88,005,056
		CR	85.3%	99.3%	109.5%	142.4%	111.5%
		CR Rel	0.82	0.95	1.05	1.36	1.07
< $1 Mil	C, D, E	Writ Prem	38,717,344	100,448,260	16,268,948	53,475,254	208,909,805
		CR	87.8%	99.1%	125.6%	132.7%	107.7%
		CR Rel	0.84	0.95	1.20	1.27	1.03
	Total	Writ Prem	84,333,255	191,605,399	42,189,570	113,714,703	431,842,926
		CR	88.3%	97.3%	110.7%	126.1%	104.4%
		CR Rel	0.85	0.93	1.06	1.21	1.00

			Total by Risk Category			
	Total	Writ Prem	137,877,240	196,619,716	97,345,970	431,842,926
		CR	88.8%	101.9%	131.5%	104.4%
		CR Rel	0.85	0.98	1.26	1.00

Iowa Actuaries Club

Example 3:
Pricing Analysis for Equine Insurer

Background:

- Equine (horse) insurance includes Mortality & Medical coverage
- Coverages sold together, but Medical experience nearly **4 times worse** than Mortality

Goals:

- Review overall adequacy of rates
- Reduce subsidization between Mortality & Medical coverages
- Enhance Medical pricing structure to better align price with risk

Data Provided:

- 5 years of historical loss experience = 102,000 horses; 9,300 claims
- Very simple rating structure based on 2 risk characteristics

Equine Analysis: Current Rates

USE OF HORSE	AGE	THOROUGHBRED	QUARTER HORSE APPALOOSA PAINT (Non-Halter)	ARABIAN HALF-ARABIAN NAT'L SHOW HORSE	WARMBLOOD LIPIZZAN AKHAL-TEKE	ANDALUSIAN LUSITANO PASO FINO PERU PASO	PONY ICELANDIC FJORD	MULE DONKEY	FRIESIAN CROSS DRAFT CROSS TN. WALKER OTHER *
Showing *(Excludes jumping)*	2-14	3.5	3.8	2.5	3.2	2.9	2.5	3.2	3.8
Pleasure / Hacking	2-14	3.8	3.6	2.5	3.5	3.0	2.5	3.2	3.8
Dressage	2-14	2.9	2.9	2.5	2.9	2.9	2.5	3.2	2.9
Driving	2-14	3.5	3.8	2.5	2.9	2.9	3.0	3.5	3.5
Show Hunter	2-14	3.5	3.5	3.5	3.5	3.5	3.3	-	3.5
Show Jumper	2-14	3.7	3.7	3.7	3.7	3.7	3.5	-	3.7
Pony Club	2-14	3.6	3.6	3.6	3.6	3.6	3.4	-	3.6
Eventing *(Training level & below)*	2-14	3.8	3.8	3.8	3.8	3.8	3.8	-	3.8
Eventing *(Preliminary level & up)*	2-14	4.1	4.1	4.1	4.1	4.1	4.1	-	4.1
Field Hunter	2-14	3.8	3.8	3.8	3.8	3.8	3.8	-	3.8
Endurance / Distance	2-14	3.0	3.0	3.0	3.0	3.0	3.0	3.2	3.0
Cutting / Team Penning	2-14	3.0	3.0	3.0	-	3.0	3.0	3.0	3.5
Reining	2-14	3.2	3.2	3.2	-	3.2	3.2	3.2	3.5
Roping	2-14	3.8	3.8	3.8	-	3.8	3.8	3.8	3.8
Barrel Racing	2-14	4.2	4.2	4.2	-	4.2	4.2	-	4.2
Foals *(24 hours to 30 days)*		8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0
Foals *(30 days and up)*		7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0
Yearlings	1	3.8	4.2	2.7	3.5	3.1	3.2	3.4	4.0
Breeding Mares	2-14	3.7	4.2	2.7	3.5	2.9	3.0	3.8	3.8
Breeding Stallions	2-14	3.7	4.2	2.5	3.5	2.9	3.0	3.8	3.8

Major Medical & Surgical ($7,500 limit) = $275 per horse (ages 2-15 years); $325 per horse (6 mths - 1 year)
Major Medical & Surgical ($10,000 limit) = $400 per horse (ages 2-15 years); $450 per horse (6 mths - 1 year)

Equine Analysis: Final Results

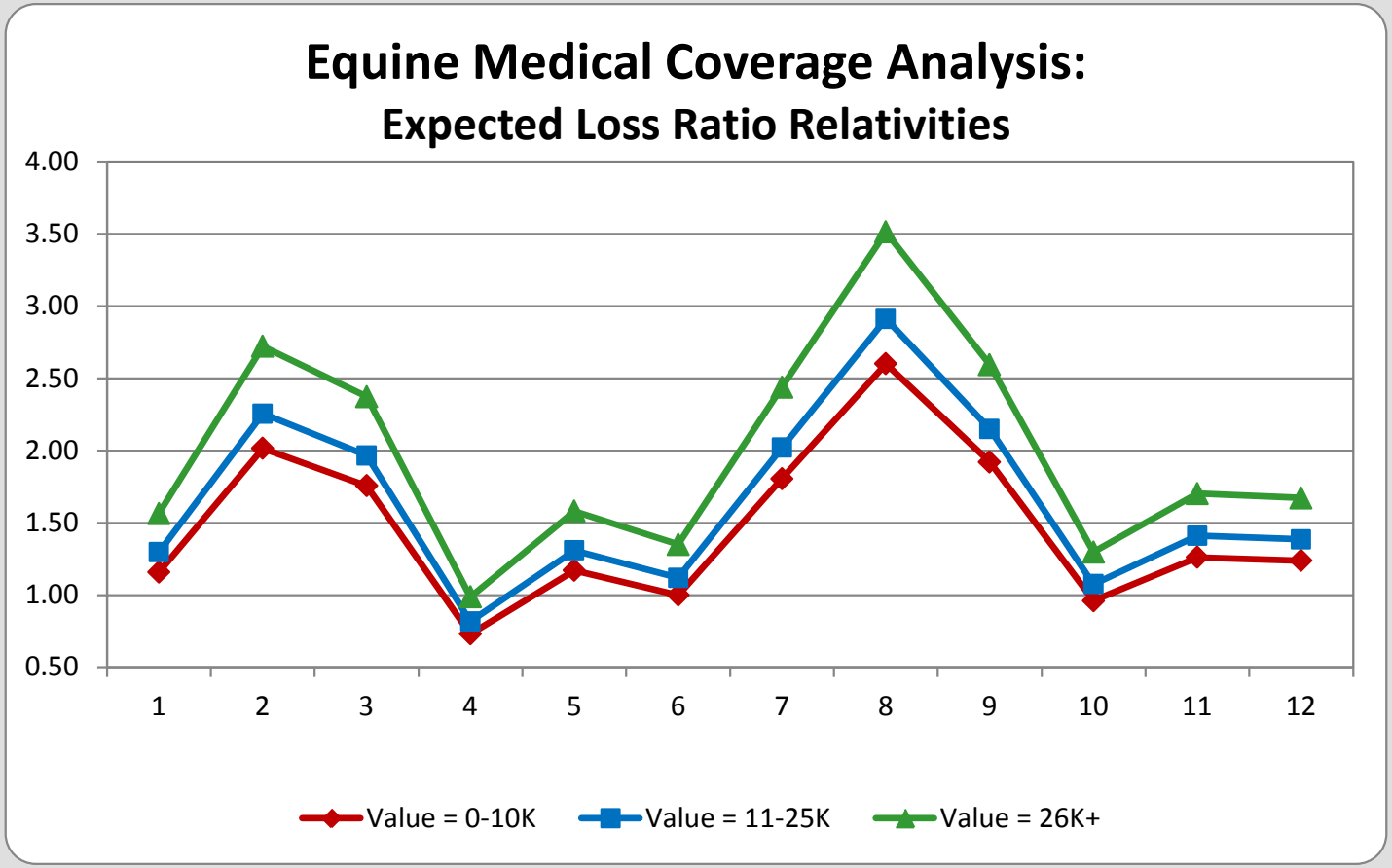
Overall Rate Level Indication

- Calculated the need for 27.4% overall increase
- Indicated -27.9% decrease for Mortality and **198.9% increase** for Medical

Generalized Linear Model (Medical Coverage only)

- Current = Horse Age & Amount of Coverage
  - ✓ 4 rates: Range = $275 - $450
- Indicated = Breed, Use, Horse Value, Horse Age, Amount of Coverage, & New vs. Renewal
- Proposed = Breed, Use, Horse Value, & Amount of Coverage
  - ✓ 36 rates: Range = $400 - $2,000

Equine Analysis: Final Results



Recent Competitions (Kaggle)

Deloitte

- "As the World Churns"
- Seeks a solution for predicting which current customers of an insurance company will leave in 12 months time, and when
- 2013; $70,000 prize

Heritage Provider Network

- Sponsored the Heritage Health Prize Competition (the "Competition")
- Goal of developing a breakthrough algorithm that uses available patient data to predict and prevent unnecessary hospitalizations
- 2012; $3 million grand prize

Recent Competitions (Kaggle)

Allstate

- Claim Prediction Challenge**:** The goal of this competition is to predict Bodily Injury Liability Insurance claim payments based on the characteristics of the insured's vehicle
- 2011; $10,000 prize

- Will I Stay or Will I Go? The goal of this competition is to predict which current customers will still be with the company in 6 months, given many of the customer's characteristics
- 2012

Employment Opportunities – Lots!

- **Nationwide**: Predictive Modeler; PC Act. Research Analytics
- **Principal Financial**: Sr. Analytics Consultant; Big Data Developer / Engineer
- **Allstate:** Sr. Predictive Modeler
  - ✓ **Desired Skills and Experience:**
    Expertise in statistical modeling techniques such as generalized linear models, tree models (CART, MART, Random Forest), survival analysis, gradient boosting methods, data visualization, cluster analysis, principal components and feature creation, validation.
- Many others …

Questions?