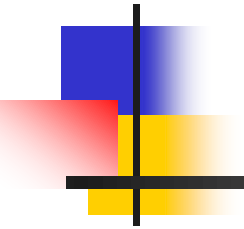
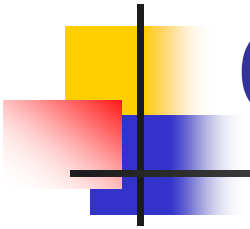
Sports Predictive Analytics: NFL Prediction Model



By
Dr. Ash Pahwa

Orange County R User's Group

April 27, 2017



Outline

- Case Studies of Sports Analytics
- Sports
- Sports Analytics
- Applications of Sports Analytics
- Sports Analytics Literature
- Data Sources
- Sports Predictive Models
- Regression Model
- Multi Variable Regression with Lasso
- NFL Prediction Model
- Prediction for Super Bowl 2016
- Prediction for NFL 2017 Playoffs

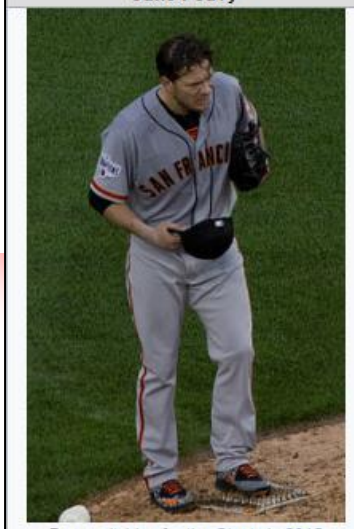
What is Sports Analytics? Which Pitcher is Better?

2007 Baseball	
Jake Peavy	John Lackey
San Diego Padres : National League	L.A. Angles : American League
ERA: 2.54	ERA: 3.01

ERA: Earned Run Average. Mean number of runs yielded per 9 innings

- American league allows Designated Hitter (DH) for pitcher
- National league does not allow Designated Hitter (DH) for pitcher. Pitcher must bat.

Jake Peavy



Peavy pitching for the Giants in 2015

Free agent

Starting pitcher

Born: May 31, 1981 (age 35)
Mobile, Alabama

Bats: Right	**Throws:** Right

MLB debut

June 22, 2002, for the San Diego Padres

MLB statistics (through 2016 season)

Win–loss record	152–126
Earned run average	3.63
Strikeouts	2,207
WHIP	1.20

Teams

- San Diego Padres (2002–2009)
- Chicago White Sox (2009–2013)
- Boston Red Sox (2013–2014)
- San Francisco Giants (2014–2016)

Career highlights and awards

- 2× World Series champion (2013, 2014)
- 3× All-Star (2005, 2007, 2012)
- NL Cy Young Award (2007)
- Triple Crown (2007)
- NL wins leader (2007)
- 2× MLB ERA leader (2004, 2007)
- 2× NL strikeout leader (2005, 2007)
- Gold Glove Award (2012)
- San Diego Padres all-time strikeouts leader

John Lackey

Lackey with the Chicago Cubs in 2016

Chicago Cubs – No. 41

Starting pitcher

Born: October 23, 1978 (age 38)
Abilene, Texas

Bats: Right	**Throws:** Right

MLB debut

June 24, 2002, for the Anaheim Angels

MLB statistics (through 2016 season)

Win–loss record	176–135
Earned run average	3.88
Strikeouts	2,145

Teams

- Anaheim Angels / Los Angeles Angels of Anaheim (2002–2009)
- Boston Red Sox (2010–2014)
- St. Louis Cardinals (2014–2015)
- Chicago Cubs (2016–present)

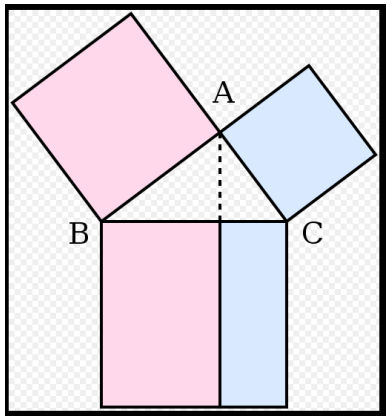
Career highlights and awards

- All-Star (2007)
- 3× World Series champion (2002, 2013, 2016)
- AL ERA leader (2007)

Which Variable is the Best Predictor of the Winning Percentage?

	A	B	C	D	E	F	G	H	I	J
1		Team	Winning Percentage	Runs Scored	Home Runs	Team Batting Average	On-Base Percentage	Batting Average Against Team	Team Earned-Run Average	
2	1	Arizona	0.556	712	171	0.25	0.321	0.262	4.13	
3	2	Atlanta	0.519	810	176	0.275	0.339	0.259	4.11	
4	3	Chicago Cubs	0.525	752	151	0.271	0.333	0.246	4.04	
5	4	Cincinnati	0.444	783	204	0.267	0.335	0.282	4.94	
6	5	Colorado	0.522	860	171	0.28	0.354	0.266	4.32	
7	6	Florida	0.438	790	201	0.267	0.336	0.285	4.94	
8	7	Houston	0.451	723	167	0.26	0.33	0.273	4.68	
9	8	Los Angeles	0.506	735	129	0.275	0.337	0.261	4.2	
10	9	Milwaukee	0.512	801	231	0.262	0.329	0.269	4.41	
11	10	New York Mets	0.543	804	177	0.275	0.342	0.255	4.26	
12	11	Philadelphia	0.549	892	213	0.274	0.354	0.276	4.73	
13	12	Pittsburgh	0.42	724	148	0.263	0.325	0.288	4.93	
14	13	San Diego	0.546	741	171	0.251	0.322	0.25	3.7	
15	14	San Francisco	0.438	683	131	0.254	0.322	0.261	4.19	
16	15	St. Louis	0.481	725	141	0.274	0.337	0.271	4.65	
17	16	Washington	0.451	673	123	0.256	0.325	0.269	4.58	
18										

Pythagorean Theorem

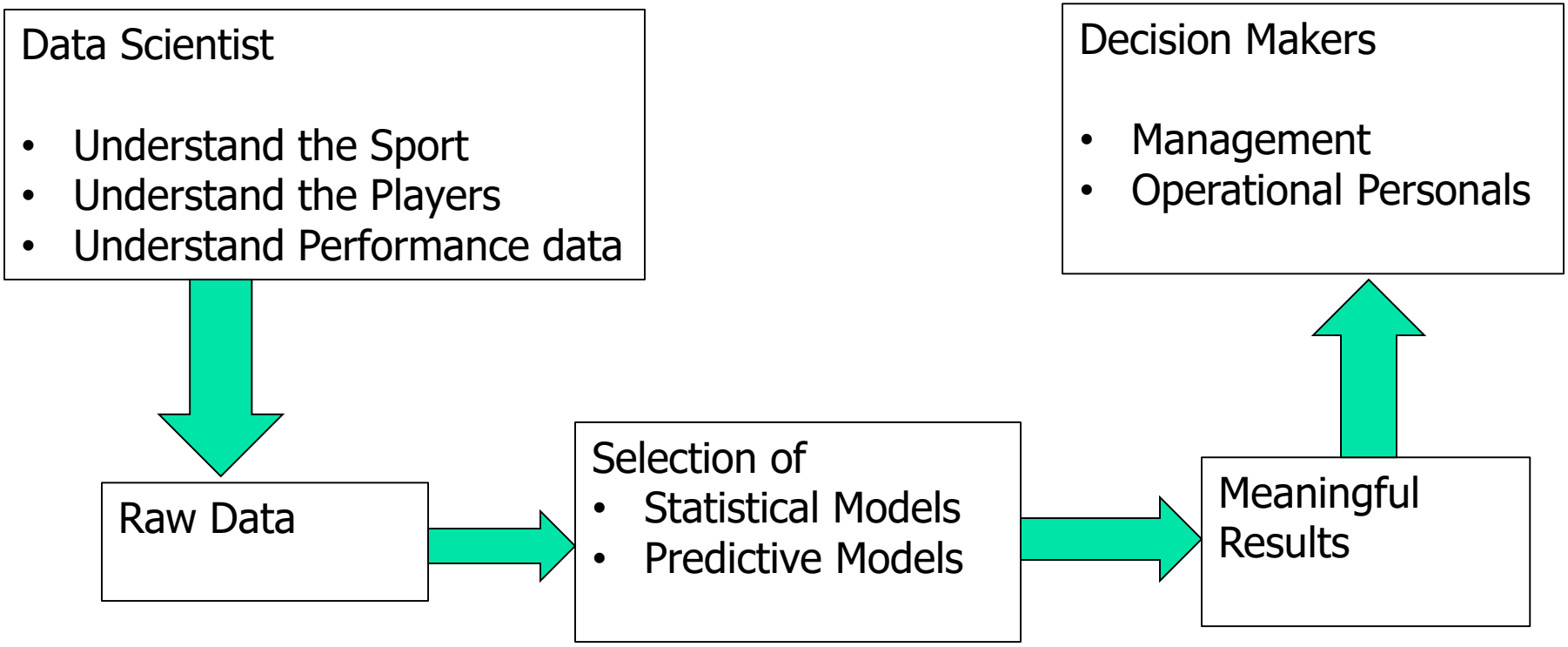


- Used in Baseball. Proposed by Bill James
- Suppose
  - 'F' represents a team's run scored
  - 'A' represents team's runs allowed
- $Pyhtagorean\ Winning\ Percentage = \frac{F^2}{F^2+A^2}$
- It is called Pythagorean theorem because it is similar to the elementary geometry theorem

Example:

- Year 2012: Detroit Tigers
  - Scored Runs = F = 726
  - Allowed Runs = A = 670
- $Pyhtagorean\ Winning\ Percentage = \frac{F^2}{F^2+A^2} = \frac{726^2}{726^2+670^2} = \frac{527{,}076}{975{,}976} = 0.54$
- Total games won = 162*0.54 = 88 games

How to Convey Information to Decision Makers



Goals of Sports Analytics

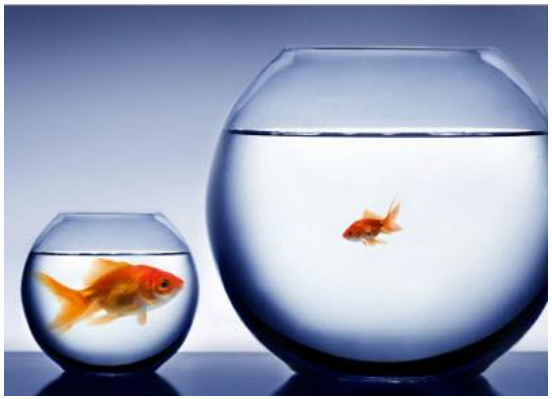
1. Apply Statistical Models to Sporting Data
2. Ratings and Rankings
3. Predictive Models
4. Player and Team Assessment

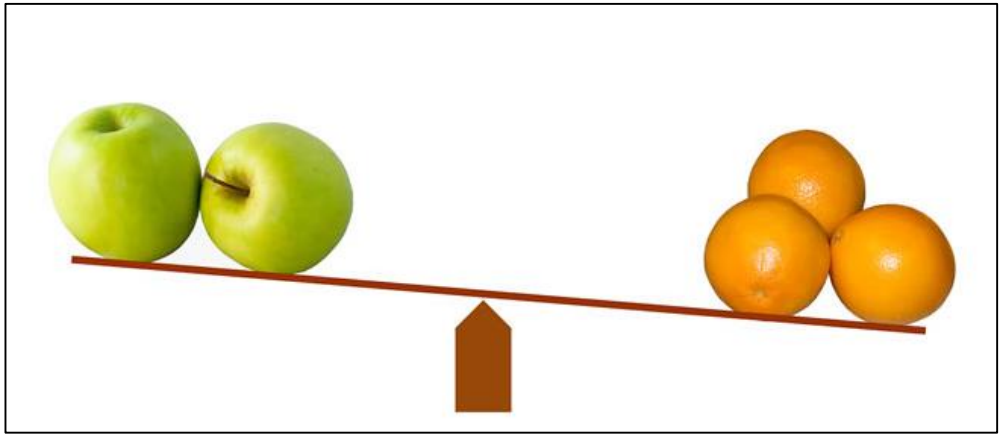
Statistical Models
Predictive Models

Indices of Central Tendencies and Variability	Statistics Used to Examine Relationships	Inferential Statistics	Ratings + Rankings	Predictive Models
Histogram (Frequency Distribution)	Normal Distribution (z-values and p-values)	Normality	Ratings + Rankings	Simple Linear Regression
Mean	Covariance	Outlier	Rank Aggregation	Multiple Linear Regression
Median	Correlation – Pearson	t-test		Polynomial Regression
Mode	Rank Correlation – Spearman	ANOVA		Logistic Regression
Range, Variance	Partial Correlation	Chi-Square		
Standard Deviation				

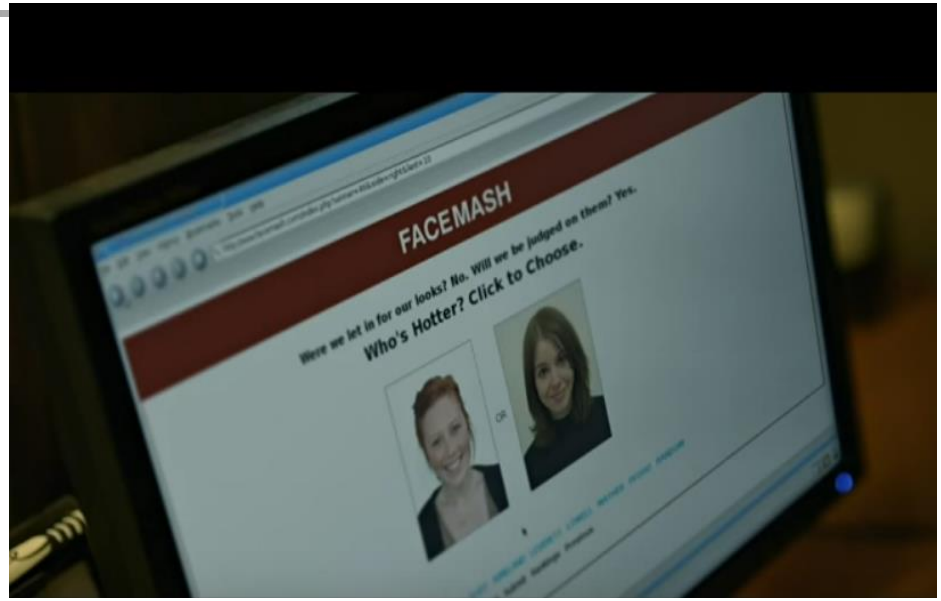
Ranking of Players and Teams?

Pair-Wise Comparison





Pair-Wise Comparison
The Social Network



Pair-Wise Comparison Can be Used for Ranking

		Duke	Diff			Miami	Diff			UNC	Diff			UVA	Diff			VT	Diff		Total Diff
Duke			0		7	52	-45		21	24	-3		7	38	-31		0	45	-45		-124
Miami	52	7	45				0		34	16	18		25	17	8		27	7	20		91
UNC	24	21	3		16	34	-18				0		7	5	2		3	30	-27		-40
UVA	38	7	31		17	25	-8		5	7	-2				0		14	52	-38		-17
VT	45	0	45		7	27	-20		30	3	27		52	14	38				0		90

Sorting the Rating list produces Ranking

	Ratings	Ranking
Miami	91	1
VT	90	2
UVA	-17	3
UNC	-40	4
Duke	-124	5

Log 5 Method

- Developed by Bill James in 1970s
- Computes the probability that Team A will beat Team B
- Log 5 formula has nothing to do with the mathematical function 'Log'

Log 5 Formula

$$p_{a,b} = \frac{p_a - p_a * p_b}{p_a + p_b - 2 * p_a * p_b}$$

- Suppose Team A true winning percentage is 10 out of 16 games
  - Percentage of true winning = p_a = 10/16 = 0.625
- Suppose Team B true winning percentage is 7 out of 16 games
  - Percentage of true winning = p_b = 7/16 = 0.438
- ---
- The probability that Team A will beat Team B
- $p_{a,b} = \frac{p_a - p_a * p_b}{p_a + p_b - 2 * p_a * p_b} = \frac{0.625 - 0.625 * 0.438}{0.625 + 0.438 - 2 * 0.625 * 0.438} = \frac{0.625 - 0.274}{1.063 - 2 * 0.274} = \frac{0.351}{0.515} = 0.681$
- ---
- The probability that Team B will beat Team A
- $p_{b,a} = \frac{p_b - p_a * p_b}{p_a + p_b - 2 * p_a * p_b} = \frac{0.438 - 0.625 * 0.438}{0.625 + 0.438 - 2 * 0.625 * 0.438} = \frac{0.438 - 0.274}{1.063 - 2 * 0.274} = \frac{0.164}{0.515} = 0.318$
- ---
- $p_{a,b} + p_{b,a} = 1$

Arpad Elo

- Physics Professor at Marquette University
  - Milwaukee, Wisconsin
- Chess Player
- Devised a method to rank chess players
- His method was adopted by
  - US Chess Federation
  - World Chess Federation

Arpad Elo

Born	Élő Árpád Imre August 25, 1903 Egyházaskesző, Austro-Hungarian Empire
Died	November 5, 1992 (aged 89) Brookfield, Wisconsin
Nationality	Hungarian American
Fields	Physics
Institutions	Marquette University
Alma mater	University of Chicago

Points Gained or Lost

Points gained/lost by player A = Points gained/lost by player B

- *Before the game*
  - *Player A rating* r_A
  - *Player B rating* r_B
  - *Difference* $d_{AB} = r_A - r_B$

- *After the game*
  - *Player A rating* r'_A
  - *Player B rating* r'_B
  - $r_A + r_B = r'_A + r'_B$

- *Points gained/lost by player A against player B*
  - $\mu_{AB} = L\left(\frac{d_{AB}}{400}\right) = \frac{1}{1+10^{\frac{-d_{AB}}{400}}}$

Example

	If Player A wins	Draw	If Player B wins
S(AB)	1	0.5	0
S(BA)	0	0.5	1

- *Before the game*
  - $Player\ A\ rating\ r_A = 2400$
  - $Player\ B\ rating\ r_B = 2000$
  - $Difference\ d_{AB} = r_A - r_B = 400$
  - $Difference\ d_{BA} = r_B - r_A = -400$

Suppose K = 32 for Chess

$$\mu_{AB} = 0.91$$
$$\mu_{BA} = 0.09$$

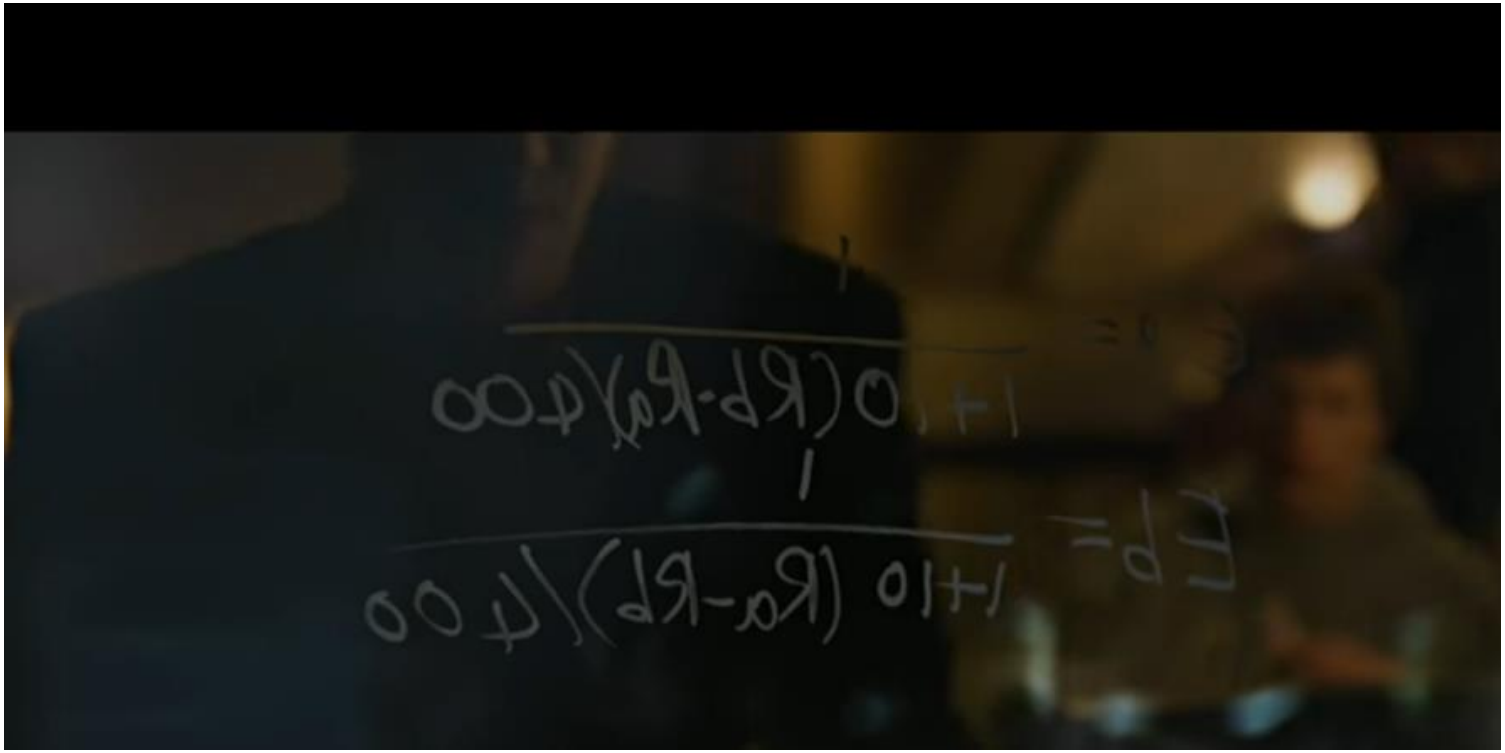
If Player A wins

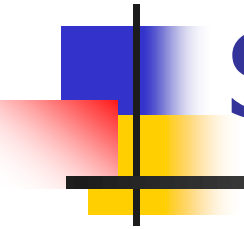
- *After the game*
  - $Player\ A\ rating\ r'_A = r_A + K(S_{AB} - \mu_{AB}) = 2400 + 32(1 - 0.91) = 2403$
  - $Player\ B\ rating\ r'_B = r_B + K(S_{BA} - \mu_{BA}) = 2000 + 32(0 - 0.09) = 1997$
- $r_A + r_B = r'_A + r'_B$
- $2400 + 2000 = 2403 + 1997$

If Player B wins

- *After the game*
  - $Player\ A\ rating\ r'_A = r_A + K(S_{AB} - \mu_{AB}) = 2400 + 32(0 - 0.91) = 2371$
  - $Player\ B\ rating\ r'_B = r_B + K(S_{BA} - \mu_{BA}) = 2000 + 32(1 - 0.09) = 2029$
- $r_A + r_B = r'_A + r'_B$
- $2400 + 2000 = 2371 + 2029$

The Social Network
Elo Formula was Used to pair-wise Comparison of Girls





Sports

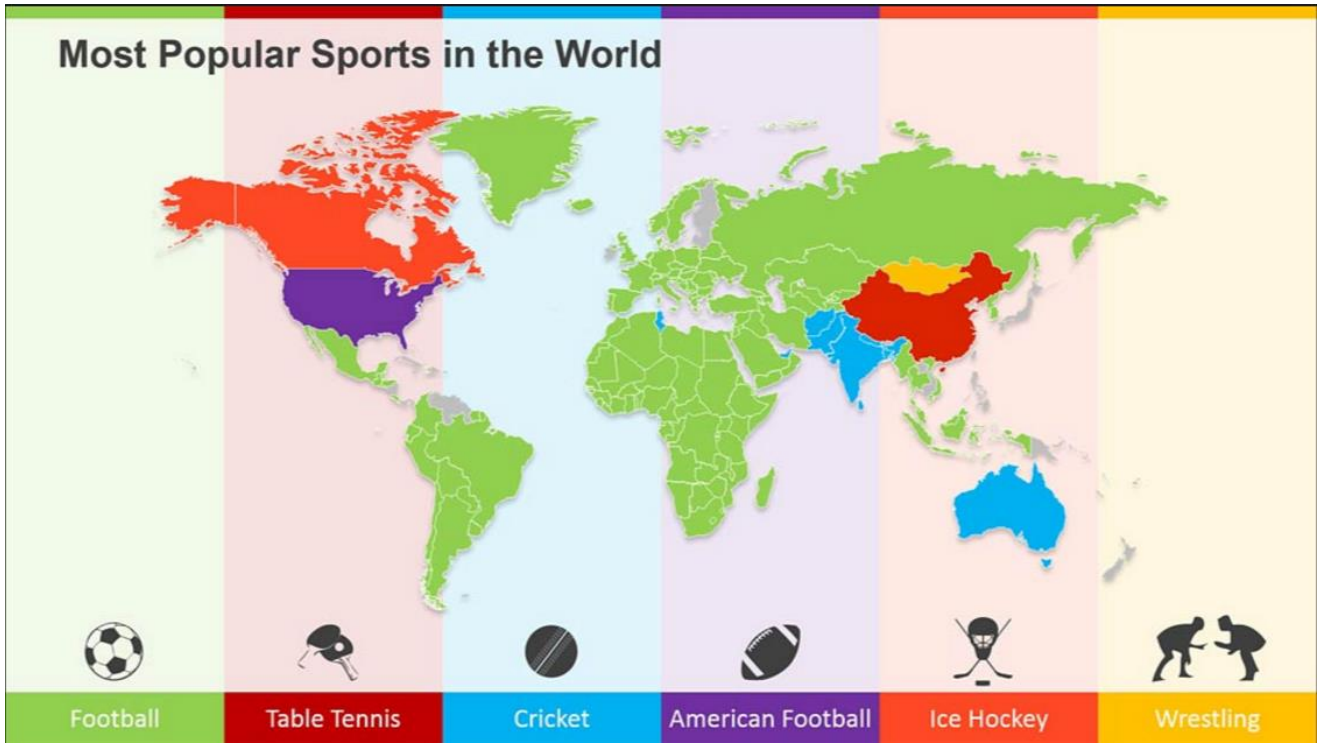
Sports

- Inherent part of Human Culture
- Sports competition dates back to the dawn of our species
- Greek Olympics : 776 BC

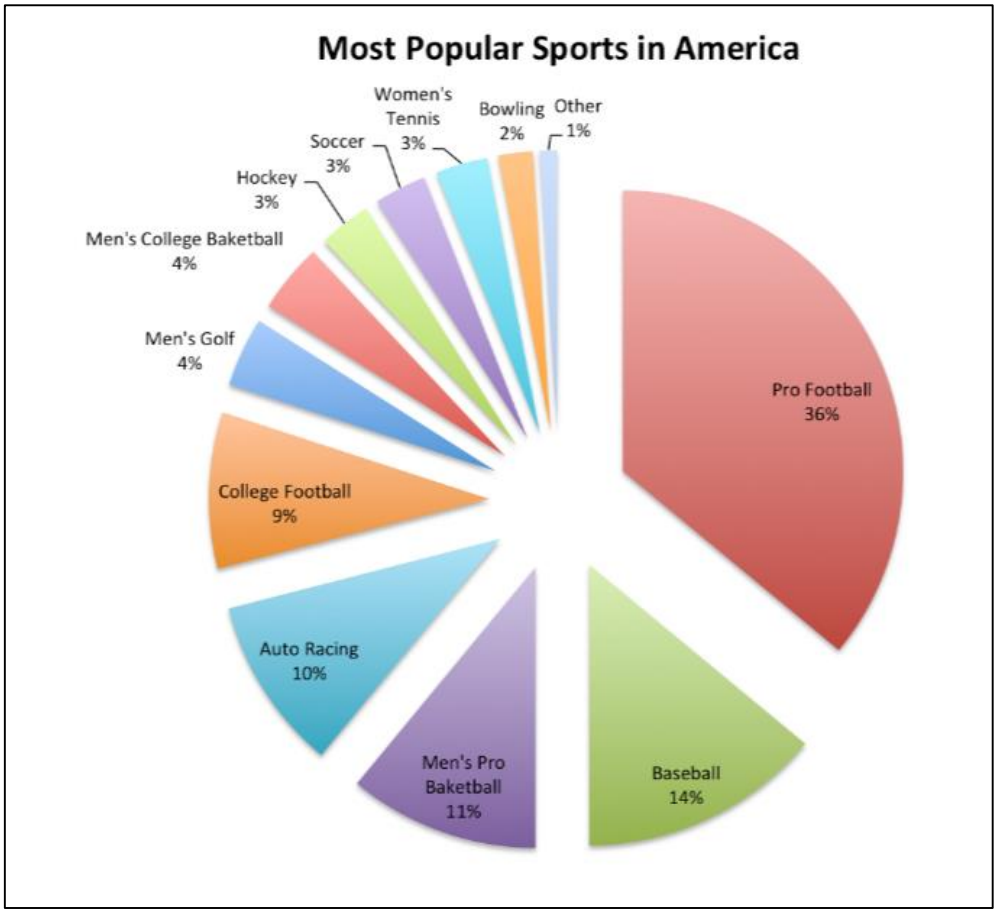
Sportsman Spirit

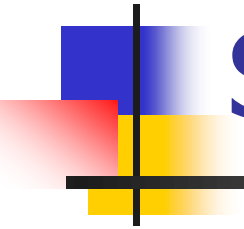
- Virtues of sports
  - fairness
  - self-control
  - courage
  - persistence
- It has been associated with interpersonal concepts of treating others and being treated fairly
- Maintaining self-control if dealing with others, and respect for both authority and opponents
- **GOOD SPORTSMEN HAVE ALWAYS BEEN HELD IN HIGH ESTEEM**

Most Popular Sports in the World



Most Popular Sports in America





Sports Analytics

Predictive Models

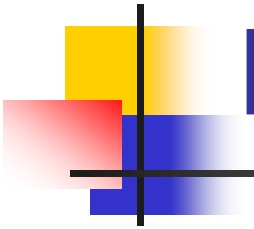
- Estimation
  - Regression
- Classification (win/loss)
  - Logistic Regression
  - Discriminant Analysis
    - Linear
    - Quadratic
  - Support Vector Machine

Goals of Sports Analytics Player

- Discovering hidden talent in a new player
- Player Evaluation
  - Assessing Player Performance
    - Which metrics is most important to assess a players' performance
  - Assessing Player Value
  - How much value a player adds to the teams' value

Goals of Sports Analytics Team

- Ranking top teams
- Accessing Team Performance
  - How to compute the value of a team
- Which Team Members are best suited to play against the opposing team
- Which strategy to use to play against a team?
  - Anticipating Opponents Behavior
- Accessing the probability of a win in a sporting event



Need for Prediction Results

- Betting on an sporting event
  - People betting on sports need to see the prediction results
  - Probability of a win
  - Point Spread
- Fantasy Sports
  - DraftKings
  - FanDuel

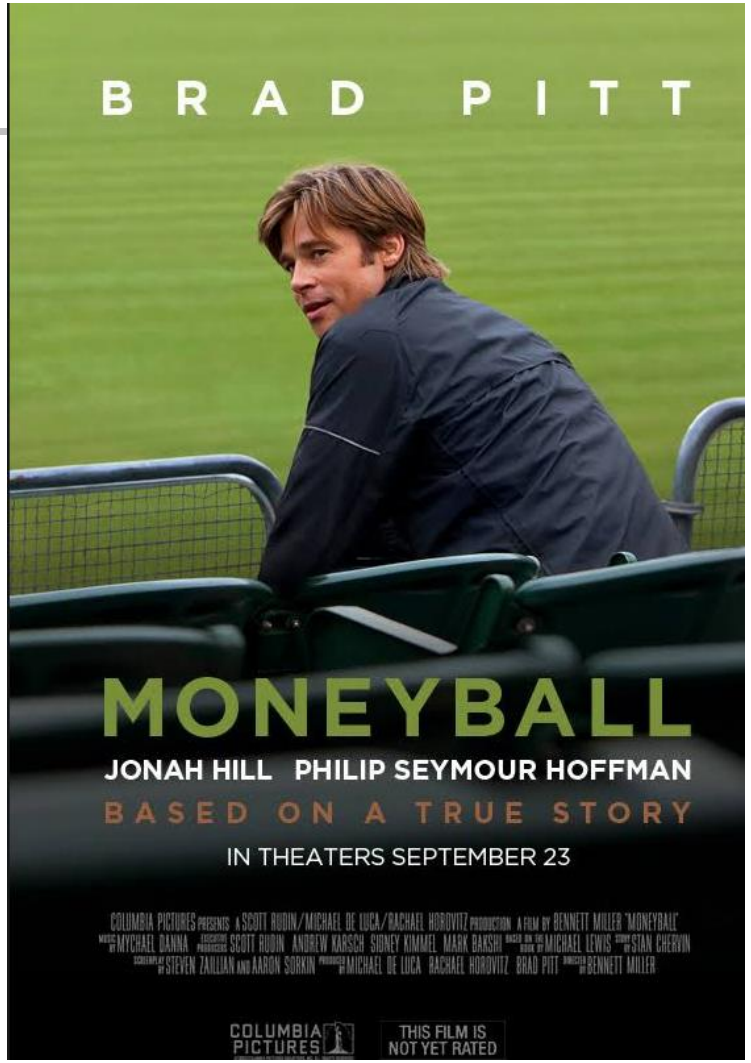
Applications of Sports Analytics

Application of Sports PA Movies

The film is based on Michael Lewis' 2003 nonfiction book of the same name, an account of the Oakland Athletics baseball team's 2002 season and their general manager Billy Beane's attempts to assemble a competitive team.

In the film, Beane (Brad Pitt) and assistant GM Peter Brand (Jonah Hill), faced with the franchise's limited budget for players, build a team of undervalued talent by taking a sophisticated sabermetric approach towards scouting and analyzing players.

They acquire "submarine" pitcher Chad Bradford (Casey Bond) and former catcher Scott Hatteberg (Chris Pratt), and win 20 consecutive games, an American League record.



Moneyball – Billy Beane & Paul DePodesta

William Lamar "Billy" Beane III (born March 29, 1962) is an American former professional baseball player and current front office executive.

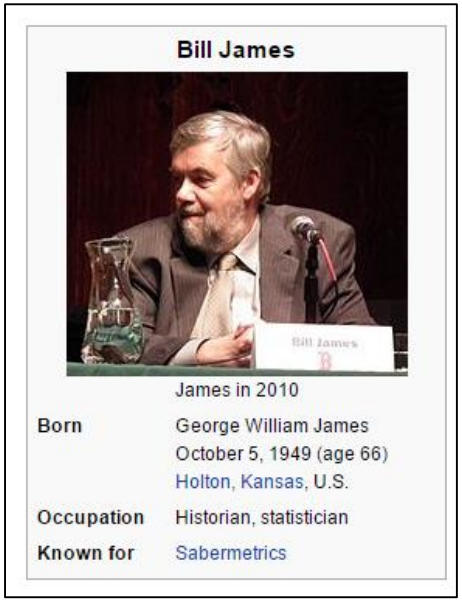
He is the Executive Vice President of Baseball Operations and minority owner of the Oakland Athletics of Major League Baseball (MLB).

The character of Brand is an invention by the filmmakers; in the excellent Michael Lewis non-fiction book upon which the movie is based, the real-life "Brand" is identified as Paul DePodesta.

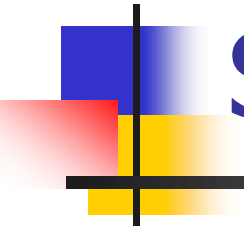
Unlike Brand, DePodesta is slender, fit and handsome. He's also Harvard-educated (not a Yalie – screenwriter Aaron Sorkin's private joke).

Application of Sports PA Boston Red Sox

- Using Predictive Analytics Strategies Boston Red Sox won 3 world series in Baseball

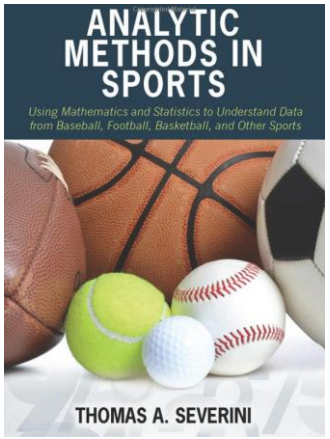


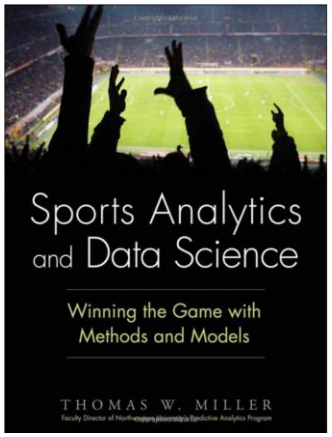
In 2006, Time named Bill James in the Time 100 as one of the most influential people in the world. He is a Senior Advisor on Baseball Operations for the Boston Red Sox.

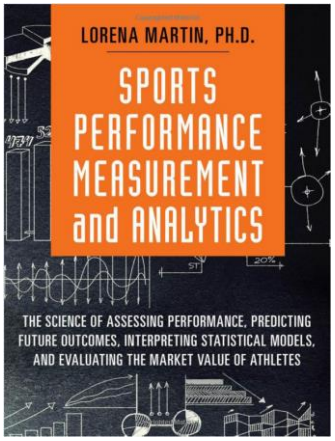


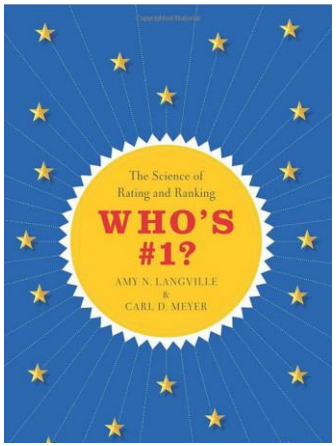
Sports Analytics Literature

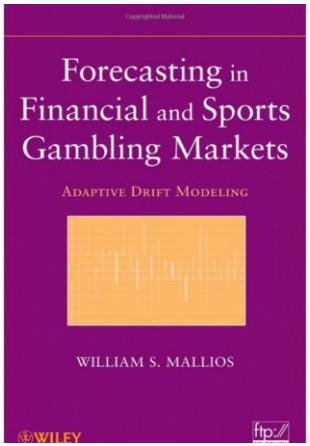
Predictive Models for Sports Literature

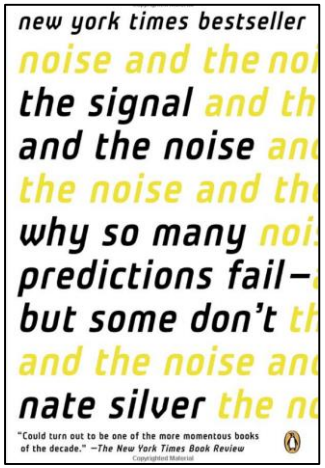








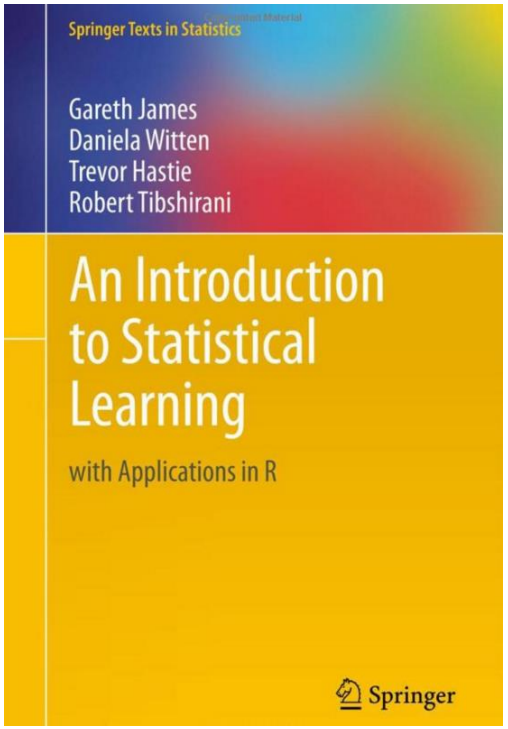


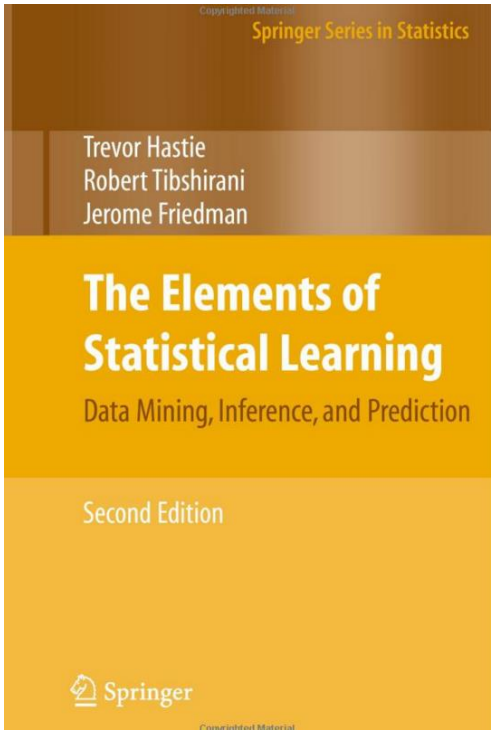


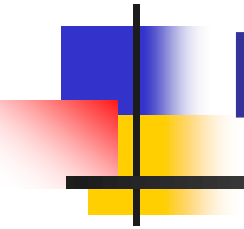
Statistical Learning

- Gareth James
- Daniela Witten
- Trevor Hastie
- Robert Tibshirani

- Stanford University



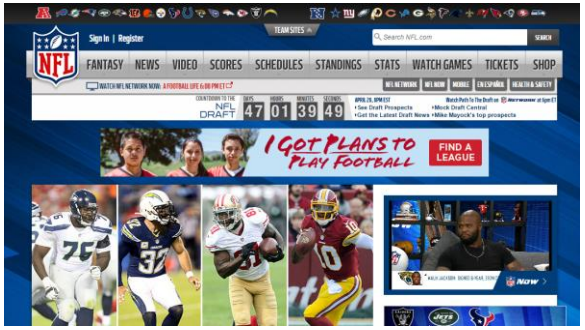


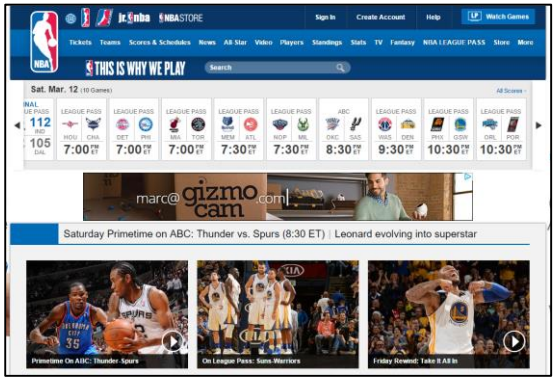


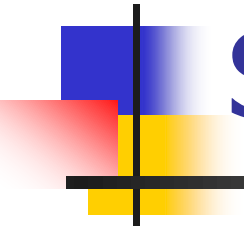
Data Sources

Data Sources

- www.NFL.com
- www.NBA.com
- www.footballOutsiders.com
- www.pro-football-reference.com
- www.soccerstats.com
- www.basketball-reference.com







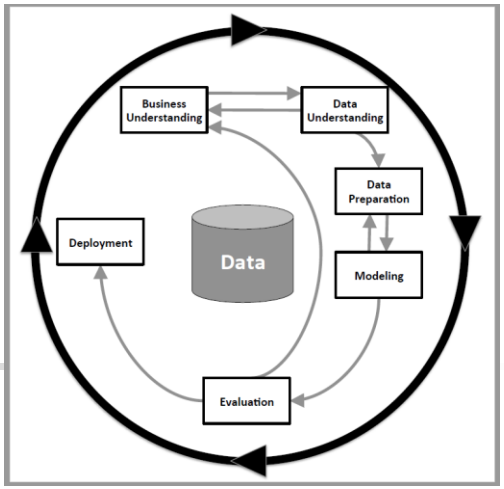
Sports Predictive Models

Goals of Predictive Analytics Application: Estimation or Classification

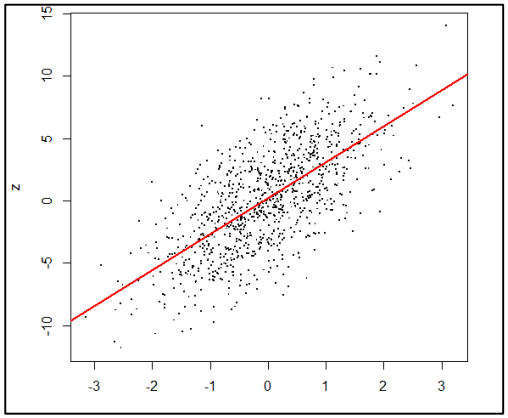
- Estimation – Regression modeling technique is used
  - Output is a number
    - House price
    - Product sales for next quarter
    - GNP growth for the next quarter
    - How many points a team will score

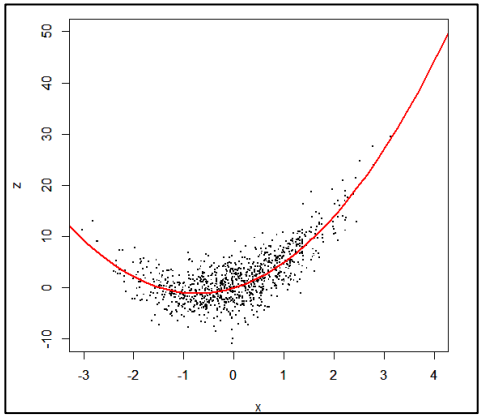
- Classification
  - Logistic Regression
  - Support Vector Machine
  - Discriminant Analysis (Linear, Quadratic)
  - Naïve Bayes, Decision Trees etc. modeling techniques are used
  - Output is a categorical variable
    - Sports team will win or lose
    - Email is junk or not
    - Which grade student will get
    - Tweet is positive or negative

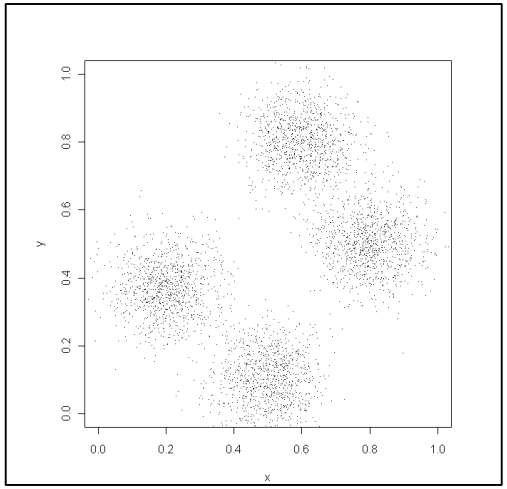
Common PA Techniques

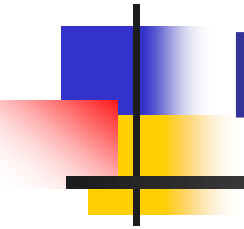


- Regression
  - Linear 2 variables
  - Linear multi variables
  - Logistic
  - Polynomial
- Clustering
- Decision Trees
- Neural Networks
- Naïve Bayes
- ARIMA
- A few more …









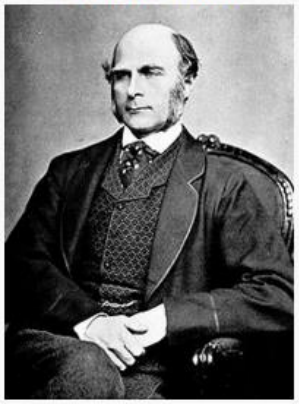
Regression Model

History of Linear Regression

- Sir Francis Galton and
- Karl Pearson
- Developed the concepts on Regression and Correlation in 1900 - 1930

Galton aged 87, with Karl Pearson.

Sir Francis Galton



Born	16 February 1822 Birmingham, West Midlands, England, United Kingdom
Died	17 January 1911 (aged 88) Haslemere, Surrey, England, United Kingdom
Residence	England
Nationality	British
Fields	Anthropology, Sociology
Institutions	Meteorological Council Royal Geographical Society
Alma mater	King's College London Trinity College, Cambridge
Academic advisors	William Hopkins
Notable students	Karl Pearson
Known for	Eugenics The Galton board Regression toward the mean Standard deviation Weather map
Notable awards	Royal Medal (1886) Darwin–Wallace Medal (Silver, 1908) Copley Medal (1910)

Karl Pearson

Portrait of Karl Pearson, by Elliott & Fry, 1890.

Born	Carl Pearson 27 March 1857 Islington, London, England
Died	27 April 1936 (aged 79) Coldharbour, Surrey, England
Residence	England
Nationality	British
Fields	Lawyer, Germanist, eugenicist, mathematician and statistician (primarily the last)
Institutions	University College London King's College, Cambridge
Alma mater	University of Cambridge University of Heidelberg
Academic advisors	Francis Galton

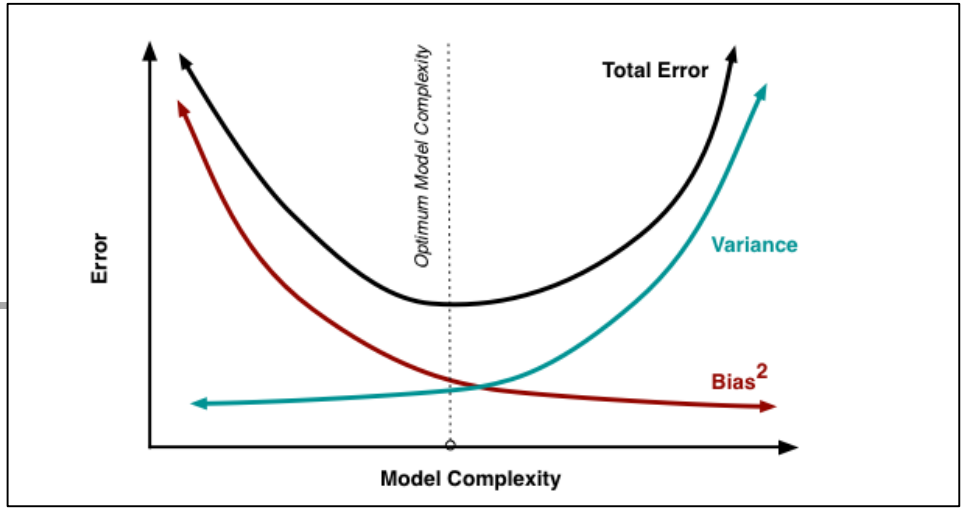
Definition: Linear Regression

- 2 Variable Regression
  - How a response variable "y" changes
  - As the predictor (explanatory)
    - variable "x" changes
- Multiple Regression
  - How a response variable "y" changes
  - As the predictor (explanatory)
    - variables "x1", "x2", … "xn" change

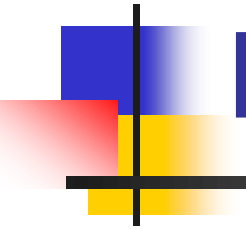
$$y = \beta_1 x + c$$

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n + c$$

Bias and Variance



- Goal is to find out a model complexity where
  - Generalization (Validation) errors are least
  - Bias + variance are least
- $Mean\ Square\ Error = Bias^2 + Variance$
- Just like Generalization Error
  - We cannot compute Bias and Variance



Lasso Regression

Least Absolute Shrinkage and Selection Operator

Cost Function of OLS + Ridge + Lasso

OLS

$$Cost\ (W) = RSS(W) = \sum_{i=1}^{N} \{y_i - \hat{y}_i\}^2 = \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2$$

Ridge Regression

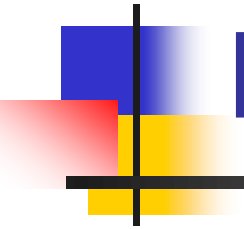
$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

Lasso Regression

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$



NFL Model

Super Bowl 2016
All Predicted Models were Wrong

	Panthers	Broncos
Nate Silver	59%	41%
A+	56.5%	43.5%

2017 NFL Prediction Wild Card : Playoff

- Predictions for all the 4 games were 100% correct

01-07-2017	Green Bay Probability: 73.18% Green Bay won 38-13	New York G.
01-07-2017	Pittsburgh Probability: 75.58% Pittsburgh won 30-12	Miami
01-06-2017	Seattle Probability: 77.36% Seattle won 26-6	Detroit
01-06-2017	Houston Probability: 56.57% Houston won 27-14	Oakland

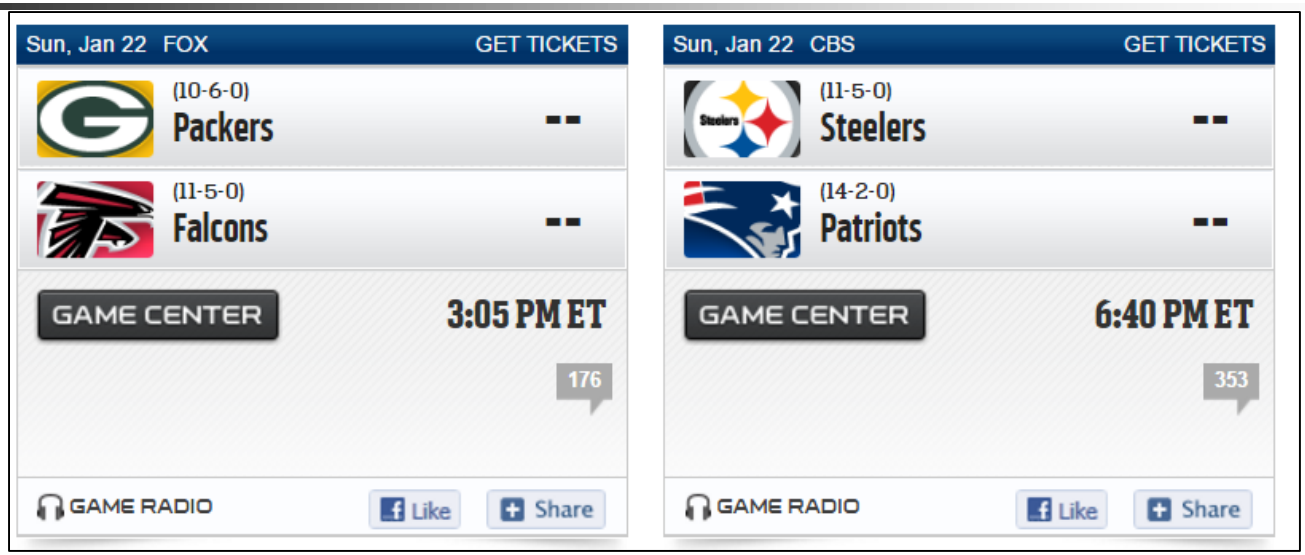
2017 NFL Prediction Divisional Title: Playoff

www.NFLPrediction.co

- Predictions
  - 2 correct
  - 2 incorrect
  - 50% correct

GAME DATE	PREDICTED WINNER	PREDICTED LOSER
01-15-2017	Dallas Probability: 63.02%	Green Bay
01-15-2017	Kansas City Probability: 59.21%	Pittsburgh
01-14-2017	New England Probability: 81.87%	Houston
01-14-2017	Atlanta Probability: 59.26%	Seattle

2017 NFL Prediction
Conference Championship: Playoff



- Atlanta Falcons vs Green Bay Packers
  - Atlanta Falcons 61%
- New England Patriots vs Pittsburgh Steelers
  - New England Patriots 70%

2017 NFL Prediction
Super Bowl

- New England Patriots vs
- Atlanta Falcons
  - New England Patriots 60%

Summary

- Sports
- Sports Analytics
- Applications of Sports Analytics
- Sports Analytics Literature
- Data Sources
- Sports Predictive Models
- Regression Model
- Multi Variable Regression with Lasso
- NFL Prediction Model
- Prediction for Super Bowl 2016
- Prediction for NFL 2017 Playoffs