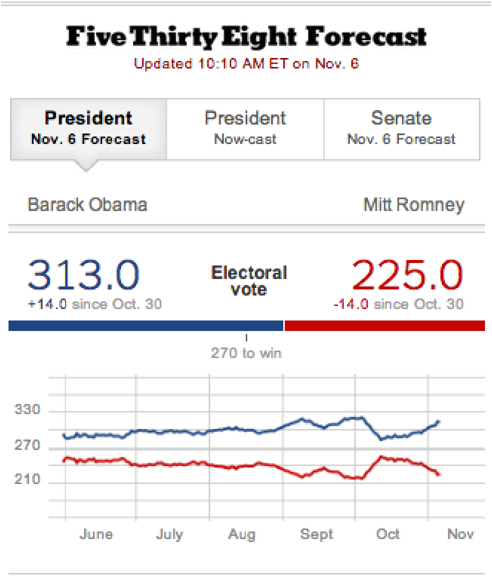
What data should you use?

Jeffrey Leek

May 18, 2016

A succcessful predictor



FiveThirtyEight Forecast

Updated 10:10 AM ET on Nov. 6

President **Nov. 6 Forecast**	President Now-cast	Senate Nov. 6 Forecast

Barack Obama

Mitt Romney

313.0

+14.0 since Oct. 30

Electoral vote

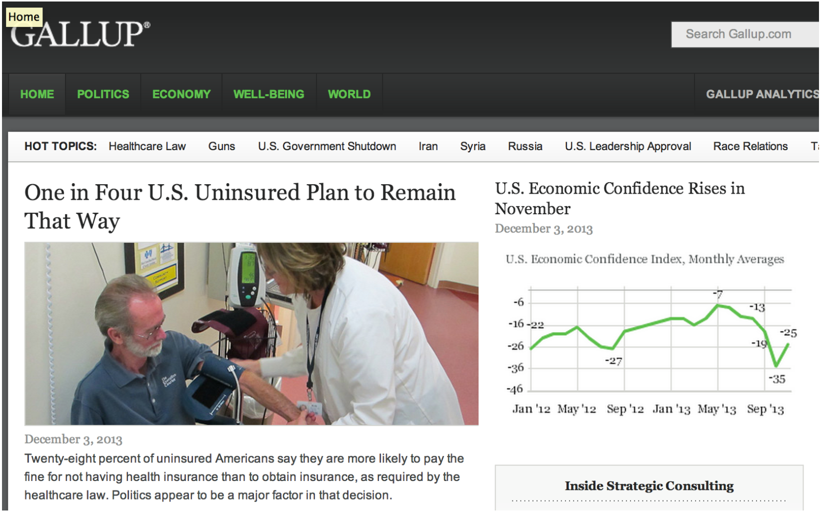
225.0

-14.0 since Oct. 30

270 to win

fivethirtyeight.com

Polling data



Home

GALLUP®

Search Gallup.com

HOME POLITICS ECONOMY WELL-BEING WORLD GALLUP ANALYTICS

HOT TOPICS: Healthcare Law Guns U.S. Government Shutdown Iran Syria Russia U.S. Leadership Approval Race Relations T

One in Four U.S. Uninsured Plan to Remain That Way

December 3, 2013

Twenty-eight percent of uninsured Americans say they are more likely to pay the fine for not having health insurance than to obtain insurance, as required by the healthcare law. Politics appear to be a major factor in that decision.

U.S. Economic Confidence Rises in November

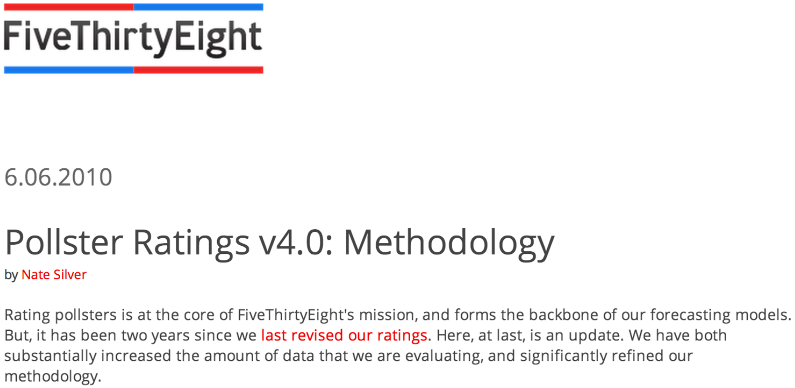
December 3, 2013

U.S. Economic Confidence Index, Monthly Averages

Inside Strategic Consulting

http://www.gallup.com/

Weighting the data



FiveThirtyEight

6.06.2010

Pollster Ratings v4.0: Methodology

by Nate Silver

Rating pollsters is at the core of FiveThirtyEight's mission, and forms the backbone of our forecasting models. But, it has been two years since we last revised our ratings. Here, at last, is an update. We have both substantially increased the amount of data that we are evaluating, and significantly refined our methodology.

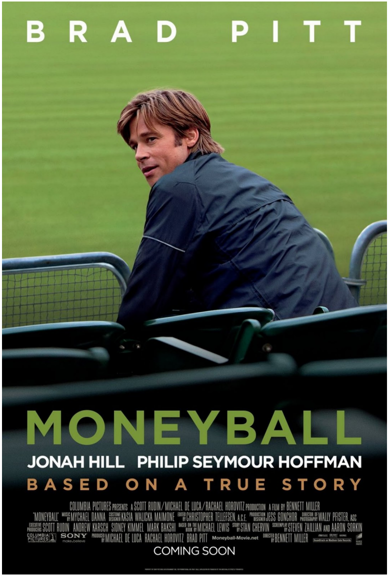
`http://www.fivethirtyeight.com/2010/06/pollster-ratings-v40-methodology.html`

Key idea

To predict X use data related to X

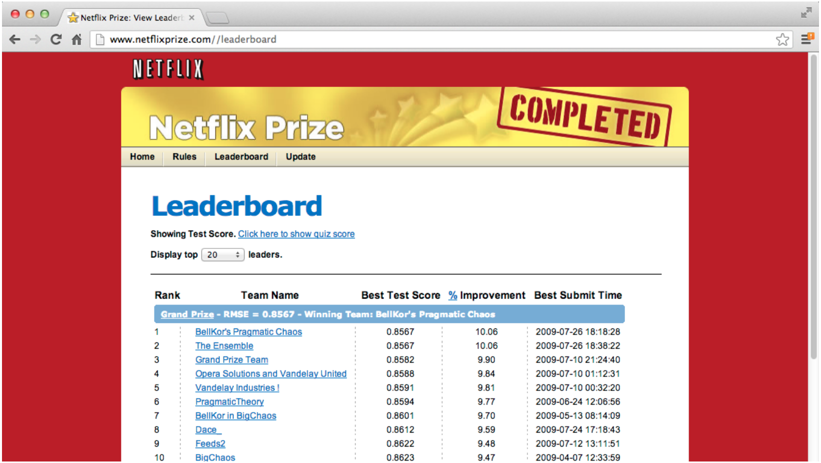
Key idea

To predict player performance use data about player performance


BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN
BASED ON A TRUE STORY
COLUMBIA PICTURES PRESENTS A SCOTT RUDIN/MICHAEL DE LUCA/RACHAEL HOROVITZ PRODUCTION A FILM BY BENNETT MILLER
"MONEYBALL" MUSIC BY MYCHAEL DANNA COSTUME DESIGNER KASIA WALICKA MAIMONE EDITED BY CHRISTOPHER TELLEFSEN, A.C.E. PRODUCTION DESIGNER JESS GONCHOR DIRECTOR OF PHOTOGRAPHY WALLY PFISTER, ASC
EXECUTIVE PRODUCERS SCOTT RUDIN ANDREW KARSCH SIDNEY KIMMEL MARK BAKSHI BASED ON THE BOOK BY MICHAEL LEWIS STORY BY STAN CHERVIN SCREENPLAY BY STEVEN ZAILLIAN AND AARON SORKIN
COLUMBIA PICTURES
SONY make.believe
PRODUCED BY MICHAEL DE LUCA RACHAEL HOROVITZ BRAD PITT
Moneyball-Movie.net
DIRECTED BY BENNETT MILLER
Soundtrack on Madison Gate Records
COMING SOON

Key idea

To predict movie preferences use data about movie preferences


Netflix Prize: View Leaderb
www.netflixprize.com//leaderboard
NETFLIX
Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update

Leaderboard

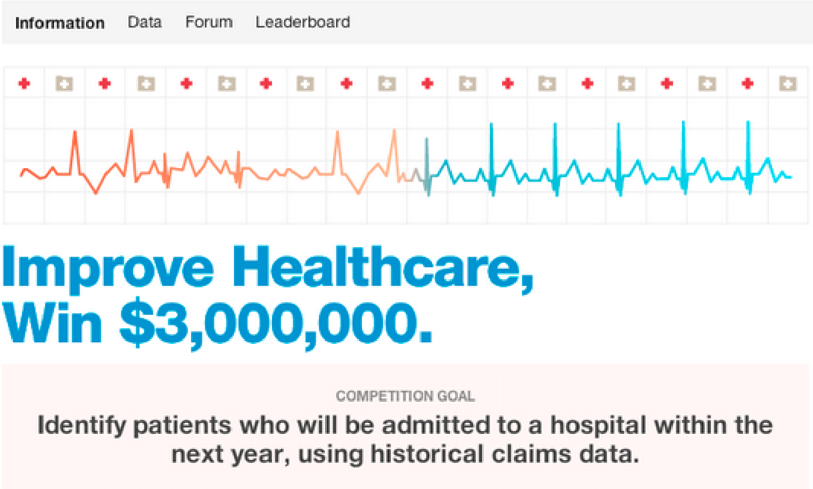
Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59

Key idea

To predict hospitalizations use data about hospitalizations



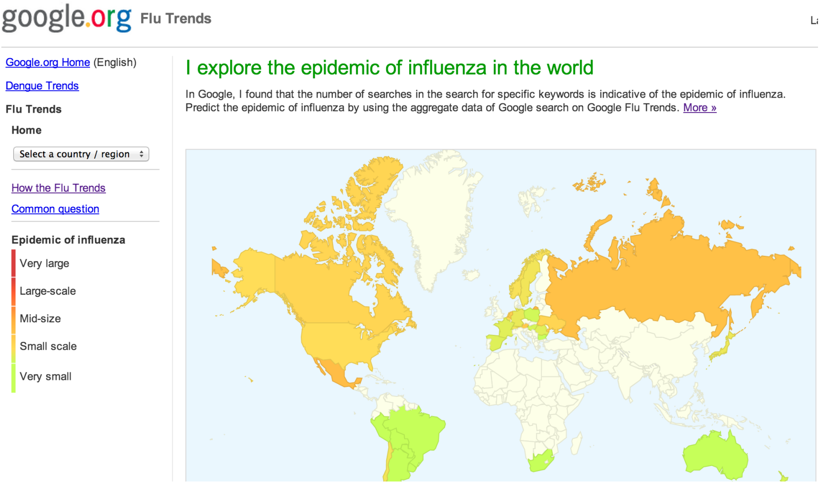
Improve Healthcare,
Win $3,000,000.

COMPETITION GOAL

Identify patients who will be admitted to a hospital within the next year, using historical claims data.

Not a hard rule

To predict flu outbreaks use Google searches



google.org Flu Trends

Google.org Home (English)

Dengue Trends

Flu Trends

Home

Select a country / region

How the Flu Trends

Common question

Epidemic of influenza

- Very large
- Large-scale
- Mid-size
- Small scale
- Very small

I explore the epidemic of influenza in the world

In Google, I found that the number of searches in the search for specific keywords is indicative of the epidemic of influenza. Predict the epidemic of influenza by using the aggregate data of Google search on Google Flu Trends. More »

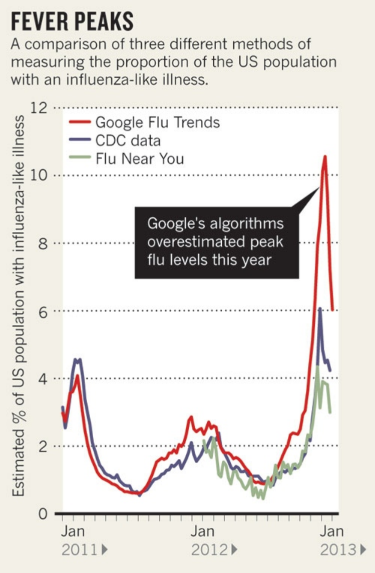
`http://www.google.org/flutrends/`

Looser connection = harder prediction

Oncotype DX® reveals the underlying biology that changes treatment decisions 37% of the time

Uncover the Unexpected℠

Data properties matter

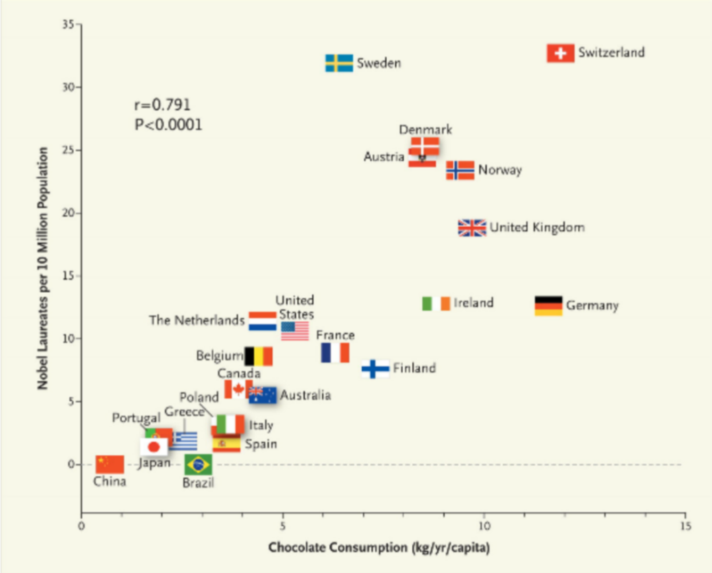


FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.

Google's algorithms overestimated peak flu levels this year

Unrelated data is the most common mistake



http://www.nejm.org/doi/full/10.1056/NEJMon1211064