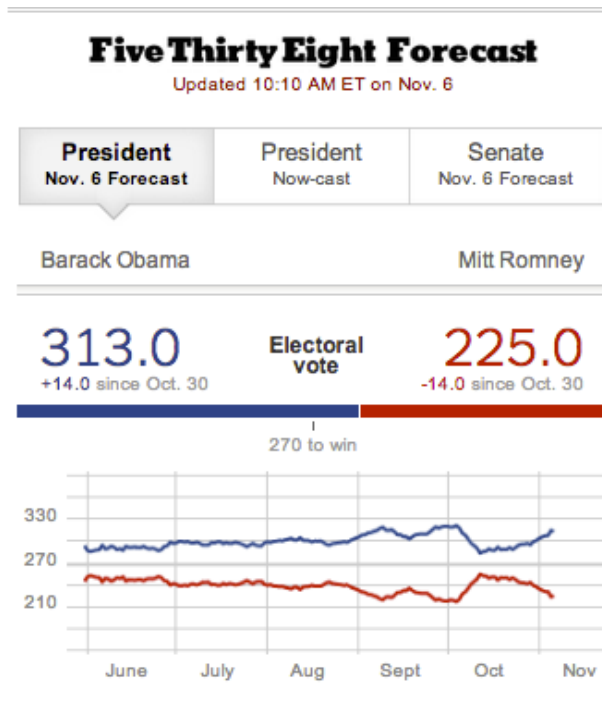




# What data should you use?

Jeffrey Leek  
Johns Hopkins Bloomberg School of Public Health

# A successful predictor



[fivethirtyeight.com](http://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



Month	Index
Jan '12	-22
May '12	-22
Sep '12	-27
Jan '13	-27
May '13	-7
Sep '13	-25

Jan '12 May '12 Sep '12 Jan '13 May '13 Sep '13

Inside Strategic Consulting

<http://www.gallup.com/>

# Weighting the data



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.

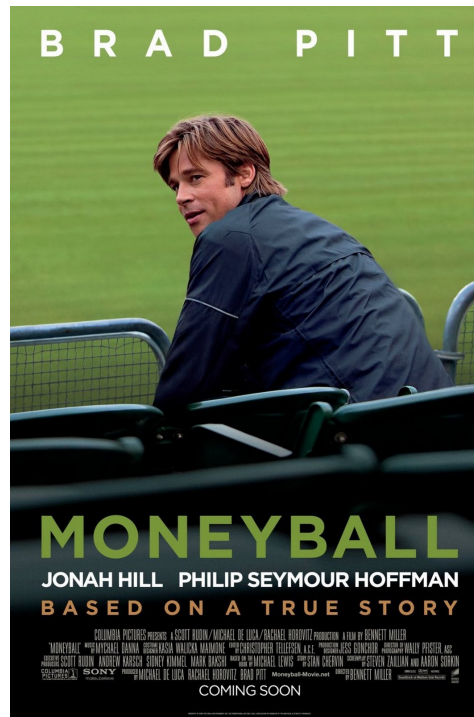
<http://www.fivethirtyeight.com/2010/06/pollster-ratings-v40-methodology.html>

# Key idea

To predict  $X$  use data related to  $X$

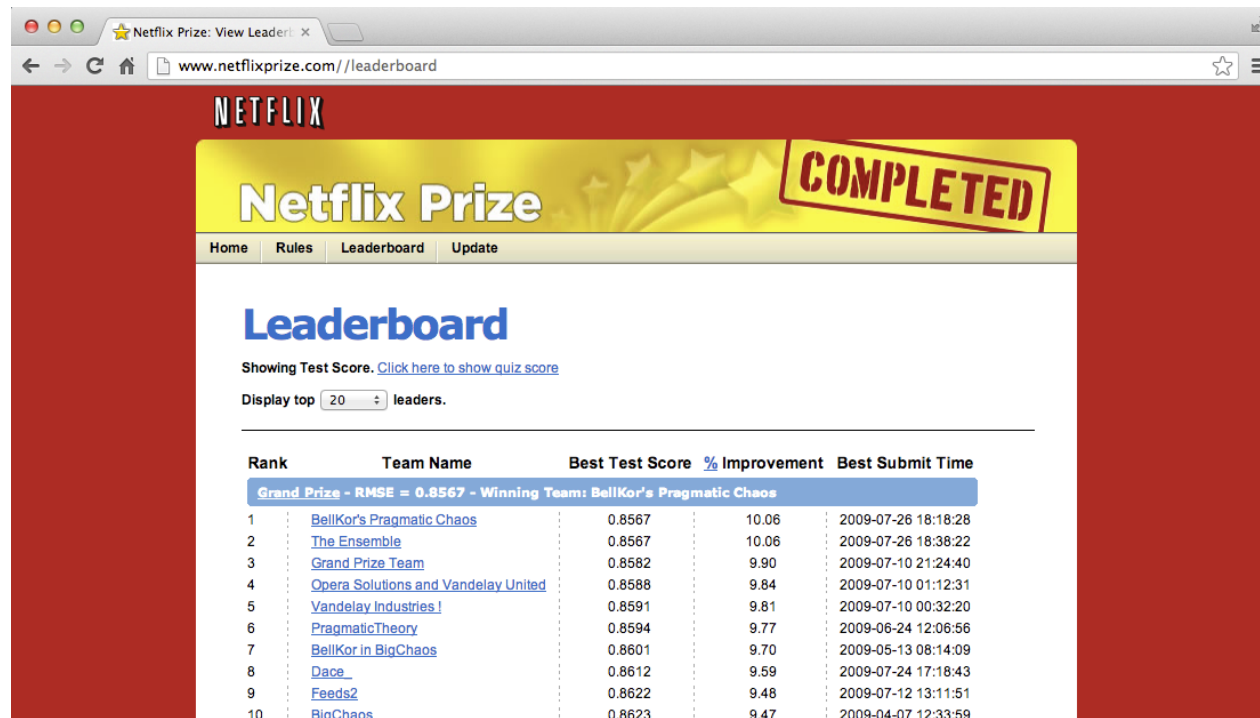
# Key idea

To predict player performance use data about player performance



# Key idea

To predict movie preferences use data about movie preferences

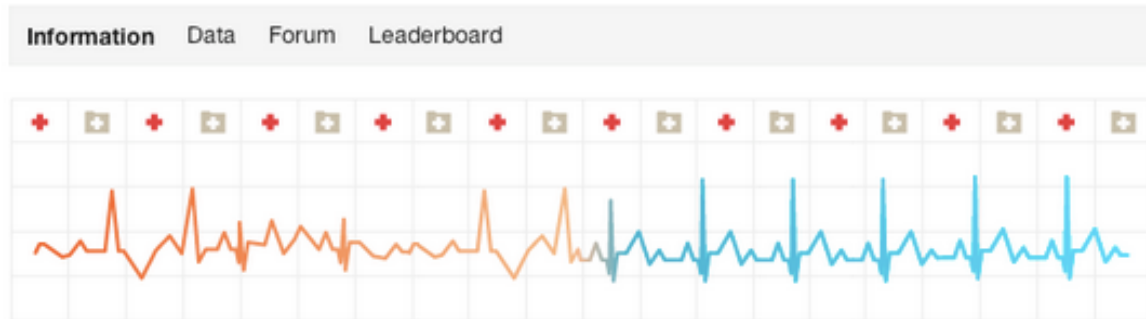


The screenshot shows the Netflix Prize Leaderboard page. At the top, there's a yellow banner with the Netflix logo, 'Netflix Prize' text, and a 'COMPLETED' stamp. Below the banner is a navigation bar with 'Home', 'Rules', 'Leaderboard', and 'Update' links. The main heading is 'Leaderboard'. Below it, a message says 'Showing Test Score. [Click here to show quiz score](#)'. A dropdown menu shows 'Display top 20 leaders'. The main content is a table with 5 columns: Rank, Team Name, Best Test Score, % Improvement, and Best Submit Time. The table lists 10 teams, with the top team being 'BellKor's Pragmatic Chaos'.

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



<http://www.google.org/flutrends/>

# Looser connection = harder prediction

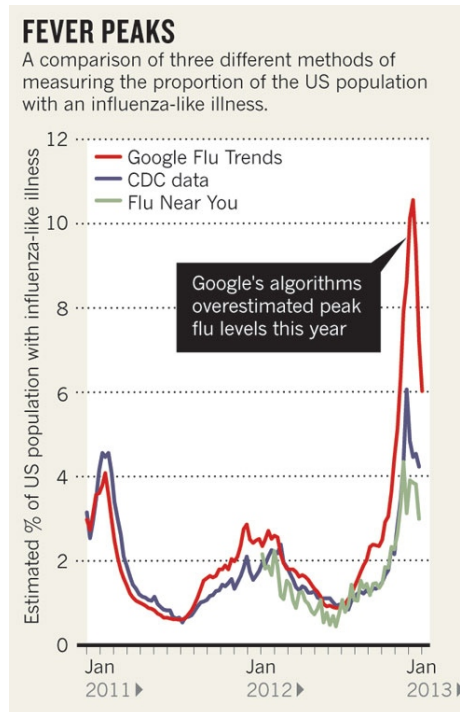
**Oncotype DX® reveals  
the underlying biology that  
changes treatment decisions  
37% of the time**

---

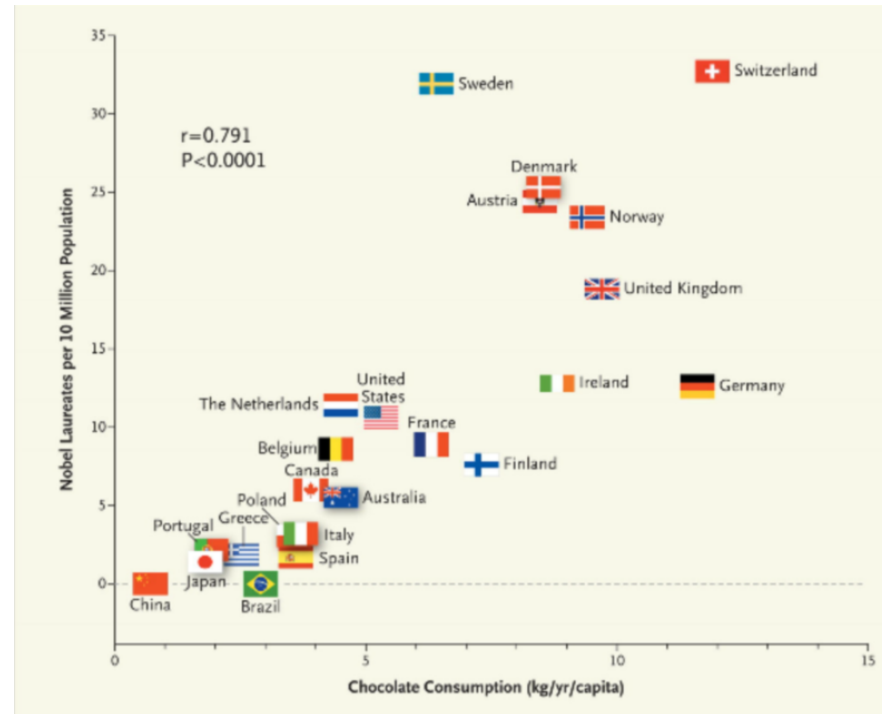
Uncover the Unexpected™



# Data properties matter



# Unrelated data is the most common mistake



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