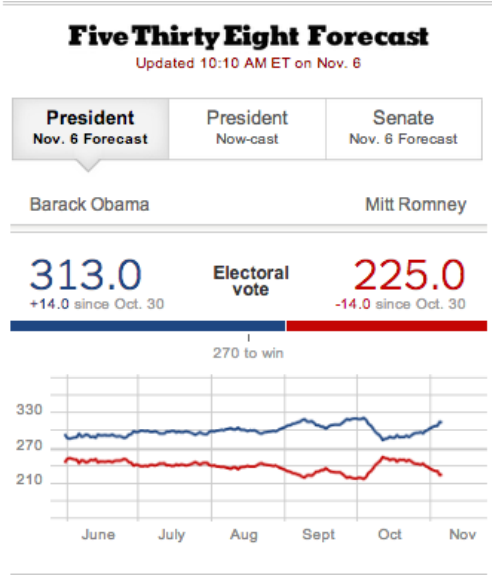
# What data should you use?

**Jeffrey Leek**
**Johns Hopkins Bloomberg School of Public Health**

# A succcessful predictor




this is an example of using the same kind of data for prediction as the result:

"who are going to vote for?"

fivethirtyeight.com

# Polling data


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.

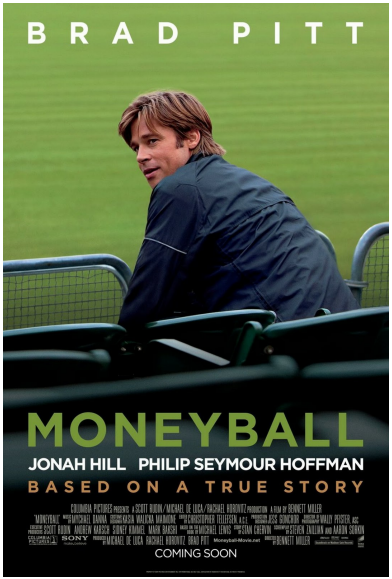
http://www.fivethirtyeight.com/2010/06/pollster-ratings-v40-methodology.html

# Key idea

To predict X use data related to X

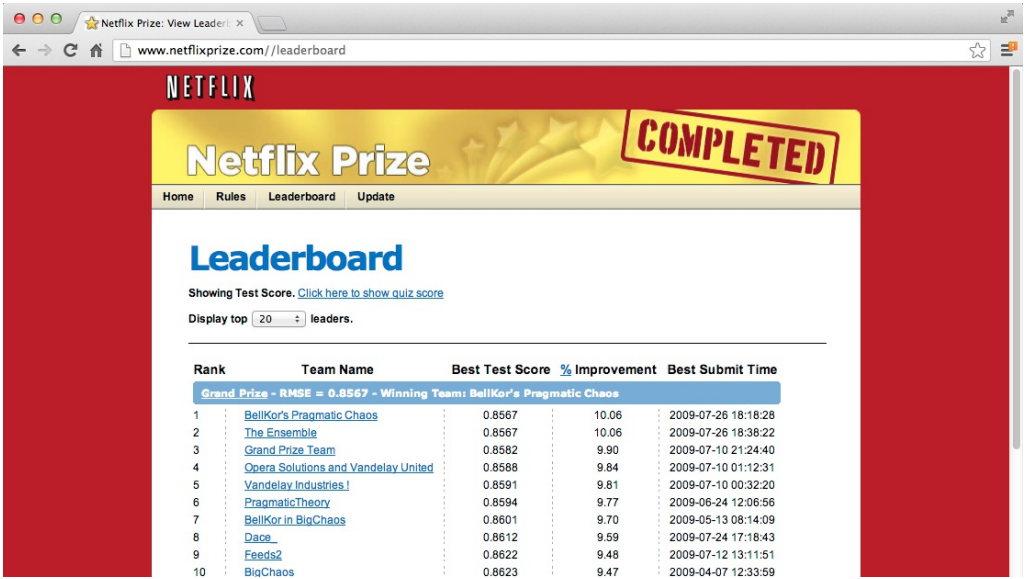
# Key idea

To predict player performance use data about player performance



# Key idea

To predict movie preferences use data about movie preferences



# Key idea

To predict hospitalizations use data about hospitalizations

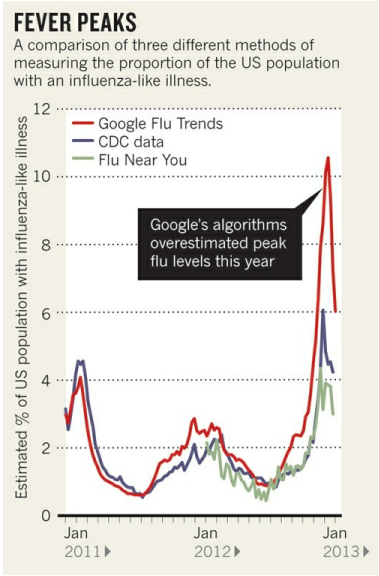
# Not a hard rule

To predict flu outbreaks use Google searches

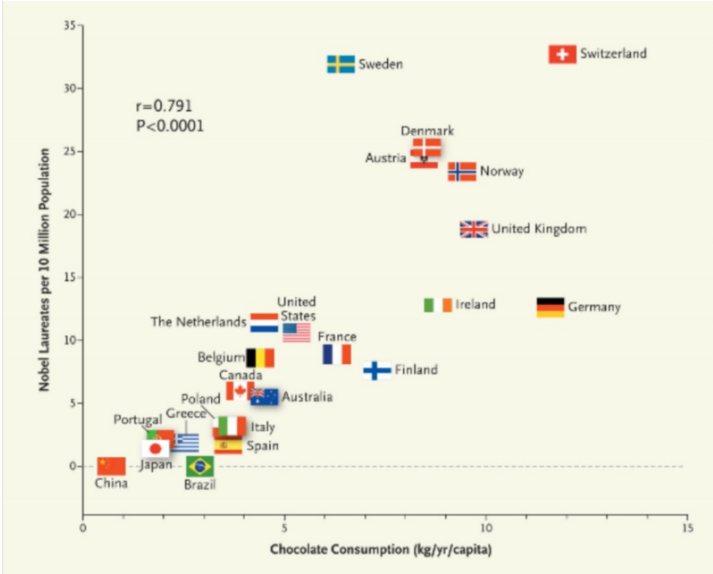

http://www.google.org/flutrends/

# Looser connection = harder prediction

# Data properties matter



# Unrelated data is the most common mistake




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