

---

# Mathematical Challenges in Predictive Analytics

**Joseph R. Zipkin, Ph.D.**

**Forecasting from Big, Noisy Data: Challenges and Techniques**

**2016 SIAM Annual Meeting**

**July 12, 2016**



DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

This material is based upon work funded and supported by the Intelligence Advanced Research Projects Activity (IARPA), in the Office of the Director of National Intelligence (ODNI) under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

---



# The Great Thinkers Agree: Prediction Is Hard



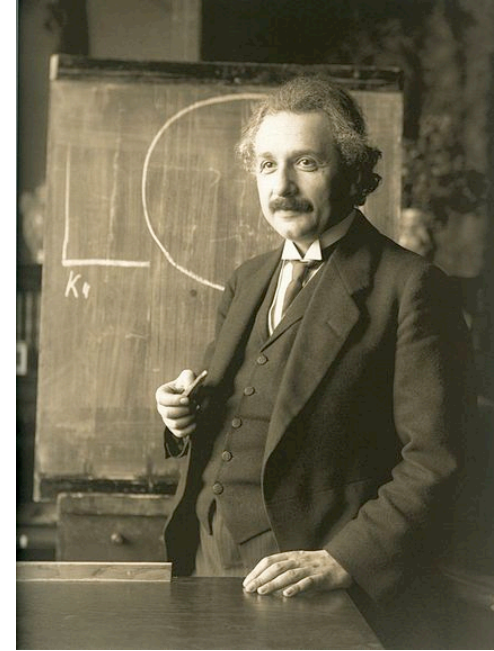
*A fool also is full of words:  
a man cannot tell what shall be;  
and what shall be after him,  
who can tell him?*

**Ecclesiastes**



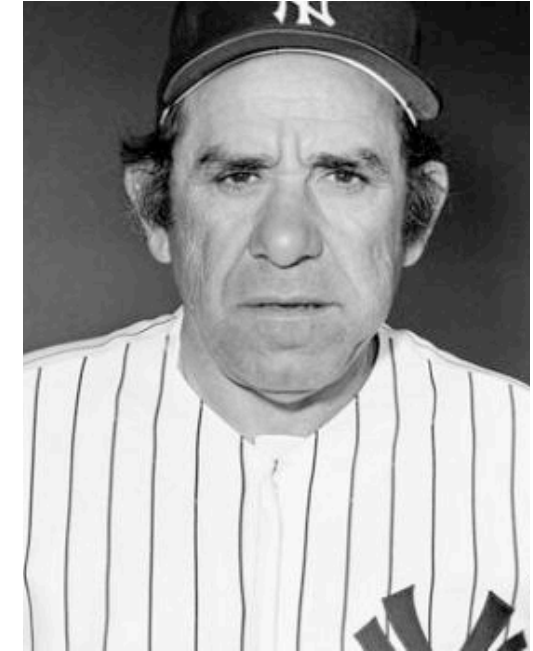
*Our reasons are not prophets  
When oft our fancies are.*

**William Shakespeare**



*When the number of factors coming into play  
in a phenomenological complex is too large,  
scientific method in most cases fails.  
One need only think of the weather, in which  
case the prediction even for a few days  
ahead is impossible.*

**Albert Einstein**



*It's tough to make predictions,  
especially about the future.*

**Yogi Berra\***



# Why Is Prediction Hard?

## Answer 1: The World Is Unpredictable



*An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and **the future just like the past would be present** before its eyes.*

**Pierre-Simon Laplace, 1814**



*It may happen that small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. **Prediction becomes impossible.***

**Henri Poincaré, 1903**



$$\sigma_x \sigma_p \geq \frac{\hbar}{2}$$

**Werner Heisenberg, 1927**



# Why Is Prediction Hard?

## Answer 2: Humans Are Bad at It

Magical thinking

Planning fallacy

Base rate neglect

Conjunction fallacy

Availability heuristic

Choose–reject asymmetry

Gambler's fallacy

Affect heuristic

Disjunction fallacy



# Predictive Analytics Can Help

- **Biases on preceding slide are violations of mathematical laws**
  - Transitive property
  - Probability axioms
  - Bayes's rule
- **Literature going back to Meehl (1954) shows models beat experts**
- **Need to control for biases when building models**
  - Demographics
  - Education
  - Quantifiability

**But predictive analytics comes with its own mathematical challenges**



# Mathematical Challenges in Predictive Analytics



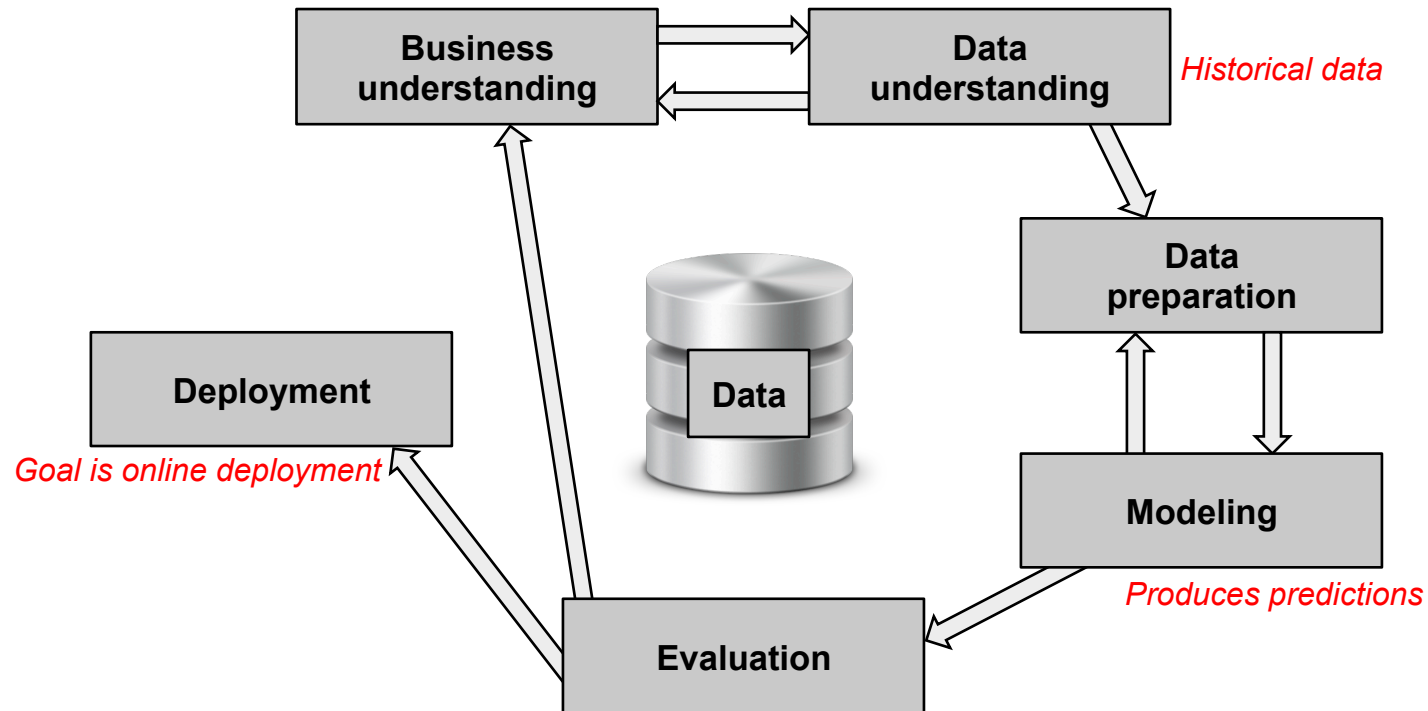
- **Motivation**
- **Predictive analytics**
- **Large, noisy feature data**
- **Evaluating prediction quality**
- **Missing or low-quality ground-truth labels**
- **Conclusion**





# What Is Predictive Analytics?

## *Predictive* The CRISP-DM Model of $\hat{A}$ Analytics

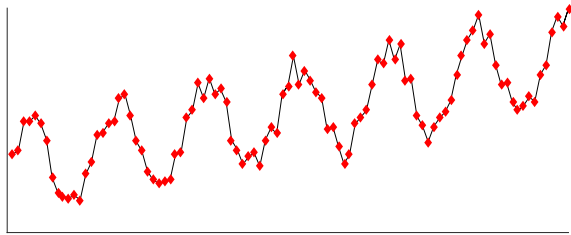


**Predictive analytics is a process, not a suite of models**

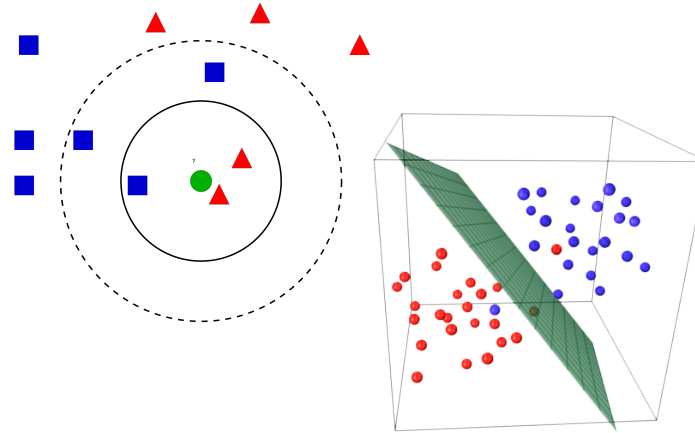


# Models

$$\varphi(B)(x_t - \mu) = \theta(B)\epsilon$$



**Time series analysis**



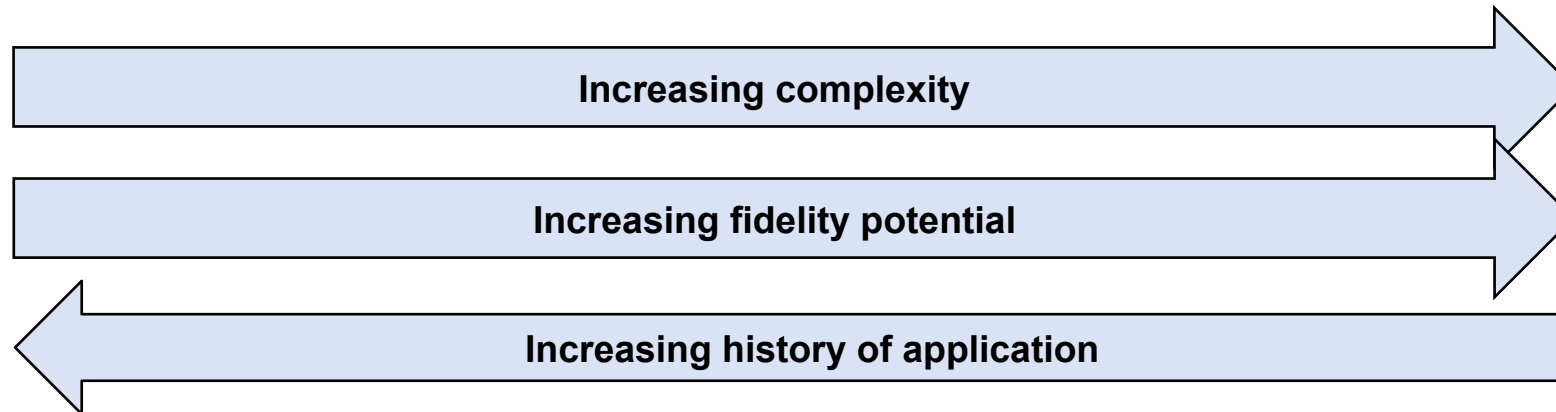
**“Machine learning” algorithms**

$$\frac{\partial A}{\partial t} = \eta \Delta A + d(\kappa) - A + A_0$$

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\nabla \rho - 2\rho \nabla \log A) - d(\kappa)\rho A + d(\kappa)\bar{B}$$

$$\kappa = \arg \min \left\{ \int_{\Omega} d(k)\rho A dx : k \geq 0, \int_{\Omega} k dx = K \right\}$$

**Custom generative models**

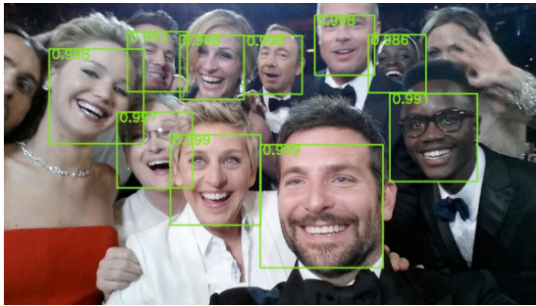






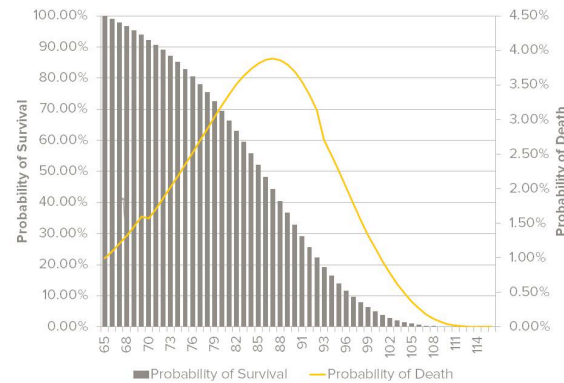
# What Is Predictive Analytics Not?

## Detection



## Risk Analysis

Impact →	1	2	3	4	5
Probability ↓	Negligible	Minor	Moderate	Significant	Severe
(81-100)%	Low Risk	Moderate Risk	High Risk	Extreme Risk	Extreme Risk
(61-80)%	Minimum Risk	Low Risk	Moderate Risk	High Risk	Extreme Risk
(41-60)%	Minimum Risk	Low Risk	Moderate Risk	High Risk	High Risk
(21-40)%	Minimum Risk	Low Risk	Low Risk	Moderate Risk	High Risk
(1-20)%	Minimum Risk	Minimum Risk	Low Risk	Moderate Risk	High Risk




## Expert Prediction





# Mathematical Challenges in Predictive Analytics

- **Motivation**
- **Predictive analytics**
-  • **Large, noisy feature data**
- **Evaluating prediction quality**
- **Missing or low-quality ground-truth labels**
- **Conclusion**



# Suppose You Want to Predict Cyber Attacks



**@FawkesSecurity: lol guys let's attack all hsbc websites at the same time**



## HSBC hit by Anonymous denial-of-service attack

Friday, October 19, 2012 Mohit Kumar

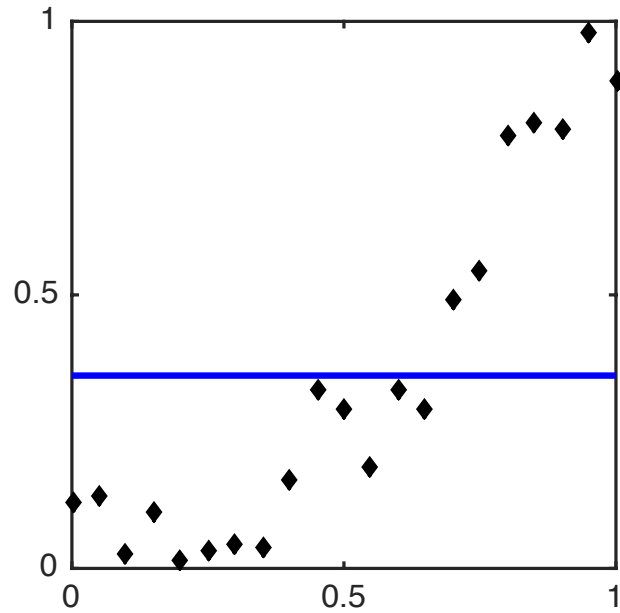
The multinational bank HSBC has blamed a denial of service attack for the downtime of many of its websites worldwide on Thursday night and the Anonymous group has been quick to take credit.

*"Banks are the sole cause of our current worldwide economic problems. They deserve to get hit. RBS, Lloyds TSB and Barclays are next," FawkesSecurity said.*

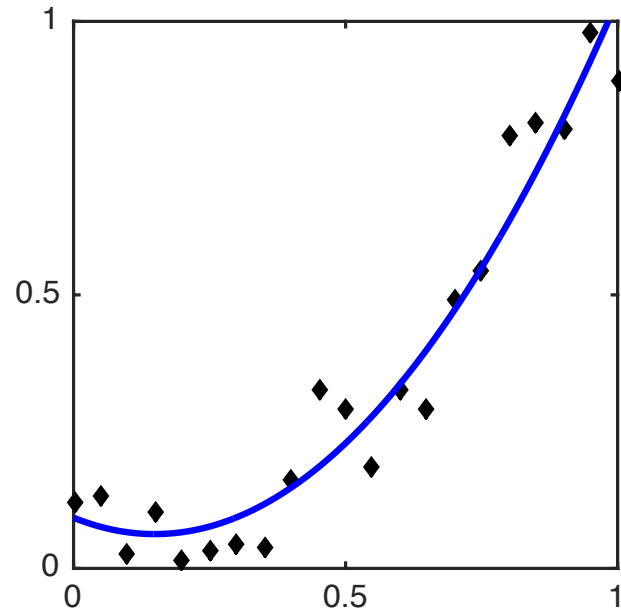
**Big, noisy data**



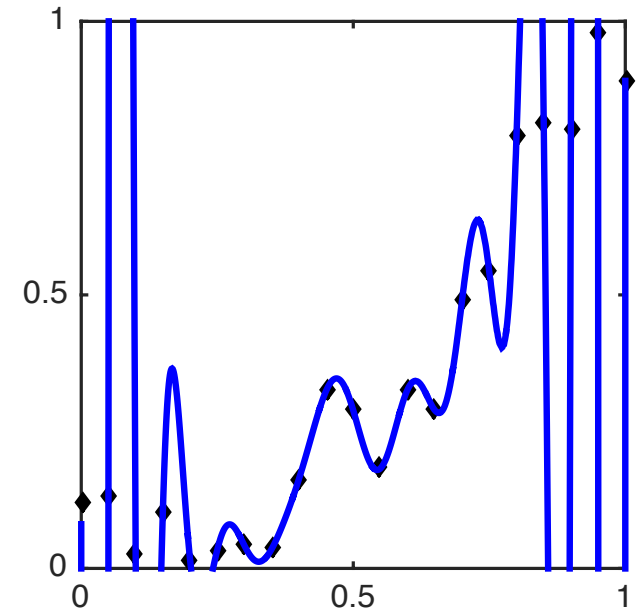
# Overfitting



**underfitted**



**well fitted**



**overfitted**



# Big Data, Big Confidence



I have so much data,  
the answer must be in here somewhere!

**Existence Fallacy**

**Optimality Fallacy**

I have so much data,  
this must be the best way to get the answer!



# Debunking the Existence Fallacy: The Answer Is Not Always in Your Big Data

Journal of Computational Science 2 (2011) 1–8

Contents lists available at ScienceDirect

**Journal of Computational Science**

journal homepage: [www.elsevier.com/locate/jocs](http://www.elsevier.com/locate/jocs)

## Twitter mood predicts the stock market

Johan Bollen<sup>a,\*</sup>, Huina Mao<sup>a,1</sup>, Xiaojun Zeng<sup>b</sup>

<sup>a</sup> School of Informatics and Computing, Indiana University, 919 E. 10th Street, Bloomington, IN 47408, United States  
<sup>b</sup> School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, United Kingdom

---

**ARTICLE INFO**

*Article history:*  
 Received 15 October 2010  
 Received in revised form 2 December 2010  
 Accepted 5 December 2010  
 Available online 2 February 2011

**Keywords:**  
 Social networks  
 Sentiment tracking  
 Stock market  
 Collective mood

**ABSTRACT**

Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error (MAPE) by more than 6%.

© 2011 Elsevier B.V. All rights reserved.

Statistical significance ( $p$ -values) of bivariate Granger-causality correlation between moods and DJIA in period February 28, 2008 to November 3, 2008.

Lag	OF	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	<b>0.085*</b>	0.272	0.952	0.648	0.120	0.848	0.388
2 Days	0.268	<b>0.013**</b>	0.973	0.811	0.369	0.991	0.7061
3 Days	0.436	<b>0.022**</b>	0.981	0.349	0.418	0.991	0.723
4 Days	0.218	<b>0.030**</b>	0.998	0.415	0.475	0.989	0.750
5 Days	0.300	<b>0.036**</b>	0.989	0.544	0.553	0.996	0.173
6 Days	0.446	<b>0.065*</b>	0.996	0.691	0.682	0.994	<b>0.081*</b>
7 Days	0.620	0.157	0.999	0.381	0.713	0.999	0.150

\*  $p < 0.1$ .  
 \*\*  $p < 0.05$ .

## Twitter mood predicts the stock market

[J Bollen](#), [H Mao](#), [X Zeng](#) - *Journal of Computational Science*, 2011 - Elsevier

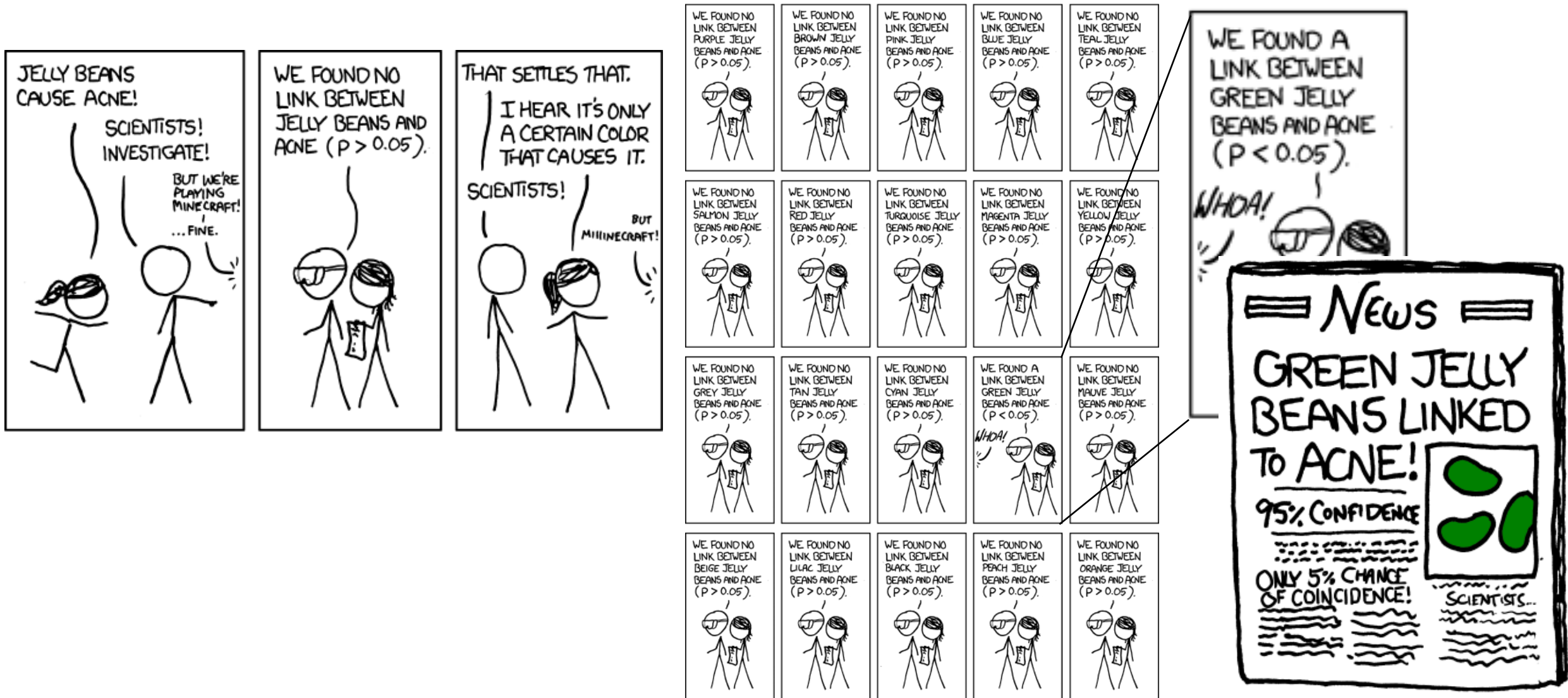
Abstract Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, ie can societies experience mood states that affect their collective decision making? By extension is the ...

[Cited by 1951](#) [Related articles](#) [All 39 versions](#) [Web of Science: 387](#) [Cite](#) [Save](#) [More](#)





# Multiple Comparison Problem







# Debunking the Existence Fallacy: The Answer Is Not Always in Your Big Data

Journal of Computational Science 2 (2011) 1–8

Contents lists available at ScienceDirect

**Journal of Computational Science**

journal homepage: [www.elsevier.com/locate/jocs](http://www.elsevier.com/locate/jocs)

## Twitter mood predicts the stock market

Johan Bollen<sup>a,\*</sup>, Huina Mao<sup>a,1</sup>, Xiaojun Zeng<sup>b</sup>

<sup>a</sup> School of Informatics and Computing, Indiana University, 919 E. 10th Street, Bloomington, IN 47408, United States  
<sup>b</sup> School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, United Kingdom

**ARTICLE INFO**

*Article history:*  
 Received 15 October 2010  
 Received in revised form 2 December 2010  
 Accepted 5 December 2010  
 Available online 2 February 2011

**Keywords:**  
 Social networks  
 Sentiment tracking  
 Stock market  
 Collective mood

**ABSTRACT**

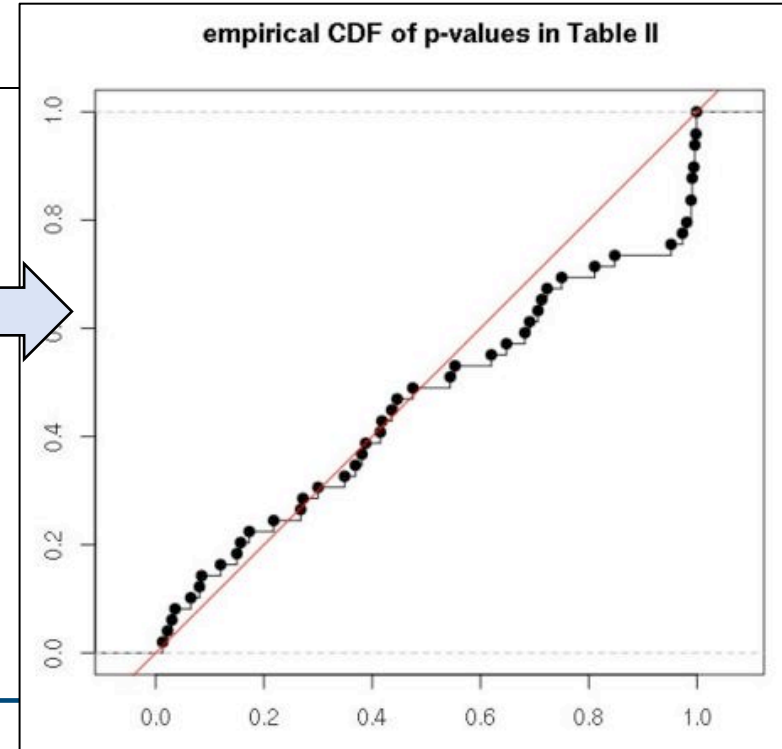
Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error (MAPE) by more than 6%.

© 2011 Elsevier B.V. All rights reserved.

Statistical significance ( $p$ -values) of bivariate Granger-causality correlation between moods and DJIA in period February 28, 2008 to November 3, 2008.

Lag	OF	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	<b>0.085*</b>	0.272	0.952	0.648	0.120	0.848	0.388
2 Days	0.268	<b>0.013**</b>	0.973	0.811	0.369	0.991	0.7061
3 Days	0.436	<b>0.022**</b>	0.981	0.349	0.418	0.991	0.723
4 Days	0.218	<b>0.030**</b>	0.998	0.415	0.475	0.989	0.750
5 Days	0.300	<b>0.036**</b>	0.989	0.544	0.553	0.996	0.173
6 Days	0.446	<b>0.065*</b>	0.996	0.691	0.682	0.994	<b>0.081*</b>
7 Days	0.620	0.157	0.999	0.381	0.713	0.999	0.150

\*  $p < 0.1$ .  
\*\*  $p < 0.05$ .



## Twitter mood predicts the stock market

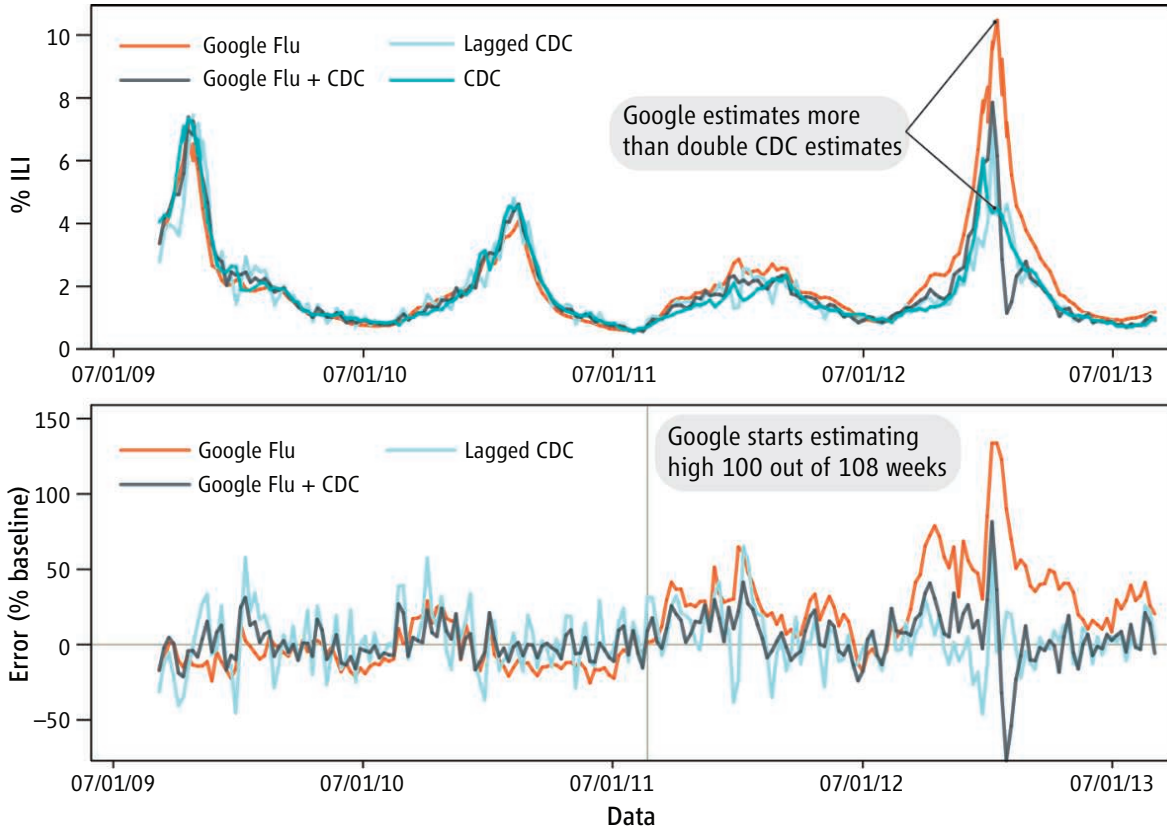
J Bollen, H Mao, X Zeng - Journal of Computational Science, 2011 - Elsevier

Abstract Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, ie can societies experience mood states that affect their collective decision making? By extension is the ...

**Cited by 1951** Related articles All 39 versions Web of Science: 387 Cite Save More



# Debunking the Optimality Fallacy: Your Big Data Is Not Always the Best Tool



- **Several papers purporting to predict flu based on search trends**
- **Google Flu Trends underperformed simple time series models using CDC data (Lazer *et al.* 2014)**
  - **Shut down in August 2015**
- **Correlations with other seasonal phenomena that don't track flu**
  - **High-school basketball**
- **Neglected known correlates**
  - **Lagged CDC data (Lazer *et al.* 2014)**
  - **Sales of over-the-counter cold remedies (Welliver *et al.* 1979)**

**Need to benchmark predictive systems against a plausible baseline**



# Mathematical Challenges in Predictive Analytics

- **Motivation**
- **Predictive analytics**
- **Large, noisy feature data**
- **Evaluating prediction quality**
- **Missing or low-quality ground-truth labels**
- **Conclusion**





# Classical confusion matrix quantities

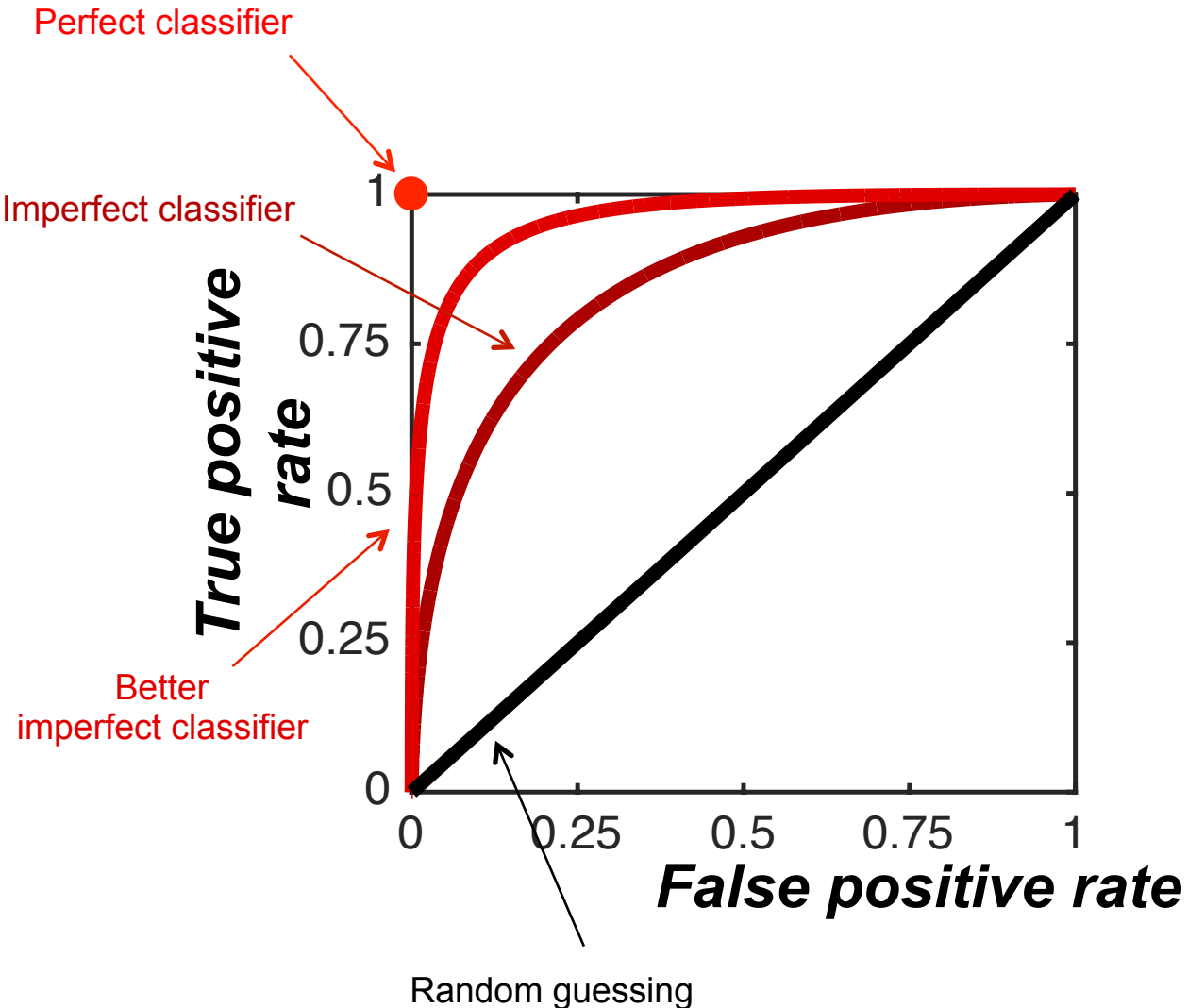
- **Compare output of your system to ground truth**
  - Need ground truth!
- **Precision**
  - $TP/(TP+FP)$
  - Proportion of detections that are true
- **Recall (True Positive Rate)**
  - $TP/(TP+FN)$
  - Proportion of true instances that are detected
- **False Positive Rate**
  - $FP/(FP+TN)$

		System Output	
		True	False
Ground Truth	True	<b><i>TP</i></b>	<b><i>FP</i></b>
	False	<b><i>FN</i></b>	<b><i>TN</i></b>

**Sweep through operating points to reveal tradeoffs among these performance measures**



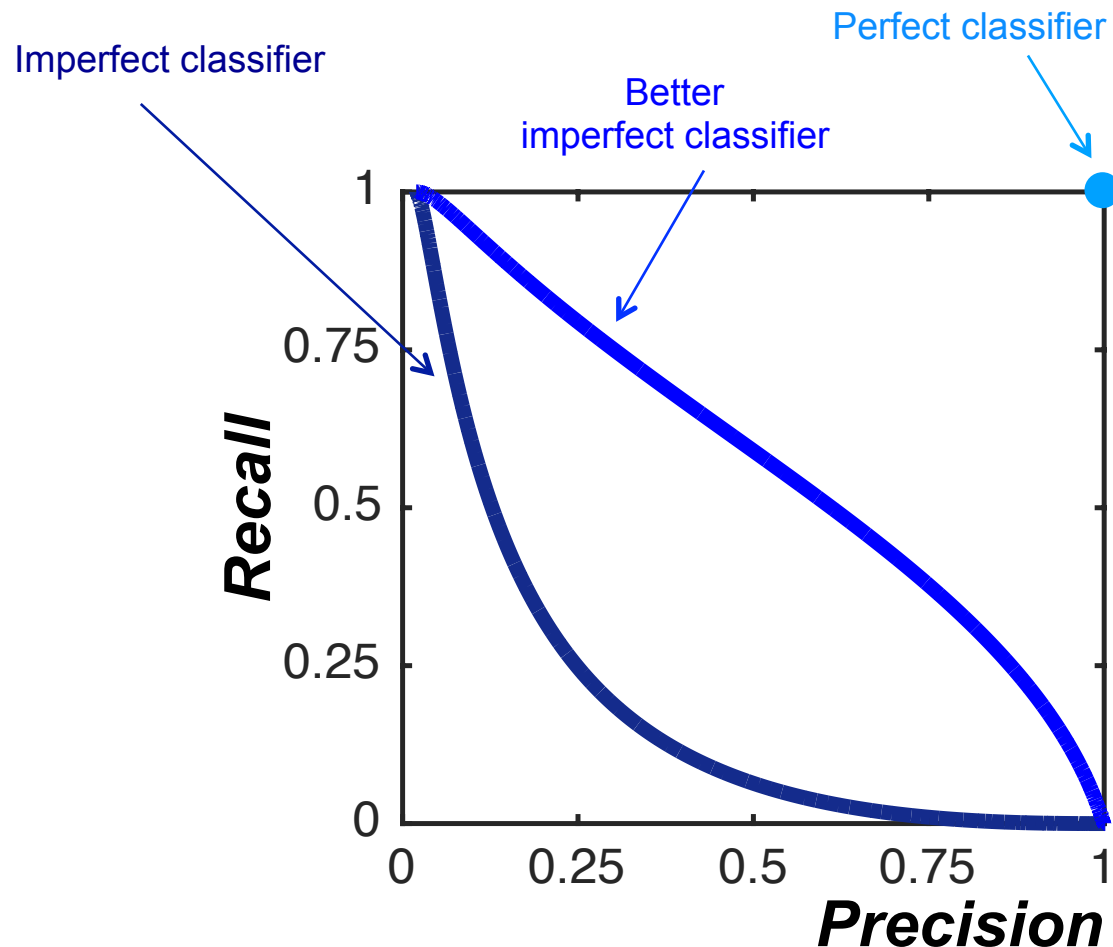
# Receiver Operating Characteristic



- Tradeoff between true positive rate and false positive rate at different operating points
- Built-in baseline
  - Identity line is random guessing
  - Not the relevant baseline when better practices exist
- Class separation between axes
  - No accounting for class imbalance



# Precision–Recall Curve

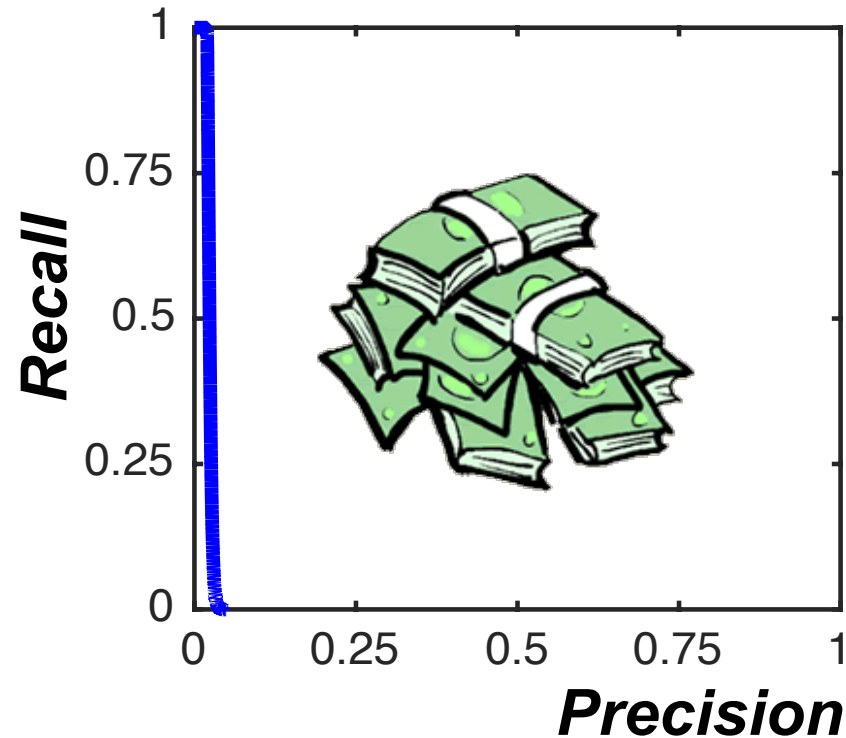


- Tradeoff between recall and precision at different operating points
- Too many alarms being false comes through in precision
  - Primary problem with class imbalance
- No natural baseline
- No need to consider ill-defined true negatives

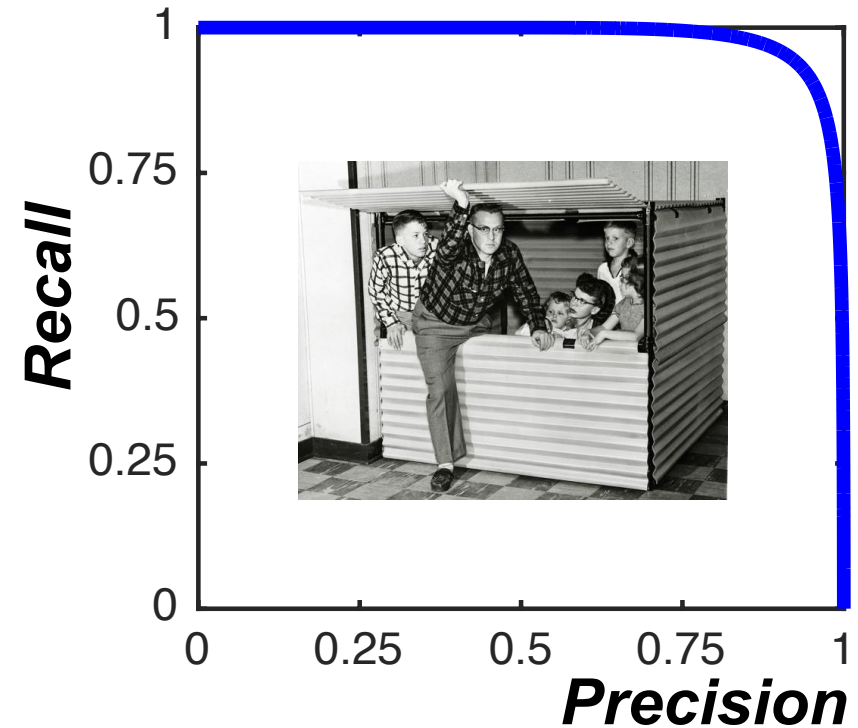


# Need for Baselines

Predicts whether customers will respond to direct-mail campaign\*



Predicts alien invasions



Need to compare parametric curves with baselines





# Probabilistic Forecasts: Brier Score

- **Forecasts should have probabilities attached to them**
  - Prevents confusion between forecasters and decision-makers
  - Enables more informative evaluation
- **Brier score/penalty:  $(p - g)^2$** 
  - $p$  is the probability attached to the forecast
  - $g$  is the ground-truth indicator variable of whether the forecasted event happened
  - Proper scoring function: Reporting true belief minimizes the penalty
- **In practice, decision-makers need to take discrete actions**
  - Binary evaluation isn't going away



# It's All About Utility

- **The ultimate measure of quality is the utility the predictive system delivers to the user**
  - Application-dependent
  - User-dependent (somewhat)
- **Metrics are useful to the extent that they illuminate utility**
- **Non-statistical metrics could be important**
  - Lead time
  - Detection improvement time



# Mathematical Challenges in Predictive Analytics

- **Motivation**
- **Predictive analytics**
- **Large, noisy feature data**
- **Evaluating prediction quality**
- **Missing or low-quality ground-truth labels**
- **Conclusion**





# Sommer and Paxson (2010) on Cyber Data Quality

2010 IEEE Symposium on Security and Privacy

## Outside the Closed World: On Using Machine Learning For Network Intrusion Detection

Robin Sommer

International Computer Science Institute, and  
Lawrence Berkeley National Laboratory

Vern Paxson

International Computer Science Institute, and  
University of California, Berkeley

**Abstract**—In network intrusion detection research, one popular strategy for finding attacks is monitoring a network’s activity for *anomalies*: deviations from profiles of normality previously learned from benign traffic, typically identified using tools borrowed from the machine learning community. However, despite extensive academic research one finds a striking gap in terms of actual deployments of such systems: compared with other intrusion detection approaches, machine learning is rarely employed in operational “real world” settings. We examine the differences between the network intrusion detection problem and other areas where machine learning regularly finds much more success. Our main claim is that the task of finding attacks is fundamentally different from these other applications, making it significantly harder for the intrusion detection community to employ machine learning effectively. We support this claim by identifying challenges particular to network intrusion detection, and provide a set of guidelines meant to strengthen future research on anomaly detection.

deployments in the commercial world. Examples from other domains include product recommendations systems such as used by Amazon [3] and Netflix [4]; optical character recognition systems (e.g., [5], [6]); natural language translation [7]; and also spam detection, as an example closer to home [8].

In this paper we set out to examine the differences between the intrusion detection domain and other areas where machine learning is used with more success. Our main claim is that the task of finding attacks is fundamentally different from other applications, making it significantly harder for the intrusion detection community to employ machine learning effectively. We believe that a significant part of the problem already originates in the premise, found in virtually any relevant textbook, that anomaly detection is suitable for finding *novel* attacks; we argue that this premise

**“[D]evising sound evaluation schemes is not easy, and in fact turns out to be *more difficult than building the detector itself* ... Arguably the most significant challenge an evaluation faces is the lack of appropriate public datasets for assessing anomaly detection systems. ... In our domain, however, we often have neither standardized test sets, nor any appropriate, readily available data.”**

**The situation has not improved six years later**



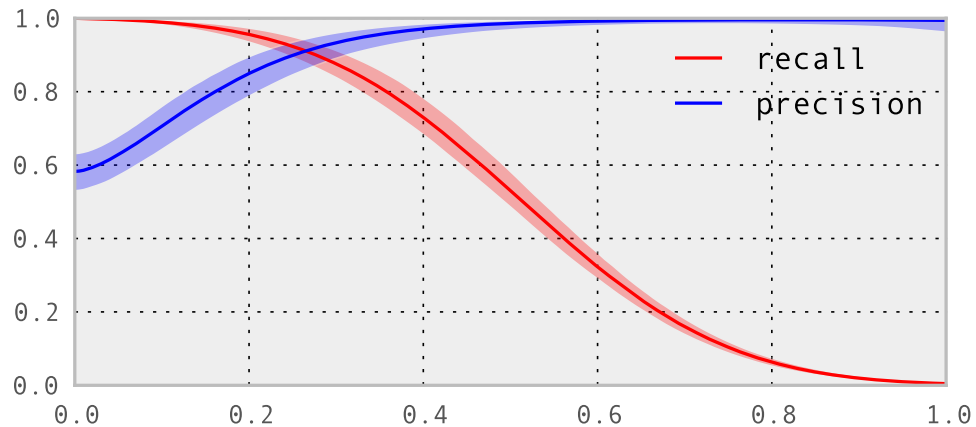
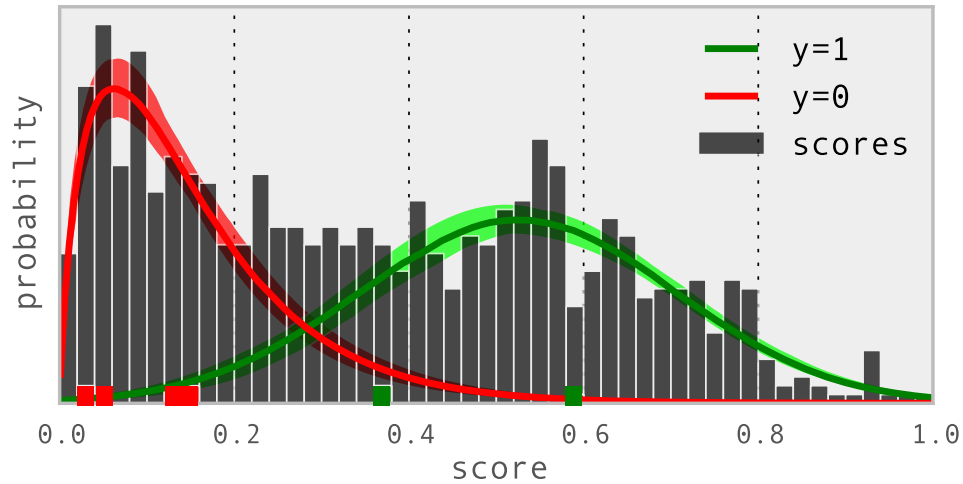
# What to Do?

**If cyber data won't change to fit machine learning...**

**... change machine learning to fit cyber data**



# Estimating Classifier Performance from Few Samples



## Welinder, Welling & Perona (2013)

- Two classes of objects characterized by a *score*
- Samples are plentiful but labeling is expensive
  - “Semisupervised” learning
- Can estimate performance curves with confidence intervals inferred from estimates of the distribution parameters

## Challenges for applying to cyber security

- Class imbalance
- Sampling only in the tail
- Poorly specified distributions



# Mathematical Challenges in Predictive Analytics

- **Motivation**
- **Predictive analytics**
- **Large, noisy feature data**
- **Evaluating prediction quality**
- **Missing or low-quality ground-truth labels**
- **Conclusion**







# Conclusion

---

- **Predictive analytics has the potential to overcome human bias in forecasting**
- **Big, noisy data multiplies the human tendency to find spurious patterns**
- **Evaluation must focus on utility for the user**
- **Plausible, meaningful baselines are essential**
- **New methods are needed for poor-quality ground truth**
- **Cyber security is an exciting application area for predictive analytics**



# Q&A





# Lineup

---

- **Amy Sliva, Charles River Analytics**  
**Predicting Events Using Diverse Ensemble Models**
- **Alexander Memory, Leidos, Inc.**  
**Probabilistic Forecasting in the Presence of Noisy and Conflicting Evidence**
- **Rebecca Cathey, BAE Systems**  
**Relational Object Analysis Drives Multi-Model Attack Prediction (ROADMAP)**



# Backup Slides

---



# References

- J Bollen, H Mao, X Zeng (2011) Twitter mood predicts the stock market. *Journal of Computational Science* 2: 1–8.
- T Gilovich, DW Griffin, D Kahneman, eds. (2002) *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.
- D Lazer *et al.* (2014) The parable of Google Flu: traps in big data analysis. *Science* 343: 1203–1205.
- PE Lehner (2015) Estimating the accuracy of automated classification systems using only expert ratings that are less accurate than the system. *Journal of Modern Applied Statistical Methods* 14(1): 122–151.
- PE Meehl (1954) *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press.
- C Shearer (2000) The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing* 5(4): 13–22.
- E Siegel (2013) *Predictive Analytics*. Wiley.
- PE Tetlock (2006) *Expert Political Judgment*. Princeton University Press.
- PE Tetlock, D Gardner (2015) *Superforecasting: The Art and Science of Prediction*. Crown Publishers.
- P Welinder, M Welling, P Perona (2013) A lazy man’s approach to benchmarking: semisupervised classifier evaluation and recalibration. *CVPR 2013*: 3262–3269.
- RC Welliver (1979) Sales of nonprescription cold remedies: a unique method of influenza surveillance. *Pediatric Research* 13: 1015–1017.



# Intellectual Property Statement

---

© 2016 Massachusetts Institute of Technology.

**Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.**