Predictive Analytics
Time Series

Philip A. Schrodt

Parus Analytical Systems
schrodt735@gmail.com

Odum Institute "Data Matters" Workshop
University of North Carolina
25 June 2014

Approach of the lectures

- Breadth, not depth!—this is more of a "bird's eye view"
  - (But not—I repeat, **not!**—a "god's eye view"!)
- A guide to vocabulary[ies], approaches and what you need to know
  - you can then follow up on all of this material in detail. If I can look it up, you can look it up
- Emphasis on practical applications: Some of the slides are recycled from presentations I've given in the U.S. policy community
  - This is a feature, not a bug
- This is the departure lounge, not the baggage claim
- All of the slides are available at

  `http://eventdata.parusanalytics.com/presentations.html`

The Debate

▶ ARGUMENT PRINT | TEXT SIZE | EMAIL | SINGLE PAGE

Why the World Can't Have a Nate Silver

The quants are riding high after Team Data crushed Team Gut in the U.S. election forecasts. But predicting the Electoral College vote is child's play next to some of these hard targets.

BY JAY ULFELDER | NOVEMBER 8, 2012

Vs.

▶ ARGUMENT PRINT | TEXT SIZE | EMAIL | SINGLE PAGE

Predicting the Future Is Easier Than It Looks

Nate Silver was just the beginning. Some of the same statistical techniques used by America's forecaster-in-chief are about to revolutionize world politics.

BY MICHAEL D. WARD , NILS METTERNICH | NOVEMBER 16, 2012

Two approaches that did not work well in the past

Two approaches that did not work well in the past

Qualitative

Two approaches that did not work well in the past

Qualitative

Quantitative

Problems with qualitative approaches

Tetlock: Experts typically do about as well as a "dart-throwing chimp"

Problems with qualitative approaches

Tetlock: Experts typically do about as well as a "dart-throwing chimp"

Except for television pundits, who do even worse. Ask President Romney.
The media want things to be dramatic. "We're all going to die! Details follow *American Idol*"

Problems with qualitative approaches

Tetlock: Experts typically do about as well as a "dart-throwing chimp"

Except for television pundits, who do even worse. Ask President Romney.
The media want things to be dramatic. "We're all going to die! Details follow *American Idol*"

Qualitative *theory* isn't much better:
Remember the hegemonic US seizure of undefended Canadian and Mexican oil fields in response to the 1973 OPEC oil embargo?

SMEs and the "narrative fallacy"

SME = "subject matter expert"

Hegel: the owl of Minerva flies only at dusk

Taleb (*Black Swan*): seeking out narratives is an almost unavoidable cognitive function and it generates a dopamine hit

This is your brain on narratives

IARPA "Anticipating Critical Events" (ACE) Project

- Five year project sponsored by IARPA: motivation is to provide a large number of systematically specified and scored probability estimates to get around the rare event problem
- Utilizes teams of volunteers, mostly non-expert
- Forecast horizon: 1 to 18 months (vs 3 to 10 years in original Tetlock research)
- Metric: Beier scores over time, with the possibility of using ensemble methods
- Consistent, rigorous and "ungameable" resolution criteria
- Five teams initially; only one—Tetlock's "Good Judgment Project"—achieved the goal and remained active after two years
- Currently also experimenting with prediction markets

IARPA ACE Objectives

- is it possible for human forecasters working in teams to exceed the accuracy of "dart throwing chimp"
- An "elitist search" for "super-forecasters" who do disproportionately well
- if this was achieved, was it possible to train individuals to do this?

Categories of ACE Questions

- Leadership Turnover and Elections in Stable Democracies
- Leadership Turnover and Social Change in Authoritarian Regimes
- Economic and Diplomatic Decisions by International Organizations
- Negotiation Processes
- Macro-economic Indicators and Financial Markets
- Military Actions, Casualty Counts, and Refugee Flows
- Legal Proceedings Within State Boundaries

Scoring

f_c: probability assigned to the event which occurs.

QSR (or Brier rule) $= 2 \times f_c - [{f_c}^2 + (1 - f_c)^2]$, accuracy ranges from -1 to +1.

LSR $= ln(f_c)$, accuracy ranges from $-\infty$ to 0.

SSR $= f_c/[{f_c}^2 + (1 - f_c)^2]^{\frac{1}{2}}$, accuracy ranges from 0 to 1.

All of these are assessed over time—that is, early correct predictions are rewarded—and the three metrics produce similar results, so the Brier score is mostly emphasized now.

Characteristics of good forecasters

High scores on the following measures

- fluid intelligence (tapped by tests of rapid pattern recognition (Raven's Progressive Matrices)
- tests of numeracy (Cokely et al., 2012; Peters et al., 2006)
- tests of cognitive impulse control (Cognitive Reflection Test; Frederick, 2005),
- measures of crystallized intelligence (specifically, geopolitical knowledge)
- measures of cognitive styles (test designed to measure "actively open-minded thinking" (Baron, 2006) and "need for cognition" (Cacioppo et al. 1984)).

Super-forecasters

Method: Assign top 2% of forecasters in each year to elite teams of super-forecasters

Result: Simple unweighted-average of the forecasts made by a group of 60 super-forecasters in year two handily surpassed (70%) the Brier score goals that the research sponsors set for the fourth year (50%)

Super-forecasters

- showed virtually no regression-to-the-mean in the subsequent year of the tournament (top 3% and 4% did)
- had better scores on both of the accuracy indicators derivable from Brier scores
- had better calibration (neither over- nor under-confident)
- had better discrimination (assigning much higher probabilities than to things that happened than to things that didn?t).

Other results

- Fuzzy evaluation—allowing for "near misses" due to chance events like insane fishing boat captains—makes the super-forecasters look even better
- Training individuals (randomly assigned to treatment groups) in probabilistic reasoning improves performance
- Ensemble methods such as weighting by past performance and "extremizing" forecasts (changing 0.7 to 0.9) appears to improve the predictions compared to individual forecasts, though the robustness of this is still unclear
- No teams were able to produce an average Brier score below 0.12: this roughly corresponds to an average distance between the estimated probability and the 0/1 occurrence of the event of around 0.25
  - That is, an accuracy of around 75%: sound familiar?

Problems with quantitative approaches

Ward, Greenhill and Bakke (2010): Models based on significance tests don't predict well because that is not what a significance test is supposed to do.

Gill, Jeff. 1999. The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly* 52:3, 647-674.

Problems with quantitative approaches

Ward, Greenhill and Bakke (2010): Models based on significance tests don't predict well because that is not what a significance test is supposed to do.

Gill, Jeff. 1999. The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly* 52:3, 647-674.

The norm in political science has been to do full-sample evaluation, whereas the norm in machine-learning has been split-sample, which is usually more robust and is certainly more credible

Prediction vs frequentist significance tests

- Significance becomes irrelevant in really large data sets: true correlations are almost never zero
- Emphasis is on finding reproducible patterns, but in any number of different frameworks
- Testing is almost universally out-of-sample
- Some machine learning methods are explicitly probabilistic—though usually Bayesian—others are not
- In "diffuse models" such as VAR, BMA, neural networks, random forests, and HMM/CRF, values of individual coefficients are usually of little interest because there are so many of them and they are affected by collinearity

Two very influential articles ca. 2000 - 1

Collier, Paul and Anke Hoeffer, 2004. Greed and grievance in civil war, *Oxford Economic Papers* 56(4): 563-595.

- Emphasize on structural opportunity for gaining recruits such as high levels of unemployment and poverty and ethnic diasporas willing to provide financial support
- De-emphasis on specific political grievances
- "Greed rather than grievance"

Two very influential articles ca. 2000 - 2

Fearon, James D. and David D. Laitin, 2003. Ethnicity, Insurgency, and Civil War, *American Political Science Review* 97(1):75-90.

- focus on weakness of state institutions
- structural aspects can favor insurgency by reducing costs of mobilization: mountainous terrain, large populations, political instability, the newness of the state, and low levels of economic development
- Democratization is not significant
- GDP/capita is negative and significant

Ward, Bakke, Greenhill 2010

Problem with both models: pattern of significant variables does not result in successful forecasts

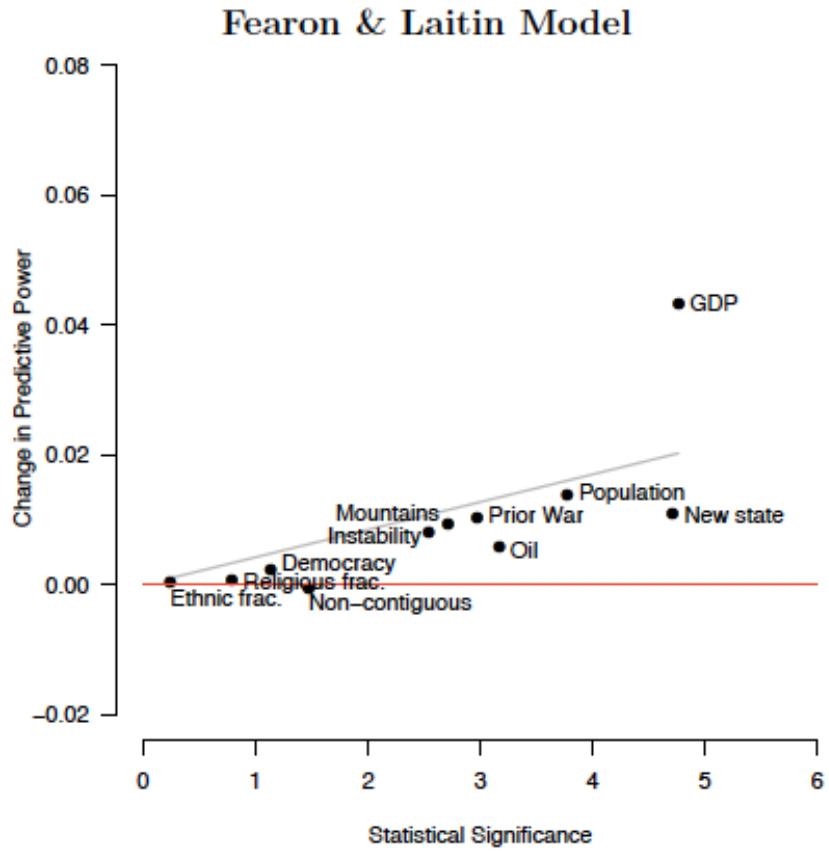
Table III: Number of Correctly Predicted Onsets and False Positives at Varying Cut-Points

	Fearon & Laitin Model		**Collier & Hoeffler Model**	
Threshold	Correctly Predicted	False Positives	Correctly Predicted	False Positives
0.5	0/107	0	3/46	5
0.3	1/107	3	10/46	20
0.1	15/107	66	34/46	110

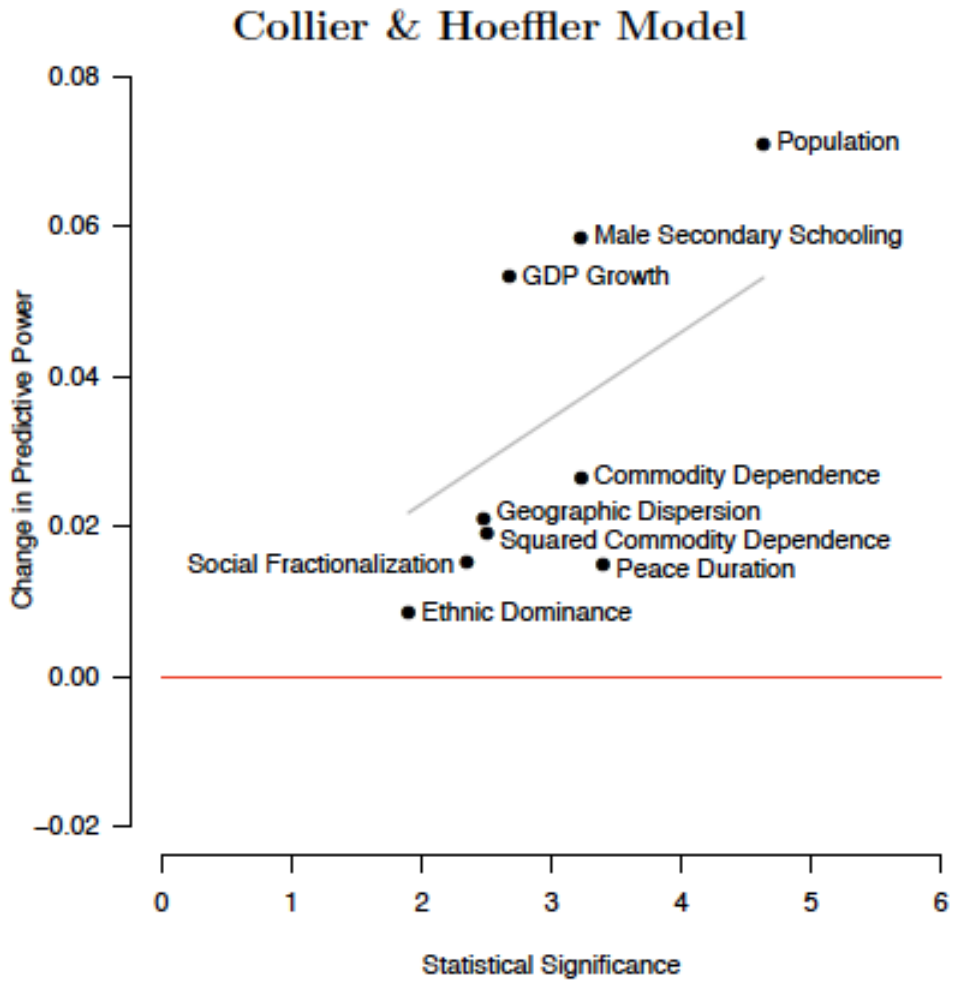
Source: Ward, Bakke, Greenhill 2010. The Perils of Policy by P-Value: Predicting Civil Conflicts. *Journal of Peace Research*

Ward, Bakke, Greenhill 2010: Prediction vs. significance

Figure 2. Comparison of Predictive Power and Statistical Significance



Ward, Bakke, Greenhill 2010: Prediction vs. significance



Role of prediction for logical positivists

Hemple: "Explanation" in the absence of prediction is "prescientific"

- Critical case: astrology vs astronomy
- More generally, mythological accounts provide "explanation" [Quine]

Prediction was simply assumed to be a defining characteristic of a good theory until relatively recently

Arguably, no philosopher of science prior to the mid-20th century would find the frequentist-based "explanation" emphasized in contemporary political science even remotely justified

- Leaving aside that frequentism is logically inconsistent and has been characterized in Meehl (1978) as "a terrible mistake, basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology"
- Hey, dude, tell us what you *really* think

Explanation, continued

Philosophers of science have long suspected that it is possible to have a seemingly sound explanation of a phenomenon that confers no predictive leverage over the phenomenon (Nagel, 1961; Toulmin, 1961). For instance, plate tectonics theory is the received explanation for earthquakes, but it confers no ability to generate accurate predictions about when earthquakes will occur. Conversely, it is possible to have remarkable predictive accuracy that rests on a deeply flawed framework. Ancient astronomers generated predictively powerful celestial charts even though they didn't have the faintest idea what planets or stars were.

. . .

How patient should we be with low-predictive-accuracy theories? When should we tune out the theorists and go with algorithms that no more understand world politics than ancient astronomers understood celestial motion? We have no off-the-shelf answer, but we resonate to Lakatos's (1970) rule for distinguishing degenerative from progressive research programs:

Additional issues in explanation vs. theory

Hume: the problem of induction

- Farmer's cat vs. farmer's turkey

Friedman: unreasonable assumptions are justified provided the predictions are accurate

- Justification for rational choice models
- Issue: the "provided predictions are accurate" part tends to be forgotten, and is far too often replaced with "provided I think the assumptions are elegant and/or make my life easier"

Success without theory: Gothic cathedrals

Note that these issues affect *observational* studies but not *experimental* studies, which is why experiments are used whenever possible.

Kahneman et al: people are really bad at statistical reasoning

- Everyone, including statisticians unless they focus very hard
- Example: managed mutual funds, which both theory and evidence indicate cannot work
- Example: opposition to "evidence based medicine" in the US, with a preference for clinical intuition even when this has been demonstrated to be less effective
- Probabilitistic weather forecasts seem to be the one major exception: rain likelihood, hurricane tracks

The Necessity of Prediction in Policy

Feedforward: policy choices must be made in the present for outcomes which may not occur for many years

Planning Times: even responses to current conditions may require lead times of weeks or months

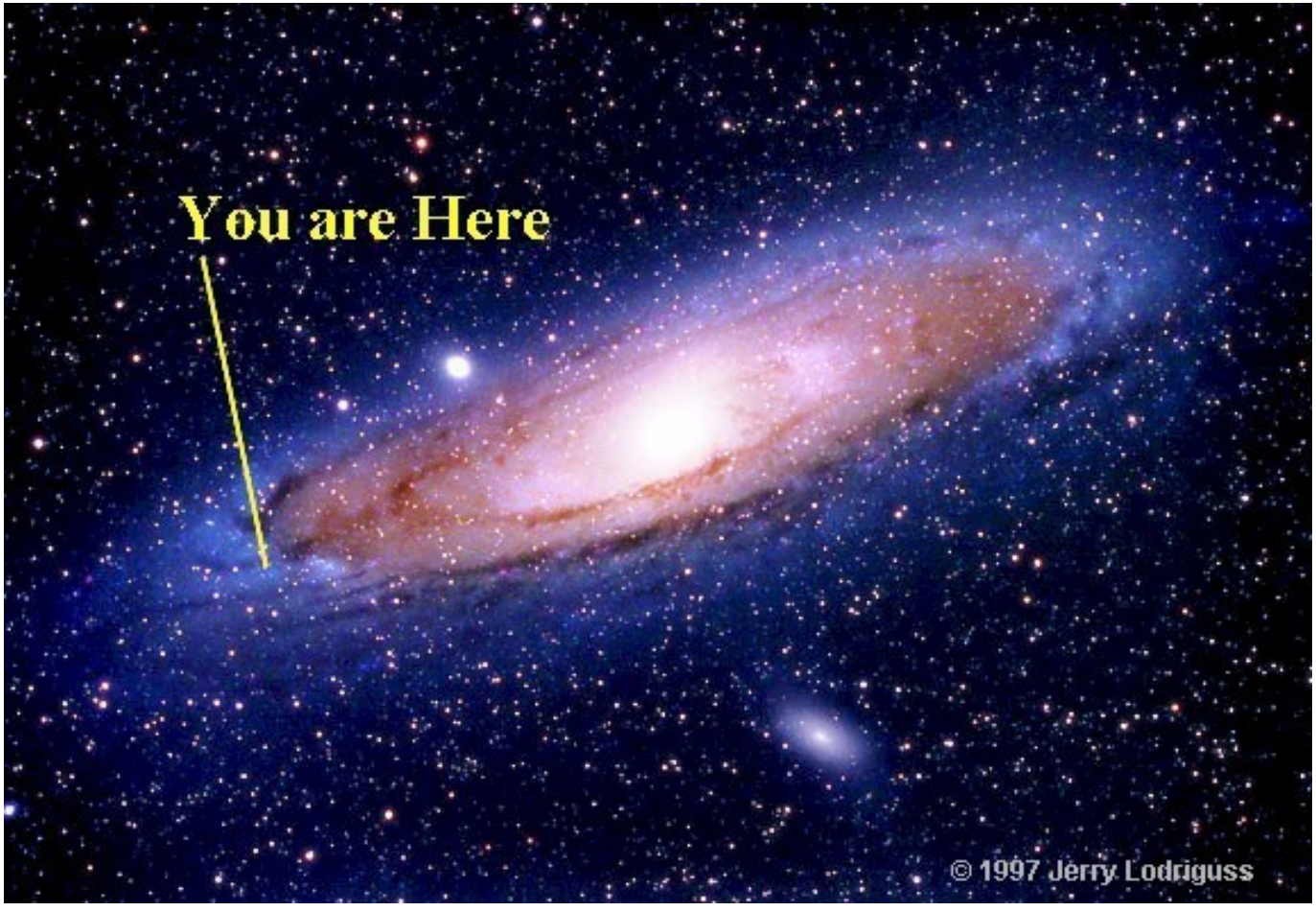
[More on this tomorrow]

The Forecaster's Quartet

- Nassem Nicholas Taleb. *The Black Swan*
  (most entertaining/obnoxious)
- Daniel Kahneman. *Thinking Fast and Slow*
  (30 years of research which won Nobel Prize)
- Philip Tetlock. *Expert Political Judgment*
  (most directly relevant)
- Nate Silver. *The Signal and the Noise*
  (high level of credibility after perfect 2012 electoral vote predictions)

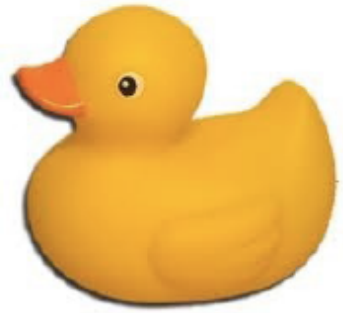
Black swans

Ideal forecasting targets are neither too common nor too frequent

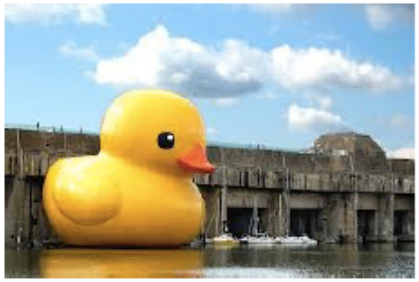


Good Judgment Project: look for events with a 10% probability

The Forecasting Zoo

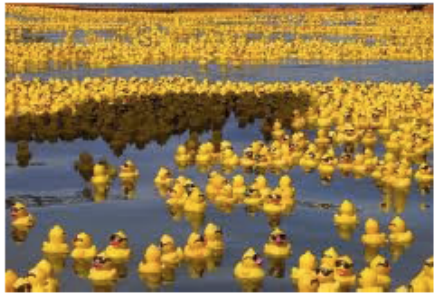


Ducks can be interesting...



Size

Variety

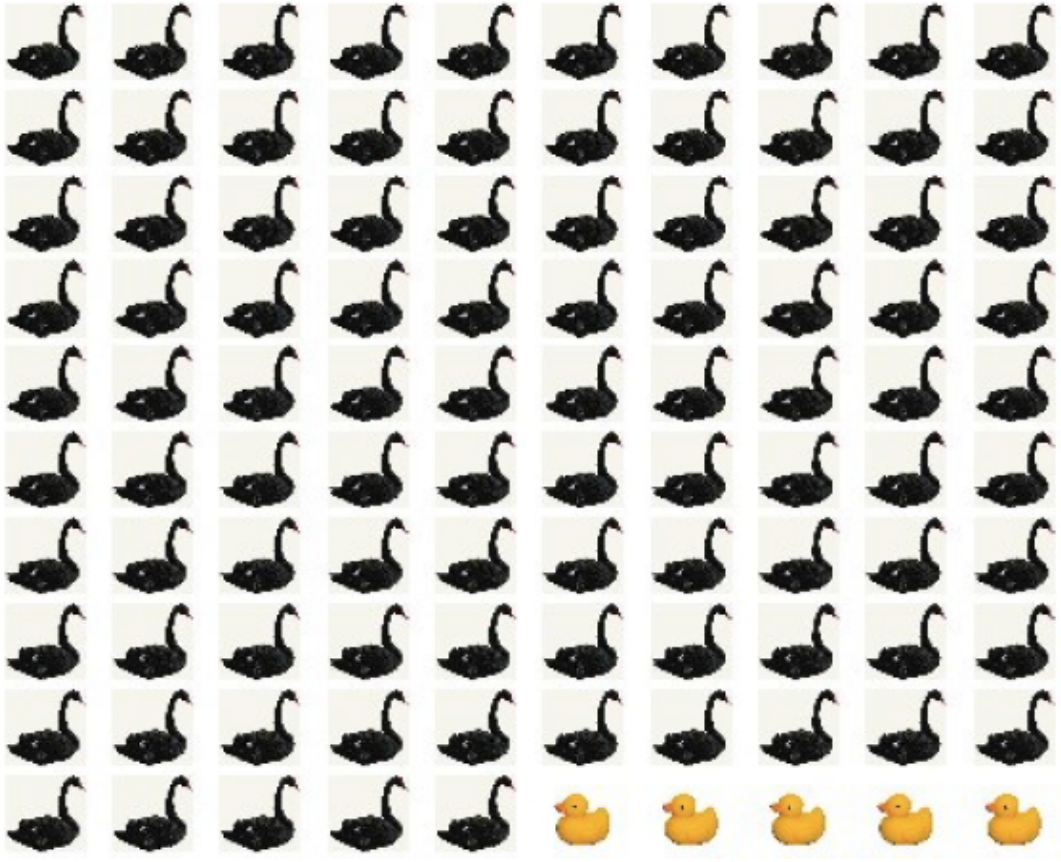


Quantity

Suspicious behaviors

And this is going too far...

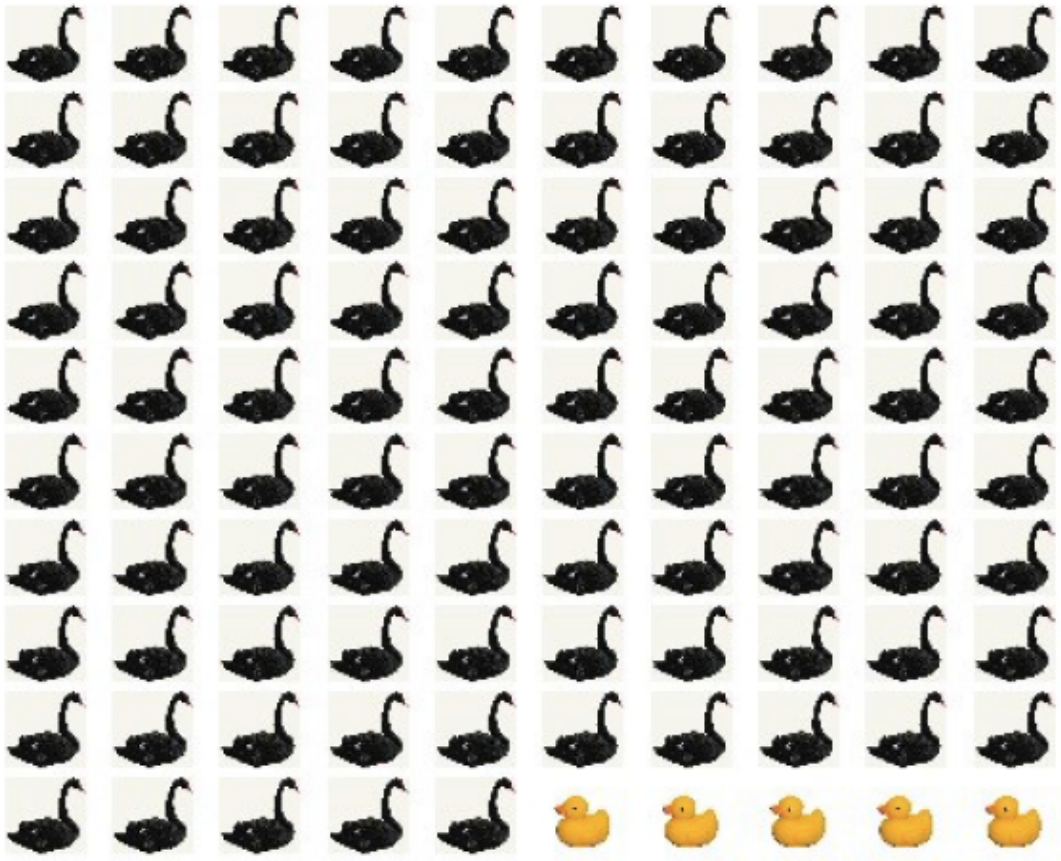
DARPA-World!



By definition, most black swans *will not occur*! So there is little point in investing a large amount of effort trying to predict them.

And this is going too far...

DARPA-World!



By definition, most black swans *will not occur*! So there is little point in investing a large amount of effort trying to predict them.
"Can your model predict a chemical attack by self-recruited Mexican jihadis working as rodeo clowns in Evanston, Wyoming? Why not?!"

Challenge: distinguishing black swans from rare events

Black swan: an event that has a low probability even conditional on other variables

Rare event: an event that occurs infrequently, but conditional on an appropriate set of variables, does not have a low probability

Medical analogy: certain rare forms of cancer appear to be highly correlated with specific rare genetic mutations. Conditioned on those mutations, they are not black swans.

Another important category: high probability events which are ignored. The "sub-prime mortgage crisis" was the result of the failure of a large number of mortgage which models had completely accurately identified as "sub-prime" and thus likely to fail. This was not a low probability event.
Upton Sinclair: It is hard to persuade someone to believe something when he can make a great deal of money not believing it.

Heterogeneous environments

- Per Pinker, Goldstein, Mueller, etc, is the system changing significantly while we are trying to model it? How far back are data still relevant?
- How different are various types of militarized non-state actors? For example, how much do al-Qaeda and international narcotics networks have in common?
- We are also using a more heterogenous set of forecasting methods, and probably do not understand their weak points as well as we understand those of regression-based models.
- Threats tend to occur in small number of "hot-spots"
  - Europe 1910-1945
  - Middle East 1965-present
  - Balkans in 1990s
  - Internal conflict in India

Note that all of these are complicated by rare events—some of which may be black swans—since it limits the number of observations we have on the dependent variable.

Theory: what can and cannot be predicted?

Is astronomy scientific?

Astronomy generally has a very good record of prediction, and from the earliest days of astronomy, successful prediction has been a key legitimating factor

- relation between star positions and events like the Nile flood
- eclipses
- orbits
- Halley's comet
- precision steering of space-craft

Is astronomy scientific?

Astronomy generally has a very good record of prediction, and from the earliest days of astronomy, successful prediction has been a key legitimating factor

- relation between star positions and events like the Nile flood
- eclipses
- orbits
- Halley's comet
- precision steering of space-craft

Nonetheless, astronomy cannot predict, nor does it attempt to predict:

- solar flares, despite their potentially huge economic consequences
- previously unseen comets
- next nearby supernova

Determinism: The Pioneer spacecraft anomaly

"[Following 30 years of observations] When all known forces acting on the spacecraft are taken into consideration, a very small but unexplained force remains. It appears to cause a constant sunward acceleration of $(8.74 \pm 1.33) \times 10^{-10} m/s^2$ for both spacecraft."

Source: Wikipedia

Irreducible sources of error-1

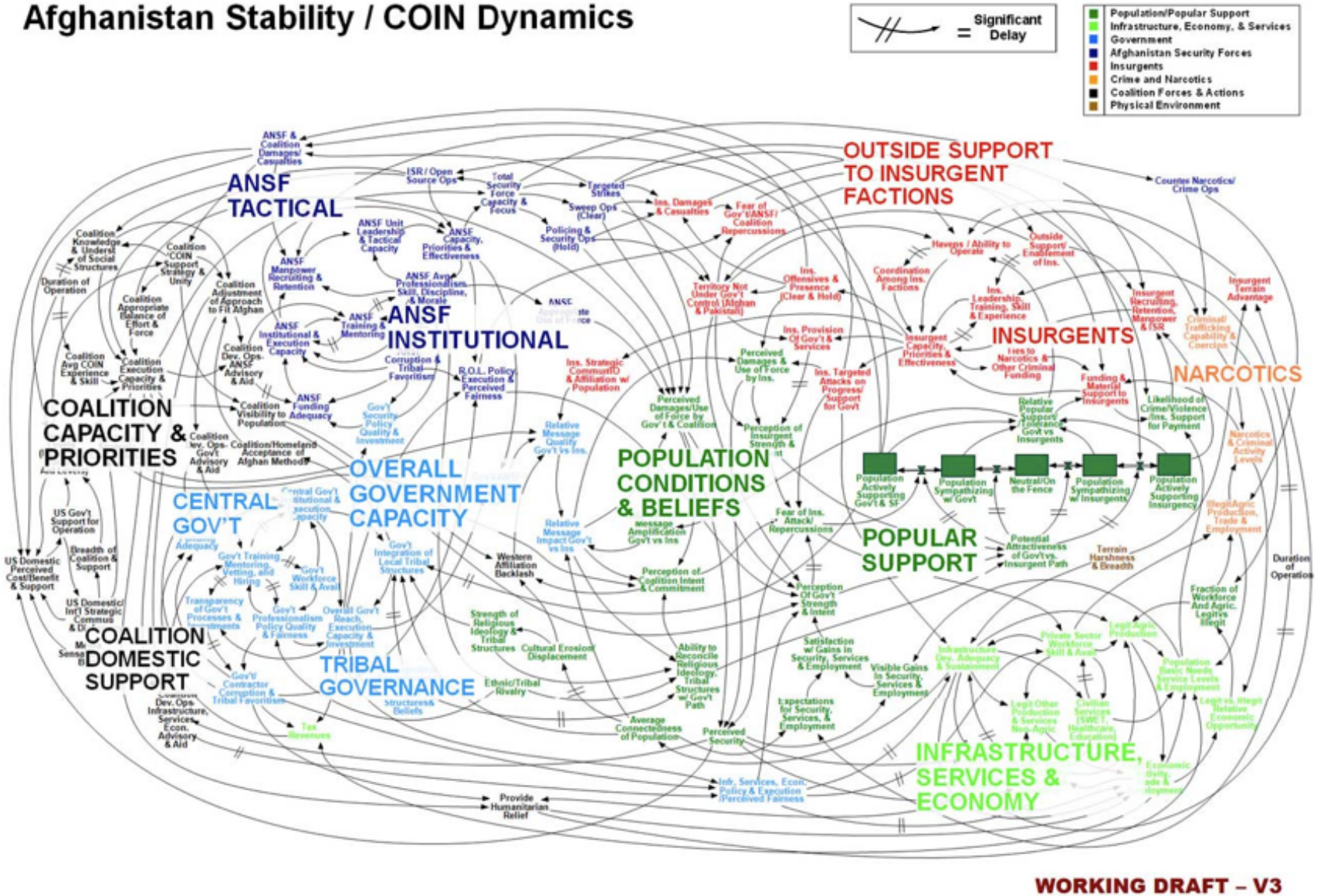
- Specification error: no model of a complex, open system can contain all of the relevant variables;
- Measurement error: with very few exceptions, variables will contain some measurement error
  - presupposing there is even agreement on what the "correct" measurement is in an ideal setting;
  - Predictive accuracy is limited by the square root of measurement error: in a bivariate model if your reliability is 80%, your accuracy can't be more than 90%
  - This biases the coefficient estimates as well as the predictions
- Quasi-random structural error: Complex and chaotic deterministic systems behave as if they were random under at least some parameter combinations .
  Chaotic behavior can occur in equations as simple as $x_{t+1} = a{x_t}^2 + bx_t$

Irreducible sources of error-2

- Rational randomness such as that predicted by mixed strategies in zero-sum games
- Arational randomness attributable to free-will
  - Rule-of-thumb from our rat-running colleagues:
    "A genetically standardized experimental animal, subjected to carefully controlled stimuli in a laboratory setting, will do whatever it wants."
- Effective policy response:
  in at least some instances organizations will have taken steps to head off a crisis that would have otherwise occurred.
- The effects of natural phenomenon
  - the 2004 Indian Ocean tsunami dramatically reduced violence in the long-running conflict in Aceh

(Tetlock (2013) independently has an almost identical list of the irreducible sources of error.)

Open, complex systems



Balancing factors which make behavior predictable

- Individual preferences and expectations, which tend to change very slowly
- Organizational and bureaucratic rules and norms
- Constraints of mass mobilization strategies
- Structural constraints:
  the Maldives will not respond to climate-induced sea level rise by building a naval fleet to conquer Singapore.
- Choices and strategies at Nash equilibrium points
- Autoregression (more a result than a cause)
- Network and contagion effects (same)

"History doesn't repeat itself but it rhymes"
Mark Twain (also occasionally attributed to Friedrich Nietzsche)

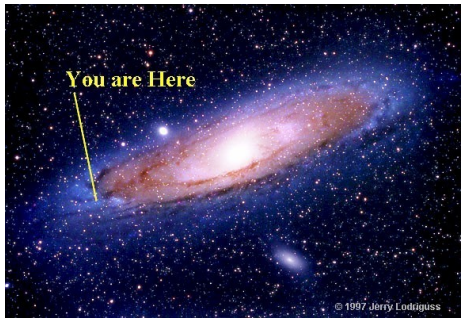
Paradox of political prediction

Political behaviors are generally highly incremental and vary little from day to day, or even century to century (Putnam).

Nonetheless, we *perceive* politics as very unpredictable because we focus on the unexpected (Kahneman).

Consequently the only "interesting" forecasts are those which are least characteristic of the system as a whole. However, only some of those changes are actually predictable.

Finding a non-trivial forecast



- Too frequent: prediction is obvious without technical assistance
- Too infrequent: prediction may be correct, but the event is so infrequent that
  - The prediction is irrelevant to policy
  - Calibration can be very tricky
  - Accuracy of the model is difficult to assess
- "Just right": these are situations where typical human accuracy is likely to be flawed, and consequently these could have a high payoff, but there are not very many of them.

Models matter

Arab Spring is an unprecedented product of the new social media

- Model used by Chinese censors of NSM: King, Peng, Roberts 2012
- Next likely candidates: Africa

Arab Spring is an example of an instability contagion/diffusion process

- Eastern Europe 1989-1991, OECD 1968, CSA 1859-1861, Europe 1848, Latin America 1820-1828
- Next likely candidates: Central Asia

Arab Spring is a black swan

- There is no point in modeling black swans, you instead build systems robust against them

Statistical and modeling challenges

Rare events

- Incorporate much longer historical time lines?—Schelling used Caesar's *Gallic Wars* to analyze nuclear deterrence
- New approaches made possible by computational advances

Analysis of event sequences, which are not a standard data type

- There are, however, a large number of available methods, and it is just possible that these will work with very large data sets such as GDELT
- This possibility will be discussed in detail in Lecture 5

Causality

- Oxford *Handbook of Causation* is 800 pages long

Integration of qualitative and qualitative/subject-matter-expert (SME) information

- Bayesian approaches using prior probabilities are promising but to date they have not really been used

Pournelle's Law:
No task is so virtuous that it will not attract idiots

- Need to establish with the media and policy-makers that not every forecast, even (or especially) those made using "Big Data" methods, is scientifically valid
  - It took the survey research community about thirty to forty years to establish professional credibility, though they have largely succeeded
- Conveying limitations of the methods against the hyper-confidence of pundits and individuals with secret models
  - Limitations of the data sources
  - Limitations of the data coding, particularly automated coding
  - Limitations of the model estimation
  - Limitations of probabilistic forecasts, particularly for rare events, even when the models are correct

Critical case: studies of climate change and conflict. As Pinker and Goldstein noted, people want to hear simple scary answers.

Levels of conflict forecasting models used in policy-making

- Structural: predict the cases (countries or regions) most likely to experience conflict
- Dynamic: predict a probability of conflict breaking out at a known point (or, more realistically, interval) in the future
- Counter-factual: predict how the change in some policy (e.g introduction of aid or peacekeepers) will affect the likelihood or magnitude of conflict

Prediction is easier than explanation; explanation is easier than manipulation. An insurance company doesn't care whether you die from a car wreck, cancer or a heart attack, they just need to know how long you are likely to live.

Statistical challenges

- Systematically dealing with measurement error and missing values rather than assuming "missing at random"
- Correctly leveraging ensemble methods which utilize multiple statistical and computational pattern recognition methods
  - PITF forecasting tournament; Bayesian model averaging
  - There are known and irreducible random elements in political behavior
- Upshot: you can't simply specify a desired rate of accuracy and assume by throwing sufficient money at the problem you will get there.

Chaos

- Deterministically generates behavior that appears random
- Attractors
- Sensitivity to initial conditions
- Parameter dependent: a well-behaved model can switch to chaotic behavior
- A simple finite-difference quadratic—in particular, the logistic model—can produce this. There is nothing mystical or complex about it.

Core issues in statistical forecasting

- Rare events
  - Predicting the mode of non-occurrence will be very accurate but not very useful
  - Limited positive cases available for estimation
- High autocorrelation
  - Predicting x_{t-1} will be very accurate but not very useful
  - Cases are not independent
- Heterogeneous subsets
  - ICEWS had China and Fiji, Indonesia and New Zealand in the same model
- Non-repeatability: observational rather than experimental
  - Stability of coefficients has not been explored extensively, and this is difficult because of rare events

Possible consequence of this: Complex models are not necessarily better

Keep it simple!

Large Scale Conflict Forecasting Projects

- State Failures Project 1994-2001
- Joint Warfare Analysis Center 1997
- FEWER [Davies and Gurr 1998]
- Center for Army Analysis 2002-2005
- Swiss Peace Foundation FAST 2000-2008
- Political Instability Task Force 2002-present
- DARPA ICEWS 2007-present
- IARPA ACE and OSI
- Peace Research Center Oslo (PRIO) and Uppsala University UCDP models

(much more on this tomorrow)

Convergent Results

- Most models require only a [very] small number of variables
- Indirect indicators—famously, infant mortality rate as an indicator of development—are very useful
- Temporal autoregressive effects are huge: the challenge is predicting onsets and cessations, not continuations
- Spatial autoregressive effects—"bad neighborhoods"—are also huge
- Multiple modeling approaches generally converge to similar accuracy
- 80% accuracy—in the sense of AUC around 0.8— in the 6 to 24 month forecasting window occurs with remarkable consistency: few if any replicable models exceed this, and models below that level can usually be improved
- Measurement error on many of the dependent variables—for example casualties, coup attempts—is still very large
- Forecast accuracy does not decline very rapidly with increased forecast windows, suggesting long term structural factors rather than short-term "triggers" are dominant. Trigger models more

Linear Regression (r^2) on Material Conflict Event Counts

Lead	Balkans	Palestine	Lebanon	West Africa
1	0.34	0.45	0.31	0.12
3	0.15	0.29	0.23	0.03 (n.s.)
6	0.06 (.04)	0.27	0.16	0.03 (n.s.)
12	0.04 (n.s.)	0.23	0.16	0.01 (n.s.)

Lead is in months. Results are significant at p¡0.0001 unless otherwise noted.
P-value is in (); n.s. = not significant at 0.10 level

Logistic Regression on Event Counts (in sample)

	Lead	Balkans	Palestine	Lebanon
50% level				
	1 month	73.7%	82.6%	75.3%
	6 month	64.3%	74.9%	68.5%
75% level				
	1 month	79.6%	79.6%	81.7%
	6 month	72.8%	79.2%	75.6%

Logistic Regression on Event Counts (1:3 out-of-sample)

	Lead	Balkans	Palestine	Lebanon
50% level				
	1 month	64.3%	57.3%	67.7%
	6 month	60.1%	- - - *	56.4%
75% level				
	1 month	66.1%	71.0%	82.3%
	6 month	61.6%	- - -	74.6%

*Palestine 6-month forecasts could not be estimated due to insufficient variance in high-conflict data points

Logistic Regression on Event Counts (1:1 out-of-sample)

	Lead	Balkans	Palestine	Lebanon
50% level				
	1 month	66.7%	64.4%	63.4%
	6 month	47.1%	38.1%	46.7%
75% level				
	1 month	85.3%	67.8%	75.4%
	6 month	87.1%	55.7%	61.3%

Hidden Markov models: Accuracy by positive and negative predictions

- "Correct"—percentage of the weeks that were correctly forecast, the percentage of time that a high or low conflict week would have been predicted correctly.
- "Forecast"—percentage of the weeks that were forecast as having high or low conflict actually turned out to have the predicted characteristic; the percentage of time that a type of prediction is accurate.

Balkans Hidden Markov Model: Accuracy for 23-Category Coding System

Experiment	%accuracy	% high correct	% low correct	% high forecast	% low forecast
P1	77.6	29.3	89.5	40.8	83.7
P3	76.0	29.0	87.9	37.9	82.9
P6	76.9	25.9	90.6	42.6	82.0
N1	54.2	92.7	45.3	28.1	96.4
N3	49.0	88.1	39.6	25.9	93.3
N6	47.7	88.5	37.4	26.3	92.8

Balkans Hidden Markov Model: Accuracy for 5-Category Coding System

Experiment	%accuracy	% high correct	% low correct	% high forecast	% low forecast
P1	74.4	46.2	81.5	38.9	85.6
P3	71.7	44.1	78.9	35.4	84.4
P6	71.4	44.2	78.8	36.4	83.8
N1	61.9	90.7	54.6	33.7	95.8
N3	57.8	87.0	50.2	31.4	93.6
N6	56.8	85.9	48.8	31.5	92.7

Difference in Accuracy between 23-Category and 5-Category Coding Systems

Experiment	% accuracy	% high correct	% low correct	% high forecast	% low forecast
P1	3.2	-16.9	8.0	1.9	-1.9
P3	4.3	-15.1	9.0	2.5	-1.5
P6	5.5	-18.3	11.8	6.2	-1.8
N1	-7.7	2.0	-9.3	-5.6	0.6
N3	-8.8	1.1	-10.6	-5.5	-0.3
N6	-9.1	2.6	-11.4	-5.2	0.1

Positive value: 23-category has higher accuracy

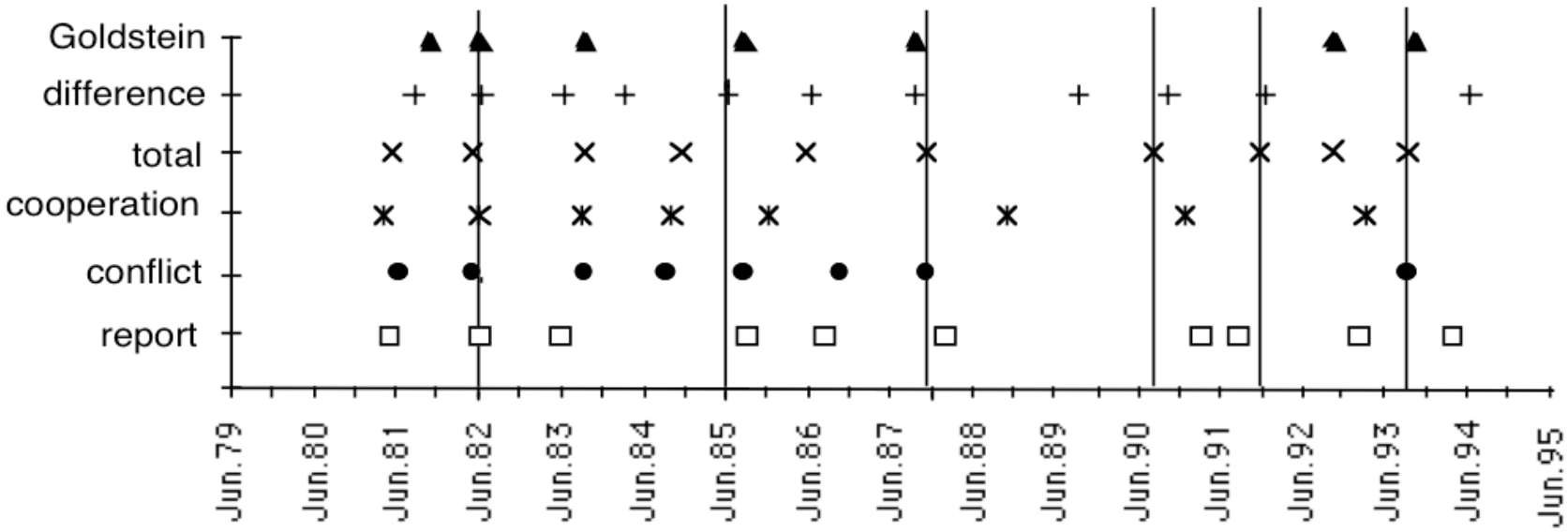
Simplifying Event Scales

Goldstein:	Goldstein weights
difference:	cooperative events = 1; conflictual events = -1
total:	all events = 1
conflict:	cooperative event = 0; conflictual events = 1
cooperation:	cooperative event = 1; conflictual events = 0
report:	1 if any event was reported in the month, 0 otherwise

Discriminant Analysis Results

Weighting scheme	%correct	variance explained	canonical correlation	Wilks' Λ	significance	# factors
Goldstein	85.6%	76.3%	0.85	0.008	<.001	6
difference	89.7%	74.7%	0.85	0.007	<.001	7
total	94.4%	83.0%	0.93	0.001	<.001	6
conflict	88.2%	76.9%	0.86	0.007	<.001	6
cooperation	92.3%	82.2%	0.91	0.002	<.001	7
report	89.2%	73.6%	0.87	0.008	<.001	7
random date	61.0%	69.5%	0.66	0.131	.37	0
random dyad	57.4%	68.8%	0.67	0.119	.18	0

Cluster Analysis Results



Why does detailed coding make so little difference?—sources of error in event data

Reporting error

- Missing events—limited reporting, censorship
- False events—rumors and propaganda

Coding error

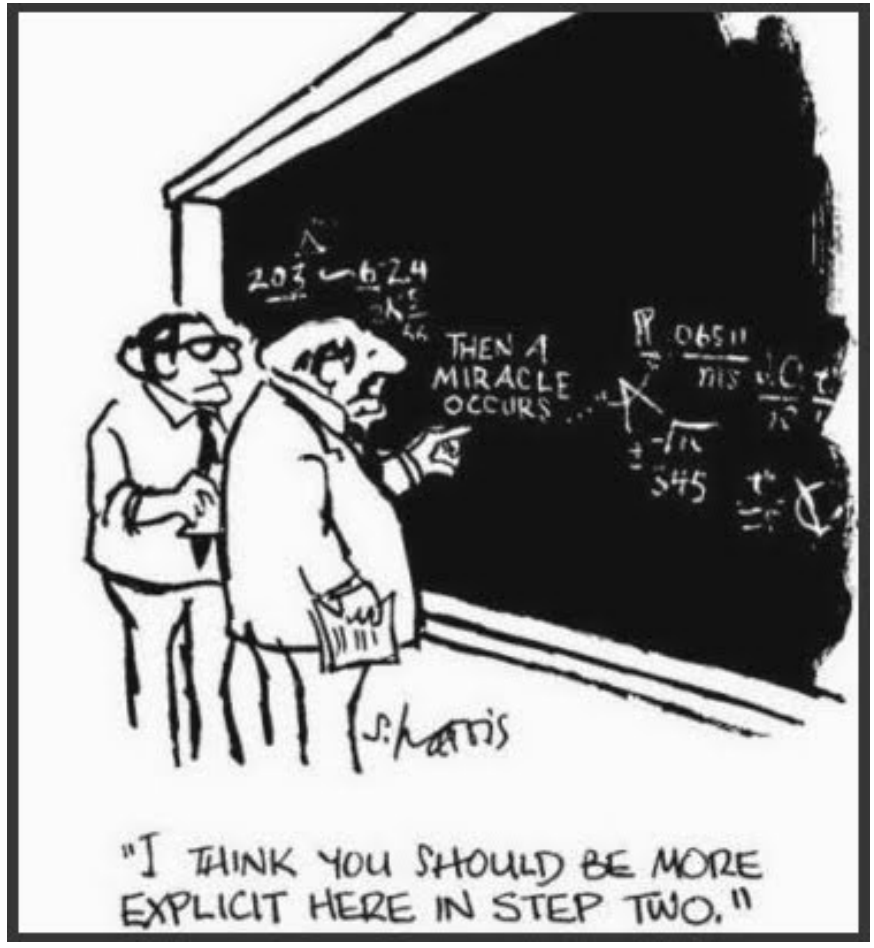
- Individual—coders are not correctly implementing the event coding system
- Systemic—event coding system does not reflect political behavior

Model specification

- model may be using the wrong indicators
- mathematical structure of the model does not produce good predictions
- models with diffuse information structuresneural networks, VAR, HMMare good at adapting to missing information

The artificial intelligence literature has consistently shown that experts over-estimate the amount of data they need
A small number of indicators will usually capture most of the available signal

Options and Cautions in Time Series Analysis



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

What could be predicted

- Levels of a continuous variable: classical time series methods
- Point predictions within a given time interval: logistic
  - This is the single most common approach, but a variety of different methods are being used
  - Poisson and negative binomial regression might be relevant here but high autocorrelation violates of the assumption of independence
- Point-prediction with a distribution
- Response of system to external shocks: vector autoregression
- Likelihood of an event as a function of time: Survival/hazard models
- Phase models: Bayesian switching models, hidden Markov, conditional random fields

Considerations in any time series model

- Lag structure in the dependent variable (autoregression): look at the autocorrelation function and the cross-correlation function
- Lag structure in the error term: if something occurs in a variable not in the equations (i.e. the "error") how long does it have an effect?
- Trend (exponential or linear): see GDELT
- Changes due to measurement, coding or method: see GDELT. Sometimes these are obvious, sometimes not.
- Outlying points with known explanations: if not filtered, these will bias the remaining estimates
- Stationarity: is the data generated by the same process for the entire interval?
- Rare events

Complicating factors in almost all conflict forecasting models

- Long time horizon eliminates most of the detailed lag effects (this could change in studies to much shorter time horizons)
- Autocorrelation is the dominant factor in the series
- Differences, however, may be almost random
- Onsets and cessations are the interesting part of the series, but they are very rare

The unreasonable effectiveness of incorrectly specified models

Most of the advanced time series methods have fairly complex underlying assumptions that are difficult if not impossible to satisfy in small-sample, heterogeneous observational situations. While they are preferable to simpler methods under those conditions, they are not—and may be worse—if the conditions are violated.

In order to adjust for this possibility, experiment with multiple models in split-sample evaluations. And don't trust your models.

The same applies for whether you are treat count or scaled data as if it was continuous:

"Box-Jenkins-Tiao" framework

Transform the data until it is stationary using some combinations of the following operations

- moving average: high-frequency filter
- differences: low-frequency filter
- lags

Problem: these models can produce good predictions but coefficients can be very difficult to interpret. In addition, they are designed for *interval level* (continuous) variables.

Slutsky-Yule Effect

MAVs induce induce cycles:

1. By definition, white noise random data has all cycles equally probable
2. MAVs filter out various frequencies
3. Whatever is left is your cycle (simple, eh?)

Granger Causality and Vector Autoregression

Y is "Granger-caused" by X when the prediction of Y by the lagged values of X and Y is better than the prediction by the lagged values of Y alone.

Vector Autoregression (VAR)
Essentially use a Granger approach, and pay no attention to the coefficient values because of the effects of autocorrelation and colinearity. Instead look at the effect of a shock to the variable. Widely used by the U.S. Federal Reserve and by John Freeman.

Problem (again): designed for interval-level variable

Count Models: Poisson

The Poisson is the probability distribution of the number of occurrences in a unit of time of a continuous time low-probability event which occurs independently.

- Derived by taking a binomial variable and letting the time interval go to zero.
- The variance of Poisson-distributed counts is equal to the mean.
- One of the earliest statistical regularities in the study of conflict was the Poisson distribution of wars over very long time scales (Richardson ca. 1930s)

Alternatives:

- Clustering: Variance is greater than the mean
- Spacing (even distribution): Variance is less than the mean

Poisson regression: Model the rate of occurrence based on covariates.

Count Models: Negative binomial

- Underlying distribution: number of successes before failure in discrete and independent Bernoulli/binomial trials
- In conflict models, assume cases are "at risk" for "failure"—either onset or cessation of violence—in each period
- Regression: Model this failure rate. This is particularly useful for events that occur on a partially-regular basis.

Count Models: potential issues

- Autocorrelation is almost certainly too high to be useful for modeling overall incidence.
- High autocorrelation also violates—big time—the assumption of independence
- Conversely, onsets and cessations may be *too rare* to provide sufficient information for an estimate

Survival/hazard models

- *Extensively* developed in medical and public health statistics, and consequently well understood with well-developed software
- Objective is estimating the shape of the survival curve, based on covariates and any of a number of possible curves.
  - This gets around the assumption of independence in the negative binomial
- Outcome is a probability at each time point, so easily suited for ROC curves and related methods
- As always, it is more difficult to work with in rare events situations, though the statistics community is familiar with these problems

Bayesian Model Averaging

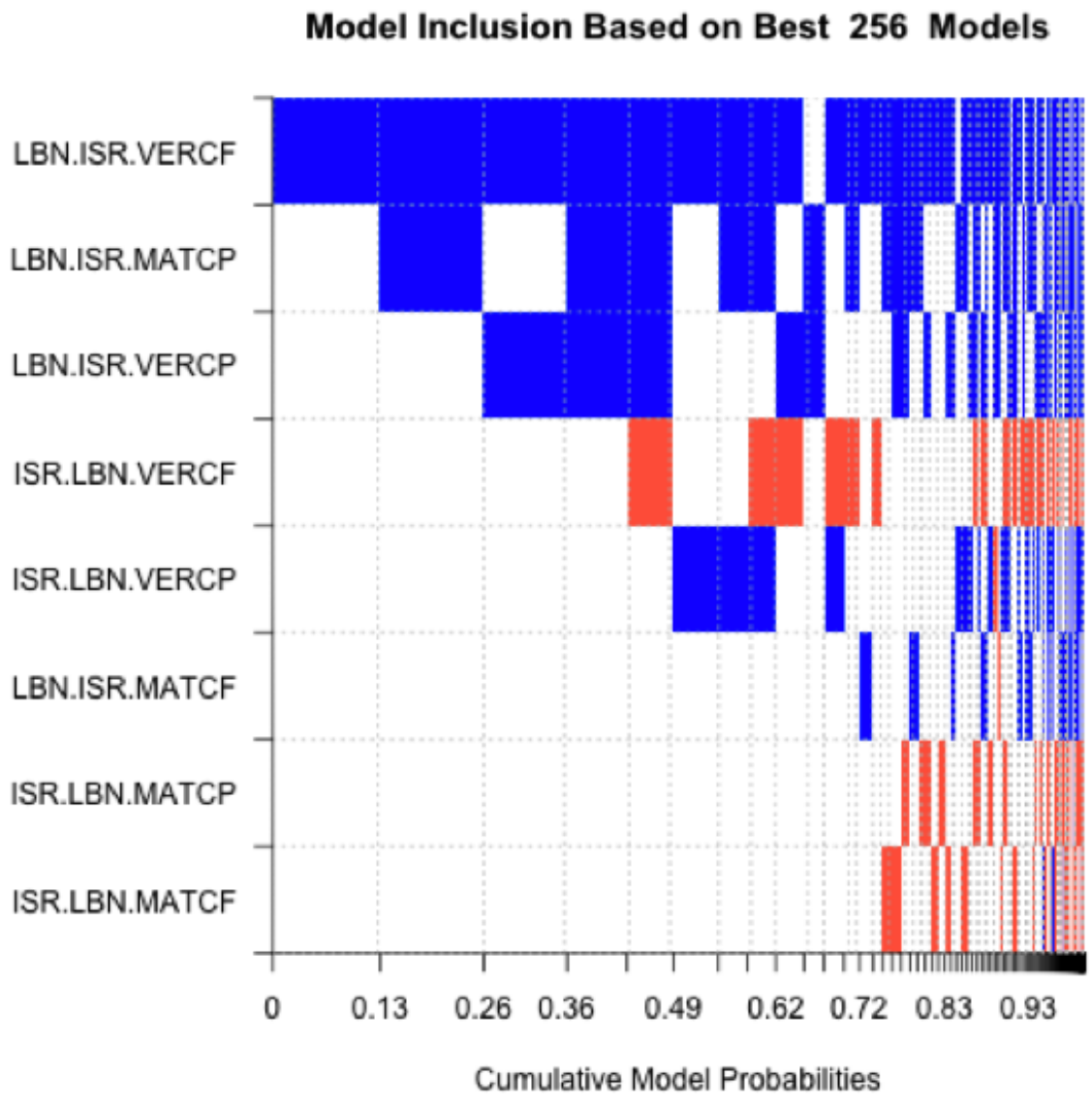
- Systematically integrates the information provided by all combinations of variables
- Result is the overall posterior probability that a variable is important
  - Without having to generate hundreds of papers and thousands of non-randomly discarded models
- Machine learning suggests that systematic assessment of models gives about 10% better accuracy with much less information, and completely eliminates the need for vaguely defined indicators
- Predictions can be made using an ensemble of all of the models
  - In meteorology and finance, these models are generally more robust in out-of-sample evaluations
- Framework is Bayesian rather than frequentist, which eliminates a long list of philosophical and interpretive problems with the frequentist approach

The problem of "controls"

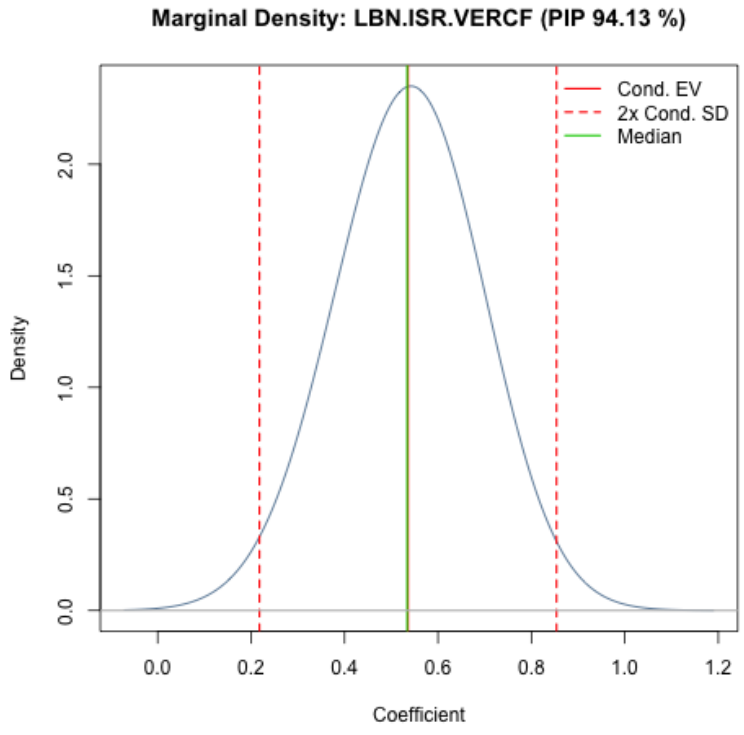
- For starters, they aren't "controls", they are just another variable
  - Often in a really bad [colinear] neighborhood
  - Nature bats last in $(X'X)^{-1}X'y$
  - For something closer to a control, use case matching or Bayesian priors
- Numerous studies over the past 50 years—all ignored (Kahneman)—have suggested that simple models are better
- In many forecasting models, there is no obvious theoretical reason for using any particular measure, so instead we have to assess multiple measures of the same latent concept: "power", "legitimacy", "authoritarianism"
  - This is a feature, not a bug
  - Regression approaches have terrible pathologies in these situations
  - Currently, we laboriously work through all of these options across scores of journal and conference papers presented over the course of years*

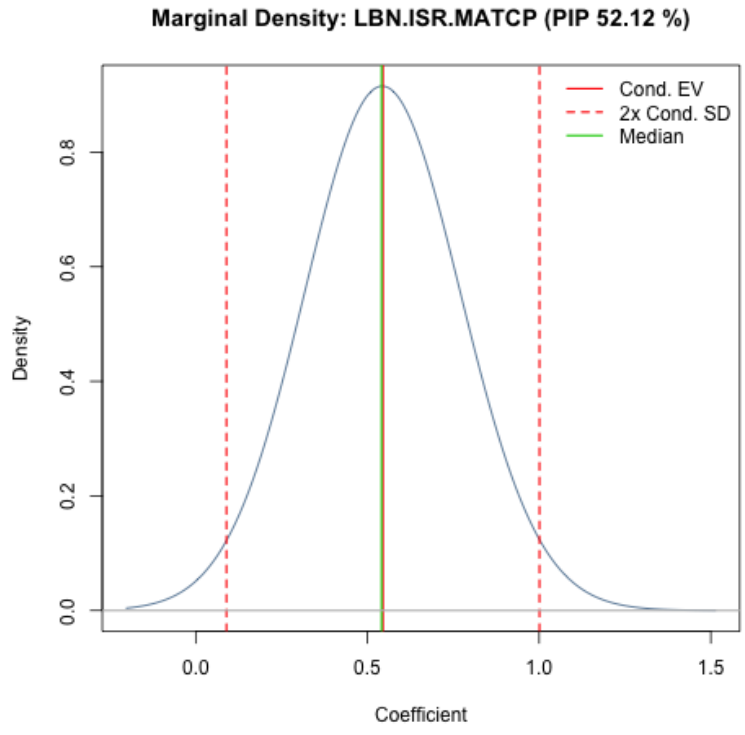
* So if BMA really catches on, a number of journals—and tenure cases—are doomed. On the former, how sad. On the latter, be afraid, be very afraid.

BMA: variable inclusion probabilities



BMA: Posterior probabilities





Thank you

Email: schrodt735@gmail.com

Slides: http://eventdata.parusanalytics.com/presentations.html

Forecasting papers:
http://eventdata.parusanalytics.com/papers.html