Forecasting Conflict
Lecture 2
Major U.S. Forecasting Projects

Philip A. Schrodt
Parus Analytical Systems
schrodt735@gmail.com

Graduate School of Decision Sciences
University of Konstanz
14 - 17 October 2013
ARPA Projects 1960s and 1970s
  - WEIS
  - COPDAB
  - Goldstein scales

National Science Foundation: DDIR, KEDS, CAMEO
  - DDIR
  - KEDS, PANDA, VRA
  - CAMEO and IDEA

State Failures Project and Political Instability Task Force
  - SFP Neural Network Models
  - PITF Core Models
  - PITF Forecasting Tournament
  - PITF Data
Overview-2

- DARPA Integrated Conflict Early Warning Systems (ICEWS)
  - ICEWS EOIs
  - Lockheed ICEWS models
  - W-ICEWS
- IARPA ACE and Good Judgment Project
  - Tetlock, *Expert Political Judgment*
  - ACE Forecasting teams
  - ACE Forecasting markets
The Debate

**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

---

**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 10, 2012
Factors encouraging technical political forecasting-1

- Conspicuous failures of existing methods: end of Cold War, post-invasion Iraq, Arab spring
- Success of forecasting models in other behavioral domains
  - Macroeconomic forecasting [maybe...]
  - Elections: Nate Silver effect
  - Demographic and epidemiological forecasting
  - Famine forecasting: USAID FEWS model
  - Example: statistical models for mortgage repayment were quite accurate
    - Moneyball
- Technological imperative
  - Increased processing capacity
  - Information available on the web
  - “Moore’s Law states that computing power doubles every 18 months. Human cognitive ability is pretty much a constant. This leads to some interesting and not always desirable substitution effects”
  
  Larry Bartels, Princeton University
Factors encouraging technical political forecasting-2

- Demonstrated utility of existing methods, which tend to converge on about 80% accuracy
  - Political Instability Task Force
  - ICEWS
  - “Big Data” analytical methods
- Decision-makers now expect visual displays of analytical information, which in turn requires systematic measurement
  - “They won’t read things any more”
Feedback: change behavior based on current conditions (or with a slight lag)
   Classical control systems

Feedforward (Casti): set behaviors based on the projected impact of the policy on a behavior in the distant future
Successful feedforward policies

- US Constitution (to 2013?)
- Marshall Plan
- Nuclear deterrence
- Euro (so far)
Feedforward failures

- European military mobilization plans ca. 1910
- U.S. policies in Iraq, Afghanistan 2003-present
- Various real estate bubbles in 2000s: US, Ireland, Spain
Why Event Data are well suited for predicting political change

- Structural indicators such as GDP, infant mortality, past or adjacent conflict change too slowly
  - They nonetheless affect the overall probability
- Social media indicators change too quickly
  - Though US government funders are completely obsessed with this at the moment. Tweet that!
- Newsworthy events are “just right”
  - And we’ve got the models to prove it
  - Which is why they are “newsworthy”
  - Structural indicators either are reflected in the patterns of events, or can be additional covariates
Possible objectives for forecasts

- Relative probabilities and watch lists
- Probabilities of specific events
- Causal relations: a change in X and the probability of Y will change
- "Actionable" relations: this is the subset of causal relations where X could realistically be changed, and is surprisingly small
Policy relevant forecast interval: 6 to 18 months
Early technical forecasting models

- Divination model of sheep liver
- Babylonia, ca. 600 BCE
- Persian conquest of Babylonia: 539 BCE
Early technical forecasting models

- Divination model of sheep liver
- Babylonia, ca. 600 BCE
- Persian conquest of Babylonia: 539 BCE
Temple of Apollo at Delphi

Sample prediction (Herodotus): “A mighty kingdom will fall”
Parus Analytical Systems Global Headquarters [proposed]
Dueling Media Assessments
This must be important: it’s in *The Economist*!

**The science of civil war**

**What makes heroic strife**

Computer models that can predict the outbreak and spread of civil conflict are being developed

Apr 21st 2012 | from the print edition
But Wired is not impressed
But Phil, the best models are classified!

Hollywood tells me so!

- Yeah, right. . .
- No systematic evidence of this: if it is true, government is spending vast resources to obscure this fact (at least from me. . .)
- Clearly isn’t operating at the policy level
- Probably some models have worked at some points in the past but they have not proven robust
- *Much* more likely: there is serious snake-oil sales going on here as well. . .
- Even if this is true, we need to reverse-engineer these to get them into the unclassified literature and acquaint policy-makers with the techniques
- (but it probably isn’t true. . .)
We know this about at least one classified model...
OCTOBER 7TH 2013


CAN YOU PLEASE GET YOUR SHIT TOGETHER?
THIS IS EMBARRASSING.

SINCERELY,
A CONCERNED CITIZEN
Challenges to integrating models into decision-making

Forecasting is hard (Tetlock) Probabilistic reasoning is hard (Kahneman, Taleb) Statistics is new compared to deterministic modeling and is still changing, even at very fundamental levels

▶ Frequentist vs Bayesian approaches
▶ New approaches made possible by computational advances

The answers aren’t simple, even if some colonel wants them to be simple

▶ Our 20th century peer competitors were trained as political ideologues; our 21st century peer competitors are trained as engineers
Event Coding systems

- **WEIS** ca. 1965
  Charles McClelland, Rodney Tomlinson, DARPA

- **COPDAB** ca. 1970
  Edward Azar

- **PANDA** ca. 1990
  Doug Bond

- **IDEA** ca. 1998
  Doug Bond, Craig Jenkins and Charles Taylor

- **CAMEO** ca. 2002
  Deborah Gerner and Philip Schrodlt
Categorization of Political Interactions

- Distinct English-language verb phrases: 5,000 to 15,000 (MUC, KEDS, PANDA projects)

- Micro-level categories: 50 to 200 (WEIS, BCOW, IDEA, CAMEO)

- Macro-level categories: 10 to 20 (WEIS, COPDAB, IPB, World Handbook)
<table>
<thead>
<tr>
<th></th>
<th>WEIS Primary Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Yield</td>
</tr>
<tr>
<td>02</td>
<td>Comment</td>
</tr>
<tr>
<td>03</td>
<td>Consult</td>
</tr>
<tr>
<td>04</td>
<td>Approve</td>
</tr>
<tr>
<td>05</td>
<td>Promise</td>
</tr>
<tr>
<td>06</td>
<td>Grant</td>
</tr>
<tr>
<td>07</td>
<td>Reward</td>
</tr>
<tr>
<td>08</td>
<td>Agree</td>
</tr>
<tr>
<td>09</td>
<td>Request</td>
</tr>
<tr>
<td>10</td>
<td>Propose</td>
</tr>
<tr>
<td>11</td>
<td>Reject</td>
</tr>
<tr>
<td>12</td>
<td>Accuse</td>
</tr>
<tr>
<td>13</td>
<td>Protest</td>
</tr>
<tr>
<td>14</td>
<td>Deny</td>
</tr>
<tr>
<td>15</td>
<td>Demand</td>
</tr>
<tr>
<td>16</td>
<td>Warn</td>
</tr>
<tr>
<td>17</td>
<td>Threaten</td>
</tr>
<tr>
<td>18</td>
<td>Demonstrate</td>
</tr>
<tr>
<td>19</td>
<td>Reduce Relationship</td>
</tr>
<tr>
<td>20</td>
<td>Expel</td>
</tr>
<tr>
<td>21</td>
<td>Seize</td>
</tr>
<tr>
<td>22</td>
<td>Force</td>
</tr>
</tbody>
</table>
Goldstein Scale [WEIS]

010: [1.0] YIELD
011: [0.6] SURRENDER
012: [0.6] RETREAT
013: [2.0] RETRACT
014: [3.0] ACCOMODATE, CEASEFIRE
015: [5.0] CEDE POWER
020: [0.0] COMMENT
021: [-0.1] DECLINE COMMENT
022: [-0.4] PESSIMISTIC COMMENT
023: [-0.2] NEUTRAL COMMENT
024: [0.4] OPTIMISTIC COMMENT
070: [7.0] REWARD
071: [7.4] EXTEND ECON AID
072: [8.3] EXTEND MIL AID
073: [6.5] GIVE OTHER ASSISTANCE
110: [-4.0] REJECT
111: [-4.0] TURN DOWN
112: [-4.0] REFUSE
113: [-5.0] DEFY LAW
170: [-6.0] THREATEN
171: [-4.4] UNSPECIFIED THREAT
172: [-5.8] NONMILITARY THREAT
173: [-7.0] SPECIFIC THREAT
174: [-6.9] ULTIMATUM
220: [-9.0] FORCE
221: [-8.3] NONINJURY DESTRUCTION
222: [-8.7] NONMIL DESTRUCTION
223: [-10.0] MILITARY ENGAGEMENT
Problems with the Goldstein scale

- It started out quite arbitrary, and the CAMEO versions are even worse
- It tends to be dominated by violence events, which mask low levels of cooperative events
- It correlates highly with the event count, and in fact simple event counts do almost as well, similar to the result that unweighted equations do well
- The data are nominal!: get over it
The number of actors who must be identified is substantially greater than the number involved in inter-state events

- Detailed geographical information—city, region and administrative unit names—may be required
- Ethnic group names may be important
- Leadership is less stable—“five minutes of fame”

Coverage in international news sources may be less consistent, with a focus on

- Major events
- Periods when a reporter happens to be in the area
- Events in major cities (or cities with 5-star hotels)
Sentences being coded may assume substantial implicit knowledge

- This is particularly true for full-story coding

In militarized conflicts, large parts of the country may be inaccessible

Activities of unidentified actors may be important: “gunmen killed two journalists. . .”
Observation:
Every article in the remaining discussion was published at least five years (!) after the original research was done.
Observation:
Every article in the remaining discussion was published at least five years (!) after the original research was done.

This is driving me crazy...
State Failures Project

- Initiated by Vice President Gore in response to failures in Balkans, Somalia, Rwanda
- Neural network models
- Genocide models
- Initially looking at about 700 variables, mostly economics; final model was much simpler
Failures of the State Failures models

- Selection on the dependent variable

- Genocide project focused on extreme events and therefore the sample was too small
  - Additional problems in confusion between empirical and legal definitions of “genocide”, hence later emphasis on “mass killings”

- Failure to statistically adjust for rare events: King and Zeng 2001

- Neural network models were needlessly complex
  - Normalization methods could not be replicated

- 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”

- 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
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

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Fearon &amp; Laitin Model</th>
<th>Collier &amp; Hoeffler Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correctly Predicted</td>
<td>False Positives</td>
</tr>
<tr>
<td>0.5</td>
<td>0/107</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>1/107</td>
<td>3</td>
</tr>
<tr>
<td>0.1</td>
<td>15/107</td>
<td>66</td>
</tr>
</tbody>
</table>

Figure 2. Comparison of Predictive Power and Statistical Significance

Fearon & Laitin Model

Ward, Bakke, Greenhill 2010: Prediction vs. significance
Ward, Bakke, Greenhill 2010: Prediction vs. significance

Collier & Hoeffler Model

- Population
- Male Secondary Schooling
- GDP Growth
- Commodity Dependence
- Geographic Dispersion
- Squared Commodity Dependence
- Peace Duration
- Social Fractionalization
- Ethnic Dominance

Change in Predictive Power vs. Statistical Significance
Political Instability Task Force

- US government, multi-agency: 1995-present
- Statistical modeling of various forms of state-level instability
- Forecasting models actively used since about 2005
  - Two year probability forecasts with roughly 80% accuracy (AUC)
  - Predominantly logistic models with a simple “standard PITF” set of variables; shifting to Bayesian approaches
  - (PITF has accumulated a set of 2700 variables but only a small number end up being important predictors)
Political Instability Task Force (AJPS 2010)

**TABLE 2 Out-of-Sample Prediction Exercise for Observed Onsets of Instability, 1995–2004**

A. Countries That Had Instability Onsets, 1995–2004. Quintile/decile in model score rankings based on 2-yr. prior data

<table>
<thead>
<tr>
<th>Year</th>
<th>Top Decile</th>
<th>Second Decile</th>
<th>Second Quintile</th>
<th>Third Quintile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>Armenia, Comoros</td>
<td>Belarus</td>
<td></td>
<td>Nepal</td>
</tr>
<tr>
<td>1996</td>
<td>Albania, Niger, Zambia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>Cambodia, Congo-Brazz.</td>
<td></td>
<td></td>
<td>Serbia/Montenegro</td>
</tr>
<tr>
<td>1998</td>
<td>Guinea-Bissau, Lesotho</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>Ethiopia, Haiti</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>Solomon IIs., Guinea*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>Cote d’Ivoire</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Central African Republic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Iran*</td>
<td>Yemen*</td>
<td></td>
<td>Thailand*</td>
</tr>
</tbody>
</table>

B. Tabulation of All Country-years, 1995–2004. Model estimates based on censored data, using only sample data from prior to year of forecast (countries w/population over 500,000, no ongoing conflict, at least two years old)

<table>
<thead>
<tr>
<th></th>
<th>Countries with Instability in $t + 2$</th>
<th>Countries Remaining Stable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted for Instability (Top Quintile)</td>
<td>18</td>
<td>233</td>
</tr>
<tr>
<td>Predicted for Stability (Not Top Quintile)</td>
<td>3</td>
<td>992</td>
</tr>
<tr>
<td>N = 1,246 Percent Classed Correctly</td>
<td>85.7%</td>
<td>81.0%</td>
</tr>
</tbody>
</table>

Number of instability onsets, 1995–2004: 21. Number of instability onsets in top quintile of model scores: 18 (86%).

*Cases added to the problem set in 2005 update.

This is ca. 2010
PITF-sponsored datasets

- Political 4 (Marshall)
- Institutions and Elections (Regan)
- Worldwide Atrocities (Schrodt)
- Non-state mass killings (Valentino)
<table>
<thead>
<tr>
<th>CONCEPT</th>
<th>SELECTED EXAMPLES OF MEASURES TESTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>state capacity</td>
<td>infant mortality, population, GDP, military personnel, polity durability</td>
</tr>
<tr>
<td>violent conflict</td>
<td>civil war, armed attacks, regional conflicts, reported fatalities in political violence, government mass killing</td>
</tr>
<tr>
<td>non-violent challenges to state authority</td>
<td>protests, strikes, government crises</td>
</tr>
<tr>
<td>government institutions</td>
<td>democracy, autocracy, factionalism, other polity measures</td>
</tr>
<tr>
<td>ethnic relations</td>
<td>ethnic diversity, elite ethnicity, state-led discrimination</td>
</tr>
<tr>
<td>demographics</td>
<td>youth-bulge</td>
</tr>
<tr>
<td>international ties</td>
<td>GATT/WTO membership, trade-openness</td>
</tr>
</tbody>
</table>

Source: Ben Valentino and Chad Hazlett, “Forecasting Non-state Mass Killings”, October 2012
A Global Model for Forecasting Political Instability

Jack A. Goldstone  George Mason University
Robert H. Bates  Harvard University
David L. Epstein  Columbia University
Ted Robert Gurr  University of Maryland
Michael B. Lustik  Science Applications International Corporation (SAIC)
Monty G. Marshall  George Mason University
Jay Ulfelder  Science Applications International Corporation (SAIC)
Mark Woodward  Arizona State University

Examining onsets of political instability in countries worldwide from 1955 to 2003, we develop a model that distinguishes countries that experienced instability from those that remained stable with a two-year lead time and over 80% accuracy. Intriguingly, the model uses few variables and a simple specification. The model is accurate in forecasting the onsets of both violent civil wars and nonviolent democratic reversals, suggesting common factors in both types of change. Whereas regime type is typically measured using linear or binary indicators of democracy/autocracy derived from the 21-point Polity scale, the model uses a nonlinear five-category measure of regime type based on the Polity components. This new measure of regime type emerges as the most powerful predictor of instability onsets, leading us to conclude that political institutions, properly specified, and not economic conditions, demography, or geography, are the most important predictors of the onset of political instability.

Source: Amer J of Pol Sci Vol 54, no. 1, Jan 2010 pg. 190
### TABLE 1  Results of Global Analysis of Onsets of Instability

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Full Problem Set</th>
<th>Civil War Onsets</th>
<th>Adverse Regime Change Onsets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Odds Ratio</td>
<td>Coefficient</td>
</tr>
<tr>
<td></td>
<td>(S.E.)</td>
<td>(95% CI)</td>
<td>(S.E.)</td>
</tr>
<tr>
<td>Regime Type (Full Autocracy as Reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial Autocracy</td>
<td>1.85***</td>
<td>6.37</td>
<td>1.94***</td>
</tr>
<tr>
<td>(0.47)</td>
<td>(2.53, 16.02)</td>
<td></td>
<td>(0.62)</td>
</tr>
<tr>
<td>Partial Democracy with Factionalism</td>
<td>3.61***</td>
<td>36.91</td>
<td>3.35***</td>
</tr>
<tr>
<td>(0.51)</td>
<td>(13.5, 101)</td>
<td></td>
<td>(0.73)</td>
</tr>
<tr>
<td>Partial Democracy without Factionalism</td>
<td>1.83***</td>
<td>6.22</td>
<td>.981</td>
</tr>
<tr>
<td>(0.54)</td>
<td>(2.17, 17.8)</td>
<td></td>
<td>(0.79)</td>
</tr>
<tr>
<td>Full Democracy</td>
<td>0.981</td>
<td>2.67</td>
<td>.545</td>
</tr>
<tr>
<td>(0.68)</td>
<td>(0.70, 10.2)</td>
<td></td>
<td>(0.92)</td>
</tr>
<tr>
<td>Infant Mortality†</td>
<td>1.59***</td>
<td>6.59</td>
<td>1.64***</td>
</tr>
<tr>
<td>(0.35)</td>
<td>(2.91, 14.9)</td>
<td></td>
<td>(0.48)</td>
</tr>
<tr>
<td>Armed Conflict in 4+ Bordering States</td>
<td>3.09***</td>
<td>22.0</td>
<td>2.81***</td>
</tr>
<tr>
<td>(0.95)</td>
<td>(3.42, 142)</td>
<td></td>
<td>(0.82)</td>
</tr>
<tr>
<td>State-Led Discrimination</td>
<td>0.657*</td>
<td>1.93</td>
<td>1.17***</td>
</tr>
<tr>
<td>(0.30)</td>
<td>(1.08, 3.45)</td>
<td></td>
<td>(0.36)</td>
</tr>
</tbody>
</table>

| N = Total (Problems, Controls) | 468 (117, 351) | 260 (65, 195) | 196 (49, 147) |

| Onsets Correctly Classified | 80.3%         | 80.0%         | 87.8%         |
| Controls Correctly Classified | 81.8%         | 81.0%         | 87.8%         |

**p < 0.001, **p < 0.01, *p < 0.05. †Odds ratios for continuous variables compare cases at the 75th and 25th percentiles.

Source: Amer J of Pol Sci Vol 54, no. 1, Jan 2010 pg. 190
PITF Forecasting Tournament

Source data: 2700 variables

- Logistic models
- Bayesian model averaging
- Random forests
- Nearest neighbor clustering
- Bayesian Markov switching model
- Hazard models

Source: Jay Ulfelder, SSRN paper
PITF Model: Non-state Mass Killings Onset

Non-State Mass Killing Onsets 1989-2009 (logit)

<table>
<thead>
<tr>
<th>Variable (PITF variable name)</th>
<th>Coeff (C-RSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government Crises (bnkv101)</td>
<td>0.9847*** (0.3570)</td>
</tr>
<tr>
<td>Population (log of bnkv4)</td>
<td>.3969*** (0.1442)</td>
</tr>
<tr>
<td>Infant Mortality (log of cnsimr)</td>
<td>1.8251*** (0.5307)</td>
</tr>
<tr>
<td>Ongoing Government Mass Killing (sftpval)</td>
<td>0.9831** (0.4030)</td>
</tr>
<tr>
<td>Constant</td>
<td>-188472*** (3.1517)</td>
</tr>
<tr>
<td>N</td>
<td>3296</td>
</tr>
</tbody>
</table>

Cluster-robust standard errors (clustered on country)

Source: Ben Valentino and Chad Hazlett, “Forecasting Non-state Mass Killings”, October 2012
PITF Model: Non-state Mass Killings Onset

Model Performance Statistics

AUC = 0.88

PITF CUT POINT = 82%
(...but with 20 false positives per true positive)
## 2011-12 Non-State Mass Killing “Watch List”

<table>
<thead>
<tr>
<th>country</th>
<th>predicted risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angola</td>
<td>.062</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>.017</td>
</tr>
<tr>
<td>Chad</td>
<td>.018</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>.028</td>
</tr>
<tr>
<td>Iran</td>
<td>.023</td>
</tr>
<tr>
<td>Cote d'Ivoire</td>
<td>.023</td>
</tr>
<tr>
<td>Kenya</td>
<td>.025</td>
</tr>
<tr>
<td>Laos</td>
<td>.016</td>
</tr>
<tr>
<td>Madagascar</td>
<td>.020</td>
</tr>
<tr>
<td>Mali</td>
<td>.025</td>
</tr>
<tr>
<td>Burma</td>
<td>.026</td>
</tr>
<tr>
<td>Mozambique</td>
<td>.026</td>
</tr>
<tr>
<td>Nepal</td>
<td>.047</td>
</tr>
<tr>
<td>Niger</td>
<td>.026</td>
</tr>
<tr>
<td>North Korea</td>
<td>.019</td>
</tr>
<tr>
<td>Sudan</td>
<td>.046</td>
</tr>
<tr>
<td>Tanzania</td>
<td>.016</td>
</tr>
</tbody>
</table>

Source: Ben Valentino and Chad Hazlett, “Forecasting Non-state Mass Killings”, October 2012
Estimation: Immune vs. at Risk
To model this mechanism we use split-population models

(a) Splitting Population
(b) Baseline hazard
Immune versus at Risk

\( \Pi = \text{splitting parameter}, \, \delta = \text{censoring indicator}, \, f(t) = \text{density function (information on when units fail)}, \, S(t) = \text{survivor function (information on how long units survive)} \)

- Countries that are right censored (have not had a coup event up to \( t \)) contribute information to the survivor function (\( S \)), but not to the probability that a coup occurs prior to \( t \).
- Countries with a coup provide information to the density function when the coup event occurs.
- \( \delta_i \equiv 1 \iff \text{uncensored, i.e., no coup} \)

\[
L(\theta | (t_1, \ldots, t_n)) = \prod_{i=1}^{N} \left\{ (1 - \pi) f(t_i) \right\}^{\delta_i} \times \left\{ \pi + (1 - \pi) S(t_i) \right\}^{1 - \delta_i}
\]
Results: modeling Adverse Regime Change Prediction

On the whole, the model does very well at predicting both risk and immunity.

Figure: PITF Adverse Regime Change in-sample predicted values
High sensitivity but also a high false-positive rate
Ulfelder Mass Killings Ensemble Model

Figure 3.1. ROC Curves for the Ensemble Forecast and Its Components from 10-Fold Cross-Validation

Figure 3.2. Kernel Density Plots of AUC Scores by Forecast Source for Each Fold from 10-Fold Cross-Validation
Ulfelder Mass Killings Ensemble Model

Figure 4.2. Top 30 Estimated Risks of Mass-Killing Onset for 2013. Ensemble forecasts shown in red, component forecasts in grey.
Automated Coding: Textual Analysis By Augmented Replacement Instructions (TABARI)

- ANSI C++, approximately 14,000 lines of code
- Open-source (GPL)
- Unix, Linux and OS-X operating systems (gcc compiler)
- “Teletype” interface: text and keyboard
  - Easily deployed on a server
- Codes around 5,000 events per second on contemporary hardware
  - Speed is achieved through use of shallow parsing algorithms
  - Speed can be scaled indefinitely using parallel processing
- Standard dictionaries are open source, with around 15,000 verb phrases for events and 30,000+ noun phrases for actors
- Coded the 200-million event GDELT dataset without crashing
Integrated Conflict Early Warning System

> Unclassified project funded by DARPA Information Processing Techniques Office
> Funding at $35-million for 2007-2011
> Largest quantitative conflict analysis project since the 1970s
> Objective is real-time forecasting of indicators of political instability in Asia with 6-24 month leads, 70
> Machine-coded event data has proven to be the core methodology for accurate forecasts
> Data covers 1997-present with 8.5-million stories from 27 sources
> Model accuracy has been assessed with a strict split-sample design

Reference:
ICEWS “Events of Interest”

- **Domestic Political Crisis**—Significant opposition to the government, but not to the level of rebellion or insurgency (for example, power struggle between two political factions involving disruptive strikes or violent clashes between supporters)

- **Rebellion**—Organized opposition where the objective is to seek autonomy or independence

- **Insurgency**—Organized opposition where the objective is to overthrow the central government

- **Ethnic/Religious Violence**—Violence between ethnic or religious groups that is not specifically directed against the government

- **International Crisis**—Conflict between two or more states or elevated tensions between two or more states that could lead to conflict
ICEWS Actor Categories

- **gov**: government agents such as the executive, police, and military
- **par**: political parties
- **opp**: armed opposition—rebels and military groups
- **soc**: society in general—civilians, businesses, professional groups
- **ios**: international actors
- **usa**: United States
ICEWS Metrics

Accuracy = \frac{\text{number of correct predictions}}{\text{total predictions made}}

Recall = \frac{\text{number of correctly predicted conflicts}}{\text{total conflicts that occurred}}

Precision = \frac{\text{number of correctly predicted conflicts}}{\text{total conflicts predicted}}
ICEWS Phase 1 Results: LM-ATL Out-of-Sample Results (DARPA Chart)

- Exceeds metrics for the maximum intensity index and 3 instability events: Rebellion, Insurgency, and Ethnic/Religious Violence: Passes Phase 1 gates
- By integrating improved versions of best models from multiple perspectives, team achieves more accurate, precise forecasts than any one model alone
ICEWS Phase 1 Event Data

- 30-gigabytes of text from Lexis-Nexis
- 25 sources
- 8-million stories
- 26-million sentences
  - Only first four sentences coded in each story
- 3-million events
- Generally two orders of magnitude greater than any prior event coding effort
Lockheed “Raven” System

This is ca. 2009
Lockheed iTrace System

This is ca. 2010
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

▶ whether it is 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
At this point, risk invoking the wrath of the Gods of Beamer by switching to document showing GJP IFPs
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.
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)).
Superforecasters

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 improve performance.

- Ensemble methods such as weighting by past performance and “extremizing” forecasts (changing 0.7 to 0.9) appears to improve over individual forecasts, though the robustness of this is still unclear.

- No teams were able to produce an average Beier 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.
Thank you

Email: schrodt735@gmail.com

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

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