Conflict Prediction: Seven Opportunities, Seven Challenges

Philip A. Schrodt

Parus Analytical Systems
schrodt735@gmail.com

European Network for Conflict Research
(ENCoRe)
Essex University
20 September 2013

“Schrodt should do everything in ‘sevens’”

Opportunities

- Totalitarian law of the universe: whatever is not forbidden is mandatory. Prediction *is* scientific [now]
- Data: small, big, fast
- You can't solve everything with more machine cycles, but it ~~never~~ rarely hurts
- Successful large-scale projects: PITF, ICEWS, ACE, ENCoRe
- (Mostly) Convergent models
- Location, location, location
- Open source, open access, open collaboration

Challenges

- Determining credible metrics
- Black swans
- Heterogeneous environments
- Absence of theories indicating what is and is not predictable
- Pournelle's Law: no task is so virtuous that it will not attract idiots
- Ethical concerns
- Internet time vs. journal time

Whatever is not forbidden is mandatory

The old challenge: “Prediction is not scientific”

Whatever is not forbidden is mandatory

The old challenge: “Prediction is not scientific”
Huh????

Whatever is not forbidden is mandatory

The old challenge: “Prediction is not scientific”
Huh????

The new challenge: “Political science (and economics) is not scientific because it can’t predict”

- See various op-eds in *New York Times* and related venues over the past eighteen months. Despite Nate Silver’s predictions.
- See complete suspension of US NSF Political Science program
- The sting of which has been somewhat mollified by subsequent efforts by the same people to suspend the entire US government

Whatever is not forbidden is mandatory

The old challenge: “Prediction is not scientific”
Huh????

The new challenge: “Political science (and economics) is not scientific because it can’t predict”

- See various op-eds in *New York Times* and related venues over the past eighteen months. Despite Nate Silver’s predictions.
- See complete suspension of US NSF Political Science program
- The sting of which has been somewhat mollified by subsequent efforts by the same people to suspend the entire US government

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.

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.

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?

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.

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?

Neither do I.

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.

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.

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

Multi-disciplinary challenges

Big Data: Machine learning researchers routinely use social science data to construct models. Many of these achieve high accuracy in split-sample tests, to the point where these researchers simply assume that things are predictable.

IARPA ACE "Good Judgment Project" (Tetlock): While *most* forecasters do no better than chance, a small number of "super forecasters" perform significantly better than chance over long periods of time and large numbers of questions. Furthermore these individuals exhibit common characteristics and strategies, and to a limited extent, these can be taught. Forthcoming article in *Economist: The Year 2014*

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)

Data!


Science
11 February 2011 | $10
scientific
data
information
climate
research
science
new
analysis
visualization
researchers
access
many
human
knowledge
example
AAAS


PRIO


ACLED
Armed Conflict Location & Event Dataset


Minorities at Risk

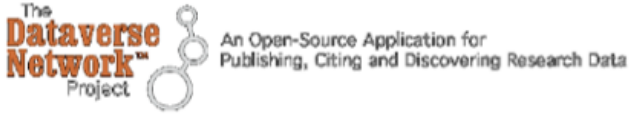

CORRELATES OF
WAR


UPPSALA
UNIVERSITET


PENNSTATE
1855


UCDP
GEOREFERENCED EVENT DATASET


CIRI Human Rights Data Project
www.humanrightsdata.org


The
Dataverse
Network
Project
An Open-Source Application for
Publishing, Citing and Discovering Research Data


Polity
IV


GTD
Global Terrorism Database


Universität
Konstanz

Though this may be going a little far...

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

Computing power

Computing Power

Control Data Corporation 3600
(ca.1965)
32 K (48-bit) RAM memory
1 processor
~1-million operations per second
Output: line printer

Computing Power

Control Data Corporation 3600
(ca.1965)
32 K (48-bit) RAM memory
1 processor
~1-million operations per second
Output: line printer

Penn State High Performance Computing Facility
15 cluster computers
100 to 2000 2.66 Ghz processors in each cluster
~50 Gb RAM accessible to each processor
130 Tb disk space
4 interactive visualization rooms

Computing Power

Control Data Corporation 3600
(ca.1965)
32 K (48-bit) RAM memory
1 processor
~1-million operations per second
Output: line printer

Penn State High Performance Computing Facility
15 cluster computers
100 to 2000 2.66 Ghz processors in each cluster
~50 Gb RAM accessible to each processor
130 Tb disk space
4 interactive visualization rooms

Motorola Razr
16 Gb RAM memory
Dual processor
~500-million operations per sec
540 x 860 color display

Open Source Software

MySQL

python

mozilla Firefox

ubuntu

source forge

T R

perl

Ruby PROGRAMMING Language

WordPress.com

APACHE HTTP SERVER

R

LibreOffice The Document Foundation

GNU run free run GNU

stackoverflow

FREE SOFTWARE FOUNDATION

Computationally-intensive methods

- Bayesian estimation using Markov chain Monte Carlo methods
- Bayesian model averaging ("*AJPS*-as-algorithm")
- random forest models
- large-scale textual databases
- machine translation
- geospatial visualization
- real-time automated coding
- remote sensing data such as nightlight density

Large Scale Conflict Forecasting Projects

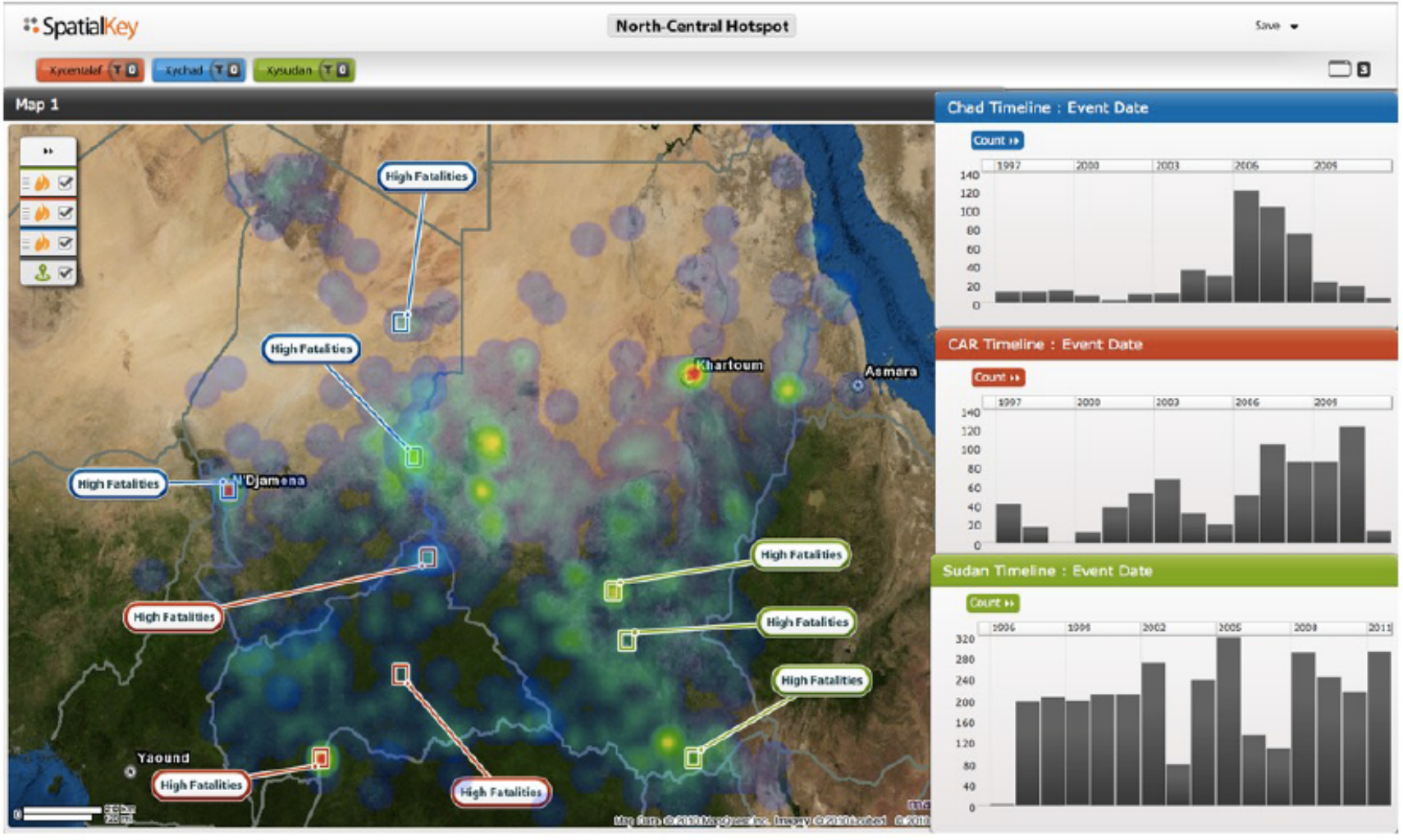
- State Failures Project 1994-2001
- Joint Warfare Analysis Center 1997
- FEWER [Davies and Gurr 1998]
- Center for Army Analysis 2002-2005
- Swiss Peace Foundation FAST 2000-2008
- Political Instability Task Force 2002-present
- DARPA ICEWS 2007-present
- IARPA ACE and OSI
- Peace Research Center Oslo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

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 generally do poorly except as *post hoc* "explanations."

Location, location, location!

ACLED Geospatial



UCDP Geospatial

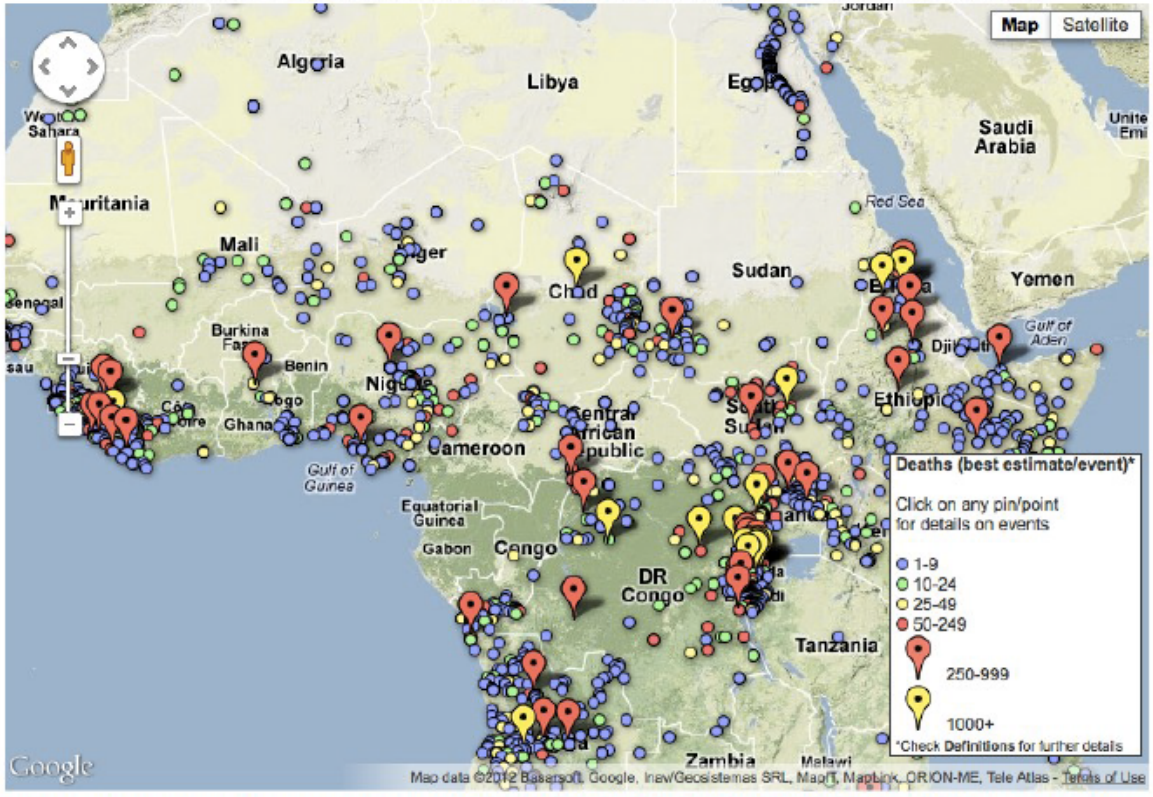
Welcome to the UCDP GED - Uppsala Conflict Data Program's Georeferenced Event Dataset

The interactive map below covers all the locations geocoded as part of GED. ***Drag and move the red and grey thumbs of the slider below to filter for particular intervals, years, or even particular months or days.*** *Use the filters to select only one of the three UCDP types of violence. Clicking on each dot or pin will display information regarding events taking place in those respective locations. Download data here.*

1989 2010

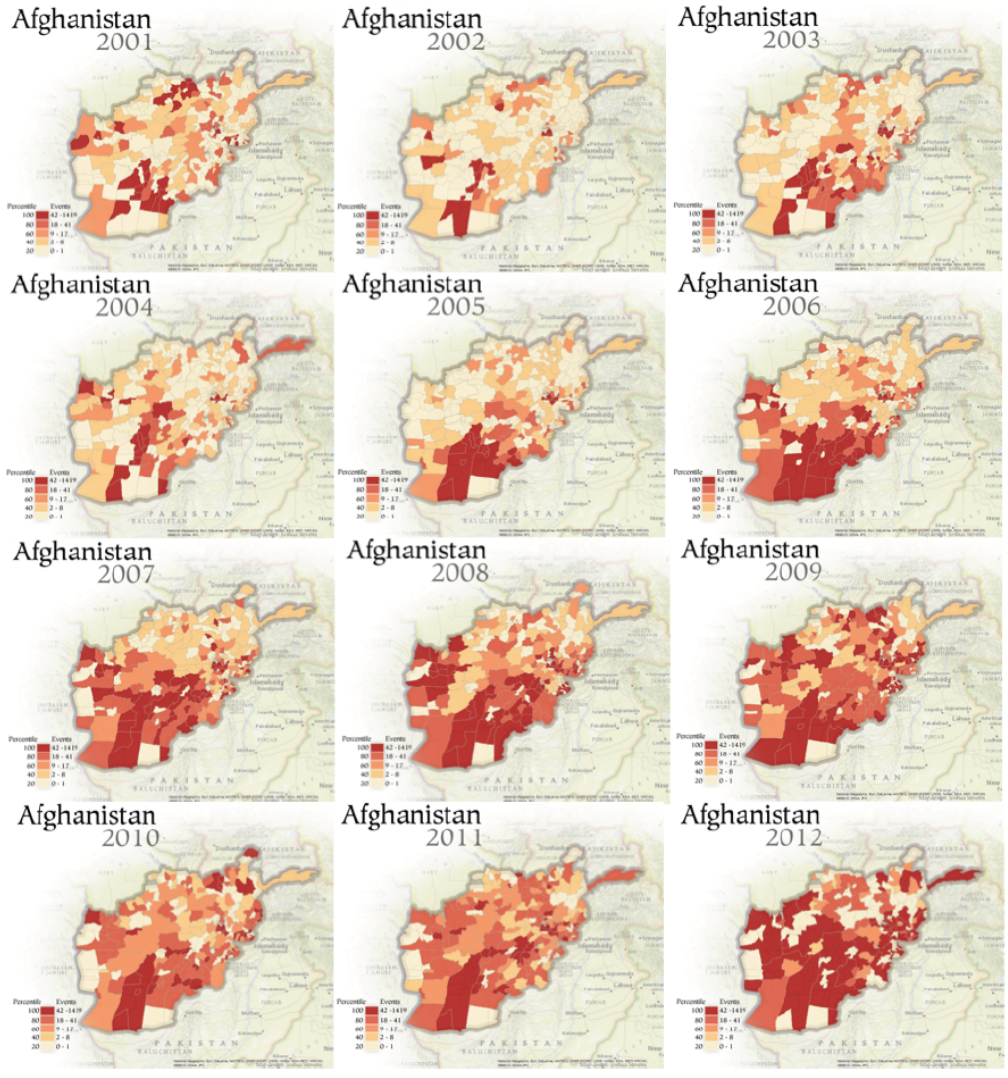
Filter for:

All violence | State-based conflicts only | Non-state conflicts only | One-sided conflicts only



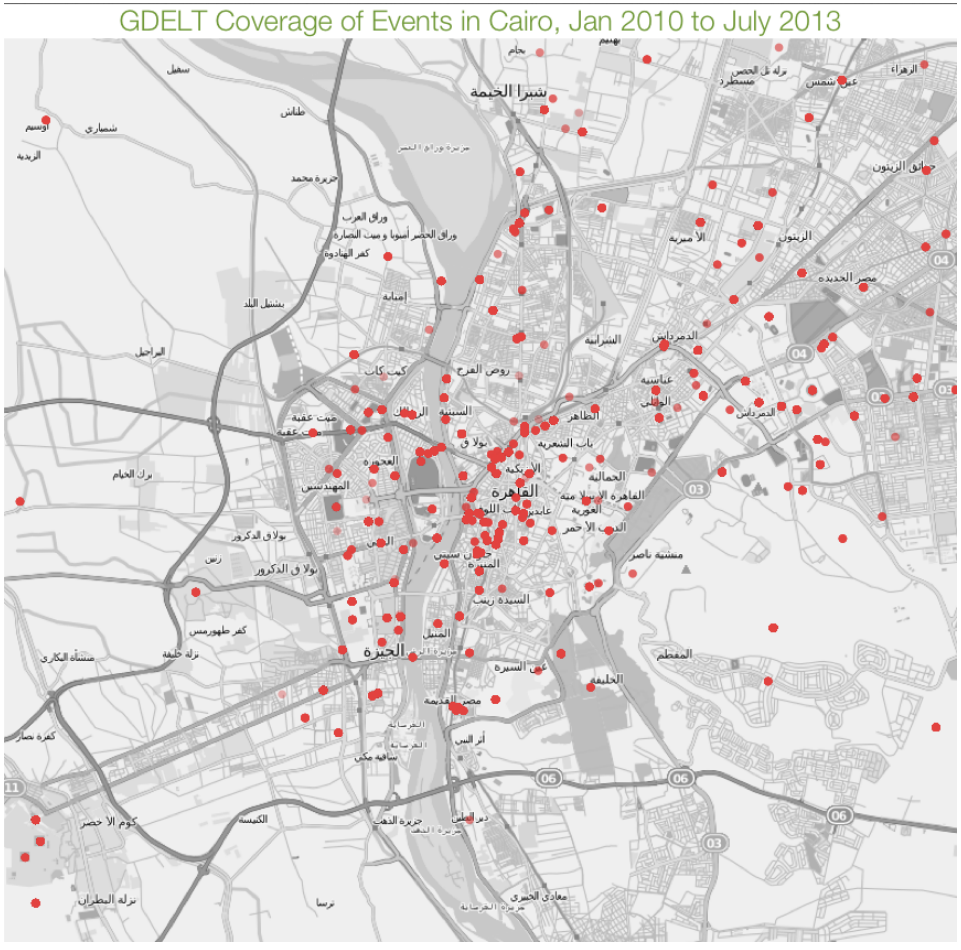
The GED is the product of two and a half years of work at the **Department of Peace and Conflict Research, Uppsala University**. The UCDP GED contains conflict data disaggregated spatially and temporally down to the level of the individual incidents of violence. For more details please see the About UCDP link above.

GDELT: Afghanistan, District-level Violence



Source: Jay Yonamine and Joshua Stevens, Penn State

GDELT: Cairo protests



Source: David Masad and Andrew Halterman of Caerus Analytics.

Open source, open access, open collaboration

- There is a strong if incomplete norm towards open sharing of data and methods
  - Unintended consequence: PITF "forecasting tournament" cannot be published in a major journal because it used proprietary data—the baseline data has 2,700 variables—that cannot be archived in replication sets. The results are, however, still available on SSRN.
  - The inability to share source texts is clearly a concern in news-report-based datasets such as GDELT and MID.
- By all available evidence, US government forecasting projects are using similar methodologies to those available in open sources; in fact they are probably lagging somewhat behind this
- We now have significant NGO and academic work, and an international "epistemic community" has developed around the topic.

CHALLENGES

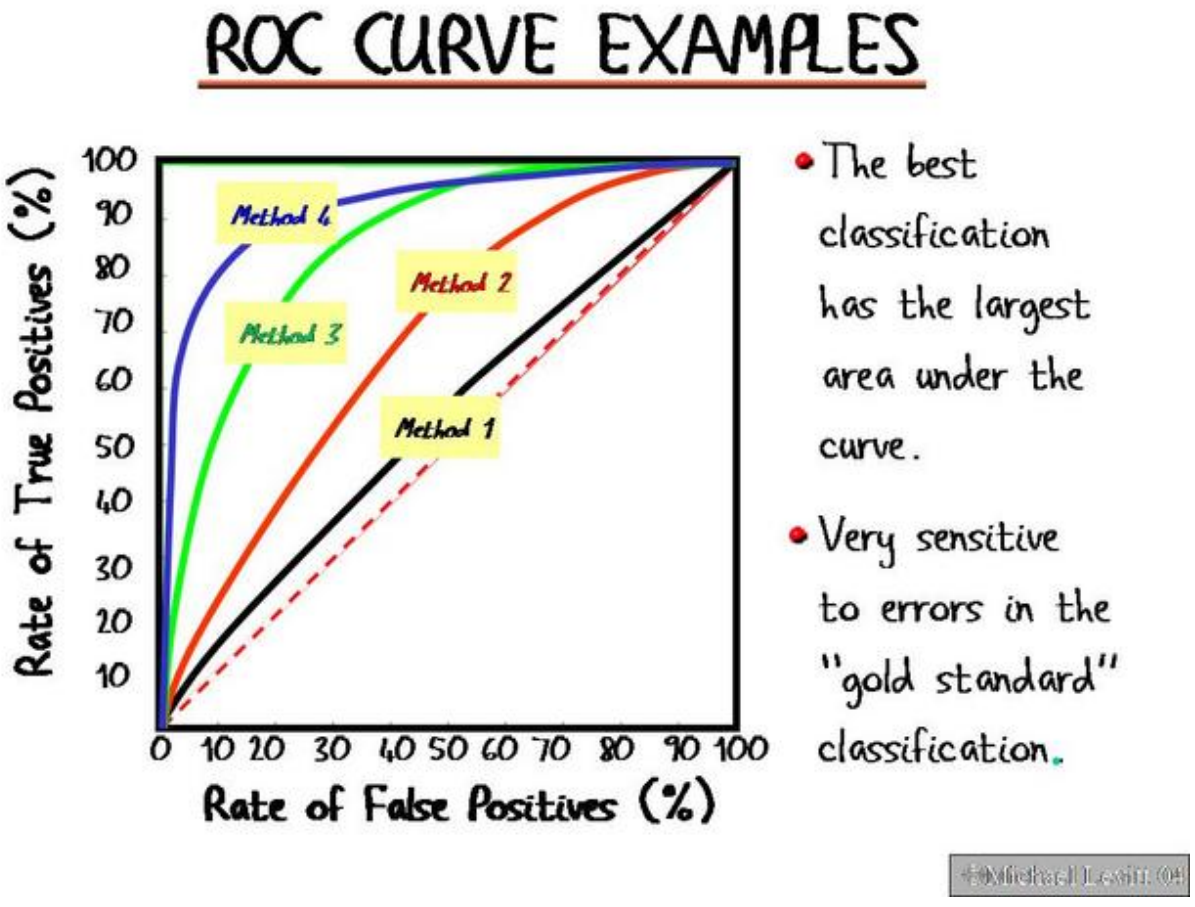
Metrics

Classification Matrix

Relationships among terms

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

ROC Curve



Source: http://csb.stanford.edu/class/public/lectures/lec4/Lecture6/Data_Visualization/images/Roc_Curve_Examples.jpg

Separation plots

TABLE 4 Rearrangement (and Coloring) of the Data Presented in Table 1 for Use in the Separation Plot

Country	Fitted Value ($\hat{p}$)	Actual Outcome (y)
B	0.364	0
F	0.422	1
D	0.728	0
A	0.774	0
E	0.961	1
C	0.997	1

FIGURE 2 Separation Plot Representing the Data Presented in Table 1

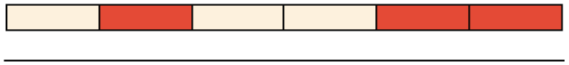
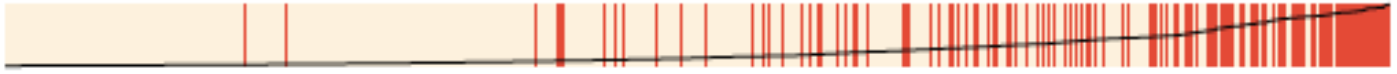


FIGURE 3 Separation Plot for a Larger Data Set

FIGURE 4 Adding a Graph of $\hat{p}$ to the Separation Plot



The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models

Brian Greenhill[1], Michael D. Ward[2], Audrey Sacks[3]

Article first published online: 20 JUN 2011

DOI: 10.1111/j.1540-5907.2011.00525.x

Issue

American Journal of Political Science

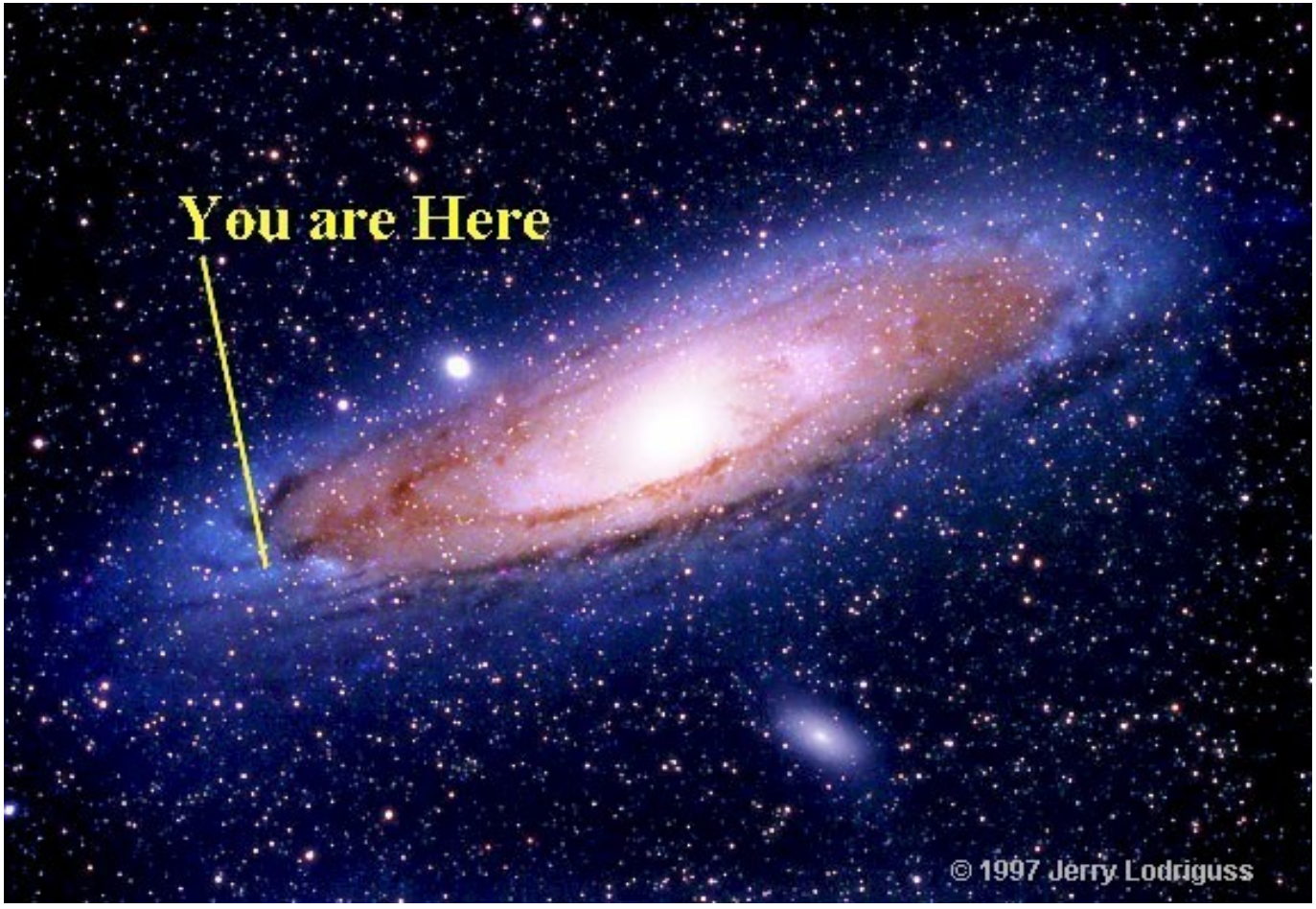
Volume 55, Issue 4, pages 991–1002, October 2011

And wait, there's still more!

- Recall / True Positive Rate/Sensitivity
- Precision / Positive predictive value (PPV)
- Specificity / True Negative Rate
- F1 score: harmonic mean of precision and recall
- Beier scores
- Posterior probabilities
- Proportional reduction of error or entropy

Black swans

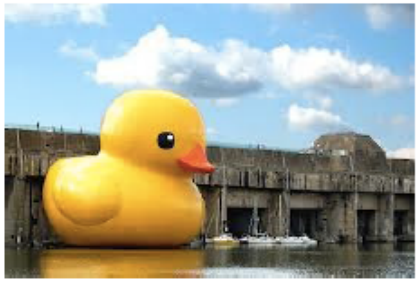
Ideal forecasting targets are neither too common nor too frequent



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

The Forecasting Zoo

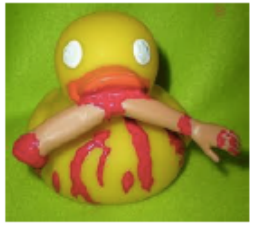
Ducks can be interesting...



Size

Variety

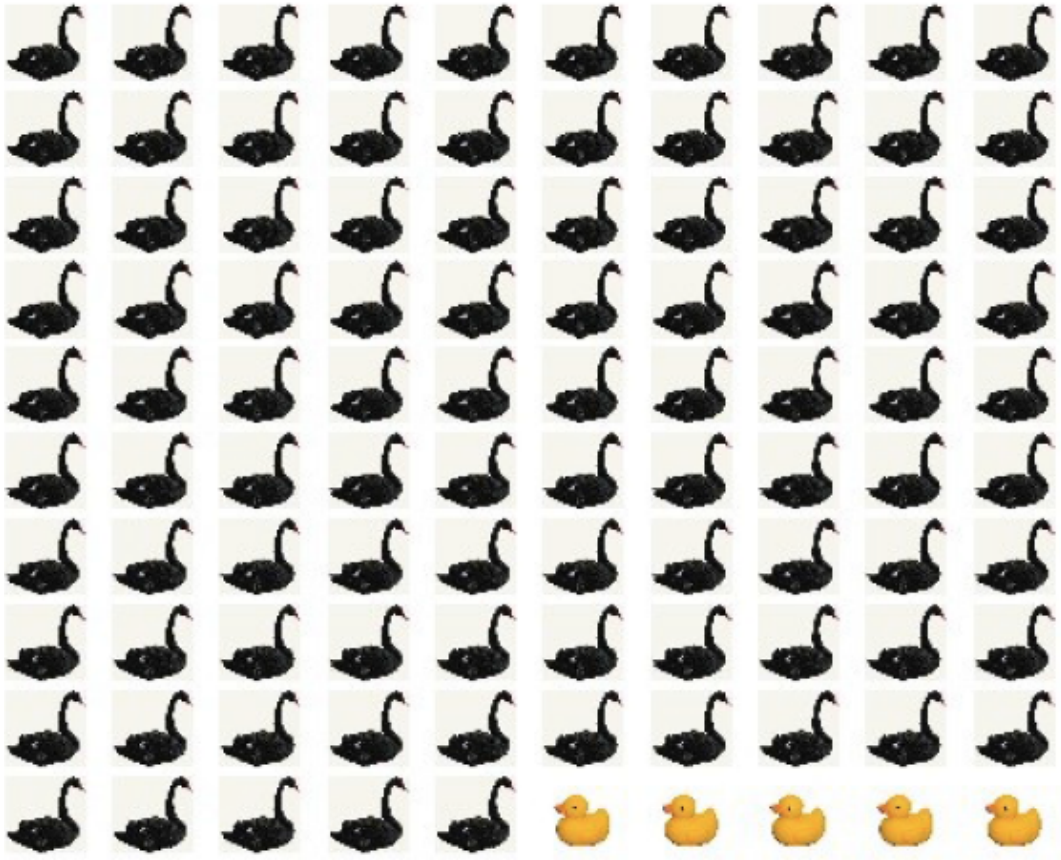
Quantity



Suspicious behaviors

And this is going too far…

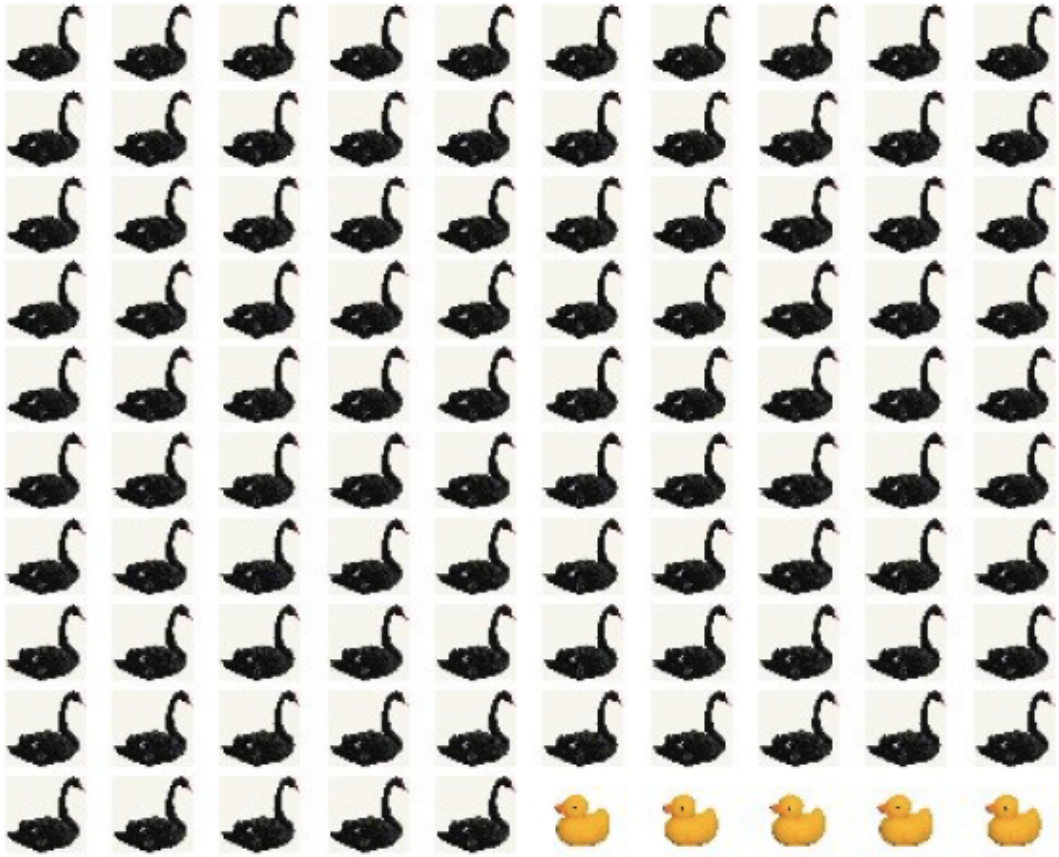
DARPA-World!



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.

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?
- Will changes in the technological environment—internet, UAVs, various monitoring technologies—change probabilities?
- 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.

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

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
- the end of the 410-year supernova peace

Irreducible sources of error

- 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%
- 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."
- Quasi-random structural error: Complex and chaotic deterministic systems behave as if they were random under at least some parameter combinations

Balancing factors which make behavior predictable

- Individual preferences and expectations, which tend to change very slowly
- Organizational and bureaucratic rules and norms
- 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

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~~ 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.

Ethical concerns

- Thus far, we've generally had the luxury of no one paying attention to any of our predictions : what if governments do start paying attention?
  - "Policy relevant forecast interval" is around 6 to 24 months
  - USAID/FAO famine forecasting model
  - It is *possible* that our models could become less accurate because crises are being averted, but I don't see that happening any time soon.
- Difficulties in getting *anyone*, including experts (see Kahneman, Tetlock), to correctly interpret probabilistic forecasts
- Possible impact on sources
  - Local collaborators
  - Journalists (cf. Mexico)
  - NGOs to the extent we are using their information

Journal time versus internet time

Journal time

Nate Jensen's September 2013 *ISQ* paper [13 Sept 2013 blog entry]

- Presumably written spring/summer 2008
- APSA IPE section "best paper" award Fall 2008
- Four rejects and R&Rs 2009-2011, with each new submit/R&R requiring *re-analysis*. Dr. Jensen is clearly a *very* patient and/or persistent individual. . . Prediction papers need to be updated: Schrodt and Gerner *APSR* 2000 also had a five-year delay
- Accepted by *ISQ* June 2012
- Published September 2013
- Available in open access (and hence to policy-makers and journalists, who generally cannot access JStore): perhaps 2015?

Observation: This is industrial archeology, not scholarly communication. In the prestige journals, one sees the field as it existed five years ago. Due to restricted access, the policy community sees the field as it existed seven to nine years ago.

Journal time

Nate Jensen's September 2013 *ISQ* paper [13 Sept 2013 blog entry]

- Presumably written spring/summer 2008
- APSA IPE section "best paper" award Fall 2008
- Four rejects and R&Rs 2009-2011, with each new submit/R&R requiring *re-analysis*. Dr. Jensen is clearly a *very* patient and/or persistent individual. . . Prediction papers need to be updated: Schrodt and Gerner *APSR* 2000 also had a five-year delay
- Accepted by *ISQ* June 2012
- Published September 2013
- Available in open access (and hence to policy-makers and journalists, who generally cannot access JStore): perhaps 2015?

Observation: This is industrial archeology, not scholarly communication. In the prestige journals, one sees the field as it existed five years ago. Due to restricted access, the policy community sees the field as it existed seven to nine years ago.

Internet time: GDELT

- August 2012: Initial data completed and provided to Penn State for internal evaluation
- September 2012: First GDELT Hackathon with about a dozen PSU political science and geography students
- March 2013: Static GDELT released on PSU site in conduction with ISA
- April 2013: Syria, Afghanistan graphics in *Guardian, Foreign Policy, New Scientist*
- June 2013: UT/Dallas server operational with daily updates
- July 2013: gdeltblog.wordpress.com
- August 2013: Beieler/Stevens protest graphic receives 150,000+ views, including a Chelsea Clinton tweet

Internet time: GDELT

- 19 September 2013: News of Javier Osorio's "Eventus ID" system: event coding in Spanish with time-series visualization of Mexican drug gang violence
- 28-29 September 2013: Second GDELT Hackthon; over 40 participants, 25+ from outside Penn State

Thank you

Email: schrodt735@gmail.com

Data: http://gdelt.utdallas.edu

Blog: http://gdeltblog.wordpress.com

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

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