PL SC 597I: Event Data Analysis
Pennsylvania State University: Spring 2013

Philip A. Schrodt
227 Pond Laboratory
schrodt@psu.edu

Course Description

> "Anything that is in the world when you're born is normal and ordinary and is just a natural part of the way the world works. Anything that's invented between when you're fifteen and thirty-five is new and exciting and revolutionary and you can probably get a career in it. Anything invented after you're thirty-five is against the natural order of things."
> Douglas Adams, *The Salmon of Doubt* (2002)[1]

> "Don't undertake a project unless it is manifestly important and nearly impossible."
> Edwin Land[2]

Classical "atomic" event data—nominal or ordinal codes recording the interactions between international actors as reported in the open press—break down complex political activities into a sequence of basic building blocks (e.g., comments, visits, grants, rewards, protests, demands, threats, and military engagements). Composite event data extend this by coding a variety of features of an event, for example the location, number of individuals involved, and the reasons the event occurred. This graduate-level course will provide an in-depth survey of contemporary political event data analysis in both modes and the tools required to generate and utilize those data, as well as issues in the analysis of event data with an emphasis on conflict forecasting.

Course Objectives

By the end of this course, you should be technically competent for the following tasks:

- Familiarity with the WEIS, COPDAB, GDELT, CAMEO and IDEA atomic data sets and coding ontologies
- Familiarity with a number of composite event data sets, including COW, ACLED, and the UCDP/PRIO family of data sets
- Familiarity with large-scale text processing—particularly reformatting—in a Unix environment, including high-performance computing environments
- Basic knowledge of automated event coding, specifically in the TABARI environment, including the enhancement of verb and actor dictionaries

[1] Better known as author of *The Hitchhiker's Guide to the Galaxy.*

[2] College drop-out and inventor of the Polaroid camera, a quaint mechanical/chemical device that produced "instant" photographs. Not as clumsy or random as a cell phone, an elegant device for a more civilized age. Nonetheless typically used to record kittens playing with yarn and birthday parties for three-year-olds.

- Text classification methods for determining whether a news report contains a codeable event
- Scaling and measurement issues relevant to using event data in conventional statistical analysis
- Basic issues in time series analysis with a specific focus on models for forecasting political conflict

Problem-Based Learning

> "A special transmission outside the scriptures;
> No dependence on words and letters;
> Direct pointing to the mind."
> Bodhidharma, First Patriarch of Zen, ca. 5C CE

> "*Try* not. Do, or do not: There is no *try*"
> Yoda, *The Empire Strikes Back*[3]

This course will make extensive use of the "problem-based learning" (PBL) instructional method, one of the various "active learning" modes that is ever-so-gradually replacing the "sage on the stage" model that dates from the period before the development of printing.[4] This emphasizes

- open-ended real-world problems: there is no single correct answer and the exercises involve all the messiness of problems encountered in actual professional environments, rather than the neatly compartmentalized approach of "toy" problems that have known solutions;
- problem solving is done in [rotating] teams: this, of course, opens the possibility of free-riding and some people—which is to say introverts, which is to say most of you—don't *like* working in teams, but the bottom line is that almost all quantitative research is now done collaboratively, and by working in teams, we can explore more difficult problems more efficiently than we can do in individual projects;
- the instructional objective is to improve *problem-solving skills*, not the mastery of a fixed canon of literature. We will be looking at some literature in the initial part of the course—and this most decidedly over-samples from the experience of the KEDS project—and I expect you to continue exploring relevant literature as we move into the PBL phase, but that literature—as occurs in an actual research project—is driven by the need to solve a specific problem, not by the need to make sure that you've read X, Y and neo-Z. To say nothing of that fact that any literature review I give will be out-dated within a year or so, and when you need to get up to date you will be using the Web anyway.
- we will be doing some of this messy work in the classroom: the current buzzword for this—along with the too-cute "guide on the side"—is "flipping" the class format so that it emphasizes material that in other pedagogical approaches is considered "homework." Our classroom is not equipped to fully take advantage of this[5] but with laptops we should be able to approximate it.

[3] http://www.youtube.com/watch?v=BQ4yd2W50No

[4] As well as ever-so-gradually displacing the "How much do I dare participate in discussion given that I haven't actually done the reading???" approach that dates to the dawn of the university seminar format, about 1200 CE (yes, really: the *disputatio*.)

[5] But hey, just how much can we complain when the Osmond Building has a bronze bust of Nikola Tesla at the door?

- This is a course in *political methodology*, and as a consequence emphases the acquisition of practical skills for the analysis of political behavior, not the theoretical understanding of political behavior. Both are important—in the phrasing of Larry Bartels, in a small village, a witchdoctor must also be a good farmer[6]—but in this particular course, we're focusing almost exclusively on technique.

Elaborating on the above point: Scientific progress can come from three sources: theory-driven, observation-driven, and technique-driven. Most accounts of the philosophy of science prevalent in the social sciences emphasize the first two to the exclusion of the third, but in fact many of the major changes in the natural sciences were the results of changes in technique: the telescope and microscope completely changed astronomy and biology, and changes in theory followed the introduction of the method, rather than theory driving the method[7].

For the past two decades, event data analysis has been driven—in some cases very dramatically—by two exponential changes: the increase in machine-readable text (now available more or less for free), and the increase in computing power that provides for the efficient processing of that text. Humans are able to code about six to ten events per hour; machine coding on a single processor currently works at about 5,000 events per second—itself an increase by a factor of about 2-million over human coding—and through simple parallel processing this can be scaled indefinitely with a near linear increase in speed.

On the *theoretical* side, most event data analysis has been applied primarily to a single—though very general—problem: conflict forecasting, and consequently this is where most of the existing literature (and most of my experience) is found. Conflict forecasting remains an important issue—particularly in applied settings—but in principle it should be possible to extend the method into other issues of theoretical concern.

But what issues? This is something we will explore, but two points

- As graduate students, the optimal topic for you to focus on is whatever will be of great interest to a large number of potential employers four to five years from now. Whatever is interesting *now* will get you through qualifying exams, but whatever is interesting in four years will get you a job.
- As social scientists, the optimal task to be working on is to be found in that small set of problems where we have the capability of solving a problem but have not done so already. This is the proverbial "cutting edge" that steers clear of topics where only marginal additional knowledge can be found,[8] and those problems that are essentially still impossible.

[6]Translation: in most environments, a political methodologist must also have a good understanding of politics and the contemporary issues of interest to the political science community as a whole.

[7]Theory sometimes drives methodological developments, but not in these two instances, nor lots of others

[8]Such problems can certainly generate *publications*—and do, in large quantities. But the fact that about 90% of political science publications are only rarely if ever cited would suggest that the contributions to knowledge are limited in such situations

Computer programming

> "That tune took me fifteen minutes to write. Fifteen minutes and twenty-five years in the business."
> George Shearing on *Lullaby of Birdland*

At present, the preparation of event data requires at the use of customized "filters" to convert the machine-readable text to a form that can be processed by an automated coding program: there are no "off-the-shelf" solutions for this, nor are there likely to be for several years. In addition, working with unstructured data and algorithmic methods generally will get you into problems which can be solved *much* more easily—and often orders of magnitude more quickly in terms of machine time—than they can be solved in database or statistical systems (which were themselves implemented as computer programs). So to do this sort of thing effectively, you need to learn some programming.

In this course I will be focusing on using Python[9]. Python is a very robust, open source, platform-independent language that can readily process text at gigabyte levels and involves codes that is typically about one-tenth the size of comparable programs in C/C++. Most of the control structures in Python are similar to those of other structured languages such as C/C++, Java, C# and perl[10] as well as the [very different] programming environments of Stata and R, and if you've not had prior exposure to computer programming, this will provide a good introduction.

About a third of the people in the class already have basic Python skills and I will be providing opportunities to apply these in real-world problems. For the remainder,[11] we will be using the Udacity online course "CS101: Introduction to Computer Science" (`https://www.udacity.com/course/cs101`) which teaches basic Python skills in the context of building a simple Web browser. Following the "flipped" classroom model, you will watch these lectures on your own and I will provide time in class for discussion and questions.

Much of the work we will be doing for this course will be done in the Unix operating system, often in the PSU high-performance computing systems. We will be working at a general level so the particular variety of Unix—both "mainframe" implementations such as BSD and AIX and the personal computer implementations Linux and Apple's OS-X "Terminal" environment—will all work equivalently. The third week of the course-within-a-course will focus on Unix and some of the utilities most widely used in text processing.

Requirements and Evaluation

1. Attendance and active participation, particularly in the problem-solving discussions. This presumably goes without saying

[9] `http://xkcd.com/353/`

[10] The primary alternative to Python is perl, an older language which is optimized for text processing, and focuses on "regular expressions," an exceedingly powerful method of processing character sequences of any type. In earlier versions of this course I used perl rather than Python as the core language—and if you already know perl, that will be quite sufficient for the coursework—but perl has a set of notoriously idiosyncratic features—`http://xkcd.com/1171/`—and is being displaced by Python is many applications.

[11] Students were provided with a non-programming option but no one chose to take that.

2. Problem-based learning exercises: these are, well, open-ended and the exact nature of what needs to be done on each problem may not be evident until we get into them
3. A conference-quality paper using event data to deal with either a methodological or substantive problem. These can be collaborative and they can also be based on a PBL exercise. Additional details will be provided in class.

All readings are on the web or available in PDF format in the *Lessons* section of the course's ANGEL site.[12] I will also be using ANGEL to post messages to the class as a whole, so if you haven't set ANGEL to forward messages to an email account that you regularly read, please do so.

There are no Twitter, FaceBook, YouTube or LOLCats tags or sites devoted to this class.

How to find me

It isn't difficult: except when I'm out of town for conferences or research consultations, I'll usually be in the office pretty much 9-6 daily. There are imaginative office hours posted outside my door, but if you need a block of time, better to contact me by email to set up a specific appointment since these might be superceded by other meetings. If my door is open, I'm fair game. If the door is closed, feel free to knock (with 8-foot windows, I rarely have the lights on, but I may be in anyway) but I may be in the middle of something and will ask you to come back later: do not be offended.

Email: schrodt@psu.edu

Phone: 814-863-8978

Departmental policies

Academic Dishonesty

The Department of Political Science, along with the College of the Liberal Arts and the University, takes violations of academic dishonesty seriously. Observing basic honesty in one's work, words, ideas, and actions is a principle to which all members of the community are required to subscribe.

All course work by students is to be done on an individual basis unless an instructor clearly states that an alternative is acceptable. Any reference materials used in the preparation of any assignment must be explicitly cited. Students uncertain about proper citation are responsible for checking with their instructor.

In an examination setting, unless the instructor gives explicit prior instructions to the contrary, whether the examination is in class or take home, violations of academic integrity shall consist but are not limited to any attempt to receive assistance from written or printed aids, or from any person or papers or electronic devices, or of any attempt to give assistance, whether the one so doing has completed his or her own work or not.

[12] ANGEL is the course management system currently used by Penn State.

Lying to the instructor or purposely misleading any Penn State administrator shall also constitute a violation of academic integrity.

In cases of any violation of academic integrity it is the policy of the Department of Political Science to follow procedures established by the College of the Liberal Arts. More information on academic integrity and procedures followed for violation can be found at:
`http://www.la.psu.edu/CLA-Academic_Integrity/integrity.shtml`

Note: In addition to these policies, it is my practice to run all major written assignments through `TurnItIn.com` to check for possible uncited material. IMHO (and practice), it is a good habit to run your own writing through this as well, just to make sure you haven't missed anything: you can get access to it through your Penn State credentials.

Note to students with disabilities:

Penn State welcomes students with disabilities into the University's educational programs. If you have a disability-related need for reasonable academic adjustments in this course, contact the Office for Disability Services. For further information regarding policies, rights and responsibilities please visit the Office for Disability Services (ODS) Web site at: `www.equity.psu.edu/ods/`. Instructors should be notified as early in the semester as possible regarding the need for reasonable accommodations.

Week 1: 11 Jan 2013
Introduction to Course; Historical Event Sets: WEIS, COPDAB, BCOW

Prior to the advent of machine-coded event data, the most commonly used event data sets for international relations research were Azar's (1982) Conflict and Peace Data Bank (COPDAB) and McClelland's (1976) World Event Interaction Survey (WEIS). Both data sets attempt to code all publicly-reported interactions by all states and some non-state actors. COPDAB includes the period between 1948 and 1978 and is available from the Inter-University Consortium for Political and Social Research (ICPSR); the public-domain WEIS set at the ICPSR covers 1966 to 1978 but a "grey set" extends this into the early 1990s. A third useful—but far less utilitized—data set is Russell Lang's BCOW (Behavioral Correlates of War) which anticipates many of the extensions in the 21st-century coding schemes.

Readings:

Schrodt and Gerner, *AEID*, Chapter 1, Appendix

Schrodt, Philip A. (2012): Precedents, Progress, and Prospects in Political Event Data. *International Interactions*38:4, 546-569 (*http://dx.doi.org/10.1080/03050629.2012.697430*)

Additional readings

Generically useful web site: `http://eventdata.psu.edu`

Azar, Edward E. , Stanley H. Cohen, Thomas O. Jukam and James M. McCormick. 1972. "The Problem of Source Coverage in the Use of International Events Data." *International Studies Quarterly* 16, 3: 373-388

Azar, Edward E. 1980. "The Conflict and Peace Data Bank (COPDAB) Project." *Journal of Conflict Resolution* 24:143-152.

Leng, Russell J. and J. David Singer. 1988. "Militarized Interstate Crises: The BCOW Typology and Its Applications." *International Studies Quarterly* 32, 2: 155-173.

Howell, Llewellyn D. 1983. "A Comparative Study of the WEIS and COPDAB Data Sets." *International Studies Quarterly* 27: 149-159.

Reuveny, Rafael, and Heejoon Kang. 1996. "International Conflict and Cooperation: Splicing the COPDAB and WEIS Series." *International Studies Quarterly* 40,2:281-305.

ICPSR MANUALS FOR WEIS, COPDAB AND BCOW:

> Azar, Edward. 1993. *Conflict and Peace Data Bank (COPDAB), 1948-1978.* (ICPSR 7767, Third Release.) Ann Arbor: Inter-University Consortium for Political and Social Research.
> **Note:** this was originally on 80-column punch cards; I've produced a PDF of the resulting text file. Page breaks don't line up all that well but otherwise it is readable.
>
> McClelland, Charles A. 1976. *World Event/Interaction Survey Codebook.* (ICPSR 5211). Ann Arbor: Inter-University Consortium for Political and Social Research.
>
> Leng, Russell J. 1987. *Behavioral Correlates of War, 1816-1975.* (ICPSR 8606). Ann Arbor: Inter-University Consortium for Political and Social Research.

Weeks 2 and 3: 18 and 25 Jan 2013
Event Coding Ontologies: CAMEO and IDEA

The classical WEIS and COPDAB coding systems had a number of weaknesses even when used for human coding, and these were exacerbated when applied to machine coding. In the last decade, two systems have superceded these: the KEDS-project CAMEO, which was originally designed for the study of international mediation but has a number of general properties designed to facilitate 21st-century automated coding, and VRA's IDEA, which was designed as a superset of all widely-used coding systems, as well as extending coding to deal with issues such as natural disasters.

The CAMEO system will be used for most of the remainder of the course, so before we get into automated coding, we are going to do some *serious* manual coding to get a handle on these systems: this will involve a combination of in-class and out-of-class coding exercises on Reuters and Agence France Press (AFP) downloads, which will be provided.

I'll spend the last part of Week 3 on a general introduction to programming and Unix.

Readings:

Deborah J. Gerner , Philip A. Schrodt, Ömür Yilmaz, and Rajaa Abu-Jabr. 2001. "Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions." American Political Science Association, Boston, August 2002.[13]

Philip A. Schrodt, Ömür Yilmaz, Deborah J. Gerner and Dennis Hermrick. 2008. "Coding Sub-State Actors using the CAMEO (Conflict and Mediation Event Observations) Actor Coding Framework." International Studies Association, San Francisco, March 2008.

Doug Bond, Joe Bond, Churl Oh, J. Craig Jenkins and Charles Lewis Taylor. 2003. "Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development. " *Journal of Peace Research* 40, 6: 733-745 (2003)

Additional resources

The *CAMEO Codebook* is available on the ANGEL site.

There are about a dozen additional papers on IDEA at `http://vranet.com/papers.html`. More generally, `http://vranet.com/` gets you to the VRA web site, which has a variety of information about the VRA approach.

Unix reference sites

There are an abundance of sites on the web that provide information on Unix at every imaginable level of detail; here are three that can get you started.[14]

`http://kb.iu.edu/data/afsk.html`

Basic introduction to the commonly used commands.

`http://www.math.utah.edu/lab/unix/unix-commands.html`

Just another nice reference site for basic Unix commands.

`http://www.calpoly.edu/ rasplund/script.html`
`http://help.unc.edu/213`

These are introductions to Unix shell scripting—which is similar to a do-file in Stata or a script in R—which you need to run HPC programs. The HPC scripts are very simple—just copy their examples—but scripting in general is one of the most powerful features of Unix and provides a way of systematically automating complex tasks.

[13] A shorter version of this eventually appears in print as Deborah J. Gerner, Philip A. Schrodt and Ömür Yilmaz. 2008. "Conflict and Mediation Event Observations (CAMEO): An Event Data Framework for a Post Cold War World." in Jacob Bercovitch and Scott Gartner, eds. *International Conflict Mediation: New Approaches and Findings.* New York: Routledge.

[14] All were accessed 23 January 2013 and were found by Googling "unix reference"

Week 4: 1 Feb 2013
Automated Coding: TABARI

At the present time, the TABARI program is the only open-source automated coding program available, and it is also the one with which I can most familiar. The basic principles underlying TABARI are, nonetheless, likely to extend to future coding systems, and in any case, TABARI can be used now.

Readings:

Schrodt. *TABARI Manual*. Chapters 2, 7, Appendix F (Joe Pull's "Ode to Coding")

Schrodt and Gerner, *AEID*, chapter 2

Schrodt and David VanBrackle. 2013. "Automated Coding of Political Event Data" in V.S. Subrahmanian (ed.), Handbook of Computational Approaches to Counterterrorism, New York: Springer Science Business Media

Additional readings

Schrodt, Philip A. and Deborah J. Gerner. 1994. "Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982-92." *American Journal of Political Science* 38, 3:825-854 .

Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. 1994. "The Machine Coding of Events from Regional and International Sources." *International Studies Quarterly* 38:91-119.

King, Gary and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57,3: 617-642. (This discusses the VRA coder, which is a fully operational proprietary coder)

Schrodt, Philip A. 2006. Twenty Years of the Kansas Event Data System Project. *The Political Methodologist* 14,1: 2-8. (rather chatty and anecdotal narrative of the first two decades of the KEDS project).

Week 5: 8 Feb 2013
Automated Coding: Machine-Readable Reports, Text Filters, Reformatting, and Named Entity Recognition

The availability of machine-readable reports is every bit as important to the contemporary development of event data as the development of automated coding. While originally limited to a small number of proprietary sources such as Lexis-Nexis (which remain important, if endlessly aggrevating...), open Web-based HTML sources and RSS feeds are likely to become increasingly important in the future. However, these are not standardized and a fair amount of processing is required to get from the "raw" information to something that can be processed. We will also deal with the general topic of

named-entity recognition and a simple implementation of that which we have used at Penn State, as well as the closely related issue of automated geolocation, which is a still-developing technology with the potential of dramatically increasing the amount of data we have available that has geographical coordinates.

Readings:

LingPipe NER Tutorial: `http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html`

David Nadeau and Satoshi Sekine. 2007 "A survey of named entity recognition and classification" [ANGEL]

Lev Ratinov, Dan Roth, Doug Downey and Mike Anderson. 2011 "Local and Global Algorithms for Disambiguation to Wikipedia" [ANGEL]

Week 6: 15 Feb 2013
Composite Event Data Sets

The shift in the focus of conflict studies from the relatively straightforward—at least as far as coding is concerned—international conflicts to substate conflicts, along with the proliferation of information available on the web (and a lot of research funding...) and an interest in geospatial methods has led to the development of a large number of new "composite" conflict data sets. This week we will collectively explore several of these.[15]

- Terrorism databases: GTD and ITERATE
- UCDP/PRIO collections
- ACLED
- full-text databases with historical coverage (Lexis-Nexis, ProQuest, etc)
- PITF atrocities/mass killings databases: Ulfelder/Schrodt and Valentino

Your group exercise—group assignments will be made later—will involve looking at the data set, evaluating what has been done with it as well as any criticisms, and then downloading the data and "doing something" with it. Then reporting on all of this.

Readings:

Thomas Bernaurer and Nils Petter Gleditsch. 2012. New Event Data in Conflict Research. *International Interactions* 38:4, 375-381 (*http://dx.doi.org/10.1080/03050629.2012.696966*)

Individual/Collaborative Assignment: Prepare a one-page description of your paper—basic question, some literature, and most importantly, which data sets and/or text sources are you going to use. Due prior to class on 22 Feb. Co-authored papers are allowed.

[15] COW/MIDS would normally also be included here but, being Penn State, everyone is already familiar with it

Week 7: 22 Feb 2013
Automated Text Classification Methods and Textual Analysis in R

The downside to the availability of news reports on the web is the "drinking from a firehose" problem: even a focused Boolean search can produce a very large number of false positives, and the time required to go through these at least can be very significant. Fortunately, text classification algorithms are a very mature field in natural language processing, and several well-understood methods exist for solving this problem. We will survey these, with an emphasis on the two most widely used—support vector machines and naive Bayes classifiers. In addition, some very robust text analysis packages are now available in *R*: we will look at *tm* ("Text Miner") and *RTextTools* as well as any other packages people have identified as useful.

Readings:

Aggarwal, Charu C. and ChengXiang Zhai. 2012. A Survey of Text Classification Algorithms. In Mining Text Data, ed. Charu C. Aggarwal and ChengXiang Zhai. New York: Springer chapter 6, pp. 77-129. [ANGEL]

Stephen Landis, Vito DOrazio, Philip Schrodt and Glenn Palmer. 2013. "Separating the Wheat From the Chaff: Application of Two-Step Support Vector Machines to MID 4 Text Classification." Working Paper, Penn State. [ANGEL]

Also download the *tm* and *RTextTools* manuals from CRAN (`http://cran.r-project.org/`) and skim these.

Week 8: 1 March 2013
Political Forecasting I: Literature and general methodologies

The classical event data systems—particularly WEIS—were originally motivated by the problem of crisis early warning. This continues to be one of the main foci, particularly in large-scale applied projects such as DARPA's Integrated Crisis Early Warning System (ICEWS) and some components of the U.S. multi-agency Political Instability Task Force (PITF). This week will review both the overall approach and some of the linear time-series methods that have been applied.

Readings:

Ward, Michael D., Brian D. Greenhill, and Kristin M. Bakke. 2010 The Perils of Policy by P-Value: Predicting Civil Conflicts. *Journal of Peace Research*:47,5. (critique of the frequentist regression-based approaches)

Goldstone, Jack A., Robert Bates, David L. Epstein, Ted Robert Gurr, Michael Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. 2010. A Global Model for Forecasting Political Instability. *American Journal of Political Science* 54, 1: 190208. (PITF circa 2009)

King, Gary and Langche Zeng. 2001. Improving Forecasts of State Failure. *World Politics* 53(4): 623-658. http://gking.harvard.edu/files/civil.pdf

Gerald Schneider, Nils Petter Gleditsch, and Sabine Carey. 2011. Forecasting in International Relations: One Quest, Three Approaches. *Conflict Management and Peace Science*, February 2011; vol. 28, 1: pp. 5-14

O'Brien, Sean. 2010. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review* 12,1:.87-104 (ICEWS, Phase I)

Bueno de Mesquita, Bruce. 2011. A New Model for Predicting Policy Choices: Preliminary Tests. *Conflict Management and Peace Science* 28: 65

Additional readings on political forecasting

This topic could, in fact, be a separate semester-length class—see Patrick's Brandt's PSCI 4396 "Predicting Political Conflict" syllabus on ANGEL—so we are just skimming the surface here. But should you want to pursue this further, some relevant sources:

Book length

Armstrong, J. S. (Ed.) (2005). *Principles of Forecasting: A Handbook for Researchers and Practitioners.* Norwell, MA: Kluwer.

Bueno de Mesquita, B. (2002). *Predicting Politics.* Columbus, OH: The Ohio State University Press.

Choucri, N., and T. W. Robinson. (Eds.). (1979). *Forecasting in International Relations: Theory, Methods, Problems, Prospects.* San Francisco: W.H. Freeman.

Davies, J. L. and T. R. Gurr. (Eds.). (1998). *Preventive Measures: Building Risk Assessment and Crisis Early Warning.* Lanham, MD: Rowman and Littlefield. (This has chapters on most of the major forecasting projects developed in the 1990s.)

Granger, C. W. J. and P. Newbold. (1986). *Forecasting Economic Time Series.* 2d ed. Orlando, FL: Academic Press. (In addition, much of the field of time series analysis deals with forecasting in various forms.)

Kahneman, Daniel. 2011. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux. (if you only read one read on this list, this is the one: it summarizes about 30 years of Kahneman's work with Amos Tversky on human decision-making—which got Kahneman a Nobel Prize[16]—and also incorporates a long discussion of Tetlock.)

May, E. R. (1973). *"Lessons" of the Past: The Use and Misuse of History in American Foreign Policy.* New York: Oxford University Press. (This book and Neustadt and May are systematic discussions of the case-study approach.)

Neustadt, R. E. and E. R. May. (1986). *Thinking in Time: The Uses of History for Decision Makers.* New York: Free Press.

[16]Not Tversky, who had died, a bad career move for individuals seeking the Nobel Prize, though not for individuals seeking success in certain genres of music, c.f. Mozart, A., Holly, B, Joplin, J, Hendrix, J, Cobain, K., Jackson, M. and drummers, Tap, S.

Schneider, Gerald, Nils Petter Gleditsch and Sabine C. Carey. 2010. Exploring the Past, Anticipating the Future: A Symposium. *International Studies Review* 12,1. (Special issue based on papers from the theme panels at ISA-2009)

Silver, Nate. 2012. The Signal and the Noise: Why So Many Predictions Fail but Some Don't. New York: Penguin.

Taleb, Nassim Nicholas. The Black Swan. Random House. (One of a series, all focusing on low probability/high consequence events, generally the tails of power-law distributions. Increasingly, the books also focus on the author's innate superiority over 99.999% of the human race, which gets a bit tiresome.)

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

Articles

Brandt, Patrick T. and John R. Freeman. 2006. Advances in Bayesian Time Series Modeling and the Study of Politics: Theory Testing, Forecasting, and Policy Analysis. *Political Analysis* 14(1):136.

Colaresi, Michael and William R. Thompson. 2002. Strategic Rivalries, Protracted Conflict, and Crisis Escalation *Journal of Peace Research* 39: 263 - 287. (retrospective, with the ICB data set, but still relevant)

Fearon, James D. and David D. Laitin, 2003. Ethnicity, Insurgency, and Civil War *American Political Science Review* 97(1):7590.

Goldstone, Jack A., Robert Bates, Ted Robert Gurr, Michael Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. 2005. "A Global Forecasting Model of Political Instability" APSA, Washington. (PITF circa 2005, with references to the various other State Failures and PITF papers and publications.)

Web sites

Political Instability Task Force: `http://globalpolicy.gmu.edu`

Green and Armstrong forecasting site: `http://www.forecastingprinciples.com/` (covers a very wide variety of methods)

Additional readings on pattern-based analyses

Schrodt and Gerner, *AEID*, chapter 4

Asal, Victor, Kihoon Choi, and Krishna Pattipati. 2009. Forecasting the Use of Violence in Ethnic-political Organizations: Middle Eastern Minorities, At Risk Minorities and the Choice of Violence. Iinternational Studies Association, New York (application of SVM classification)

Beck, Nathaniel, Gary King and Langche Zeng. 2000.Improving Quantitative Studies of International Conflict: A Conjecture.*American Political Science Review* 94(1): 21-36 (`http://gking.harvard.edu/files/improv.pdf`) (neural network approaches)

Hudson, Valerie M., Philip A. Schrodt and Ray D. Whitmer. 2008. Discrete Sequence Rule Models as a Social Science Methodology: An Exploratory Analysis of Foreign Policy Rule Enactment Within Palestinian-Israeli Event Data. *Foreign Policy Analysis* 4,2: 105-126. Also see New Kind of Social Science web site: `http://www.nkss.org/`

O'Brien, Sean P. 2002. Anticipating the Good, the Bad, and the Ugly: An Early Warning Approach to Conflict and Instability Analysis *Journal of Conflict Resolution* 46: 791 - 811.

Schrodt, Philip A. and Deborah J. Gerner. 1997. Empirical Indicators of Crisis Phase in the Middle East, 1979-1995 *Journal of Conflict Resolution* 41: 529 - 552. (crisis phase detection rather than forecasting per se)

Weidmann, Nils and Michael Ward. 2009. Predicting Conflict via Machine Learning. International Studies Association, New York.

Elbadawi, Ibrahim and Nicholas Sambanis. 2002. How Much War Will we see?: Explaining the Prevalence of Civil War *Journal of Conflict Resolution* 46: 307 - 334.

Enders, W. and Sandler, T. 2005. After 9/11: Is it all different now? *Journal of Conflict Resolution.* 49(2): 259277.

Additional readings on problems with frequentist analysis

Achen, Christopher. 2002. Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science* 5: 423-450

Brady, Henry E., and David Collier, eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards.* Lanham, MD: Rowman and Littlefield.

Freedman, David A. 2005. *Statistical Models: Theory and Practice.* Cambridge University Press (2005)

Freedman , David A., David Collier, Jasjeet Sekhon and Philip B. Stark, eds. 2009. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences.* Cambridge: Cambridge University Press.

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

Schrodt, Philip. 2006. Beyond the Linear Frequentist Orthodoxy. *Political Analysis* 14,3: 335-339. Also see `event data.psu.edu/7DS`

[lest these critics be considered disgruntled post-modernist outsiders, note that Achen, Brady, Gill and Schrodt have been presidents of the Society for Political Methodology; Freedman was a Fellow of the American Statistical Association and won an award from the National Academy of Sciences for his statistical work. Just an observation.]

SPRING BREAK: 3 - 9 March

Week 9: 15 March 2013
Political Forecasting II: Introduction to time-series approaches

Time series has not been taught for a while in Political Science, so this week will be a general—breadth, not depth—introduction to some of the methods commonly found in event data analysis.

Readings:

Brandt, Patrick T. Michael Colaresi, and John R. Freeman. 2008. The Dynamics of Reciprocity, Accountability, and Credibility. *Journal of Conflict Resolution* 52: 343 - 374.

Other Resources

I am planning to just provide you an overview of the various options, and these topics could easily involve a two-semester, or longer, set of coursework. This cuts across many different disciplines and there is a vast amount of material on the web, increasingly in open access formats. The older approaches, involving adjustments for autoregression, trend, seasonality, smoothing and variations on spectral analysis—are fairly stable; some of the newer approaches, particularly the Bayesian, are still active research areas.

A couple of overview web sites:

`http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm`

`http://www.statsoft.com/textbook/time-series-analysis/`

The most common "encyclopedic" source on methods up to about 1990: Hamilton, James D. 1994. Time Series Analysis. Princeton University Press

Event history and survival models: Janet Box-Steffensmeier and Bradford S. Jones. 2004. Event History Modeling: A Guide for Social Scientists. Cambridge.

Week 10: 22 Mar 2013
Reconciling Multiple Sources and Measures

This deals with three very open-ended issues: how should multiple news sources (e.g. Reuters and AFP) be reconciled, and how should event data be aggregated. On the second issue, we will consider the characteristics of the three main alternatives: scale totals, scale means, and event counts. We may also do some exercises comparing some of the data sets, in particular KEDS, IDEA ("10-million dyadic events"), ICEWS and GDELT. We will also look at Will Lowe's *R* package *events*, which is an aggregation program for TABARI-formatted data.

Readings:

Schrodt and Gerner, *AEID*, chapter 3

Schrodt, Philip A. 2007. Inductive Event Data Scaling using Item Response Theory. Twenty-Fourth Political Methodology Summer Conference, Pennsylvania State University.

Shellman, Steven. 2004. Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Political Analysis*12(1): 97-104.

Week 11: 29 Mar 2013
Problem-Based Exercise Week I: Improving on the ICEWS and PITF Forecasts Using GDELT

Note: We will actually do the three exercises in parallel—with different groups—and then presenting them on the 19th. This list is still tentative pending further discussion and organization of groups.

5 Apr 2013: No class, International Studies Association meetings

Week 12: 12 Apr 2013
Problem-Based Exercise II: Extending the CAMEO categories: Financial and Economic Events

Week 13: 19 Apr 2013
Problem-Based Exercise Week III: Classification and Feature Extraction in Protest Events

Week 14: 26 Apr 2013
Presentation of Research in Conference Format: See Figure 1

Paper Deadline

Paper is due 3 May 2013. If you are under some deadline for the submission of grades, please let me know, though this also may require the paper to be turned in somewhat earlier.

Political science students: Paper should be a PDF produced in LaTeX. Others: use whatever system you'd like, but paper should be a PDF. No need to submit paper copies.

One Final Thought...

> "This is the departure lounge, not the baggage claim."
> Cliff Ketzel (University of Kansas) on teaching political science

Last Update: October 7, 2013