Seven Deadly Sins of Contemporary Quantitative Political Analysis

Philip A. Schrodt
Political Science
Pennsylvania State University
schrodt@psu.edu

Presentation at the
University of Kentucky
3 November 2011

!!! CAUTION !!!

This presentation is being performed by a highly trained and tenured (←!!!) senior academic in a closed and carefully controlled environment.

DO NOT ATTEMPT THIS APPROACH IN A JOB TALK!

We cannot be responsible for the consequences if you do so, though we promise to tip generously when we encounter you serving our vente decaf skinny soy latte, no whip

Seven Deadly Sins

- Kitchen sink models that ignore the effects of collinearity;
- Pre-scientific explanation in the absence of prediction;
- Reanalyzing the same data sets until they scream;
- Using complex methods without understanding the underlying assumptions;
- Interpreting frequentist statistics as if they were Bayesian;
- Linear statistical monoculture at the expense of alternative structures;
- Confusing statistical controls and experimental controls.

The story so far...

- Originally presented at methodology roundtable at APSA
  - Roundtable, so paper not listed on the program
- Top 3 download for most of September
- Top 10 download for 2010
- Top 5% of SSRN downloads for 2010
- Very easy to find on the web

Reaction as a function of age

- < 35: Love it
- >45: Hate it

Reaction as a function of age

- < 35: Love it
- > 45: Hate it
- 35-45:
  Oh, crap. Uh, well, you're probably right, but what is this going to do to my career?

Seven Deadly Sins

1. Greed: Kitchen-sink models and the problem of collinearity
2. Pride: Pre-scientific explanation in the absence of prediction
3. Sloth: “Insanity is doing the same thing over and over again but expecting different results.”
4. Lust: Using complex methods without understanding the underlying assumptions
5. Wrath: If the data are talking to you, you are a Bayesian
6. Gluttony: Enough already with the linear models!
7. Envy: Confusing statistical controls and experimental controls

and...The Four Horsemen of Reductionism: Rational Choice, Game Theory, Systems Dynamics and Agent-Based Models

Seven Deadly Sins

1. Kitchen sink models that ignore the effects of collinearity;
2. Pre-scientific explanation in the absence of prediction;
3. Reanalyzing the same data sets until they scream;
4. Using complex methods without understanding the underlying assumptions;
5. Interpreting frequentist statistics as if they were Bayesian;
6. Linear statistical monoculture at the expense of alternative structures;
7. Confusing statistical controls and experimental controls.

Four problems I will consider today...

- Frequentism
  - except when it isn't a problem
- Pre-scientific “explanation” without the validation of prediction
- Excessive reliance on linear models
  - yes, collinearity as well
- Why we are not doomed
  - What we are already doing right
  - What we could do better
- And because it is on my mind and I have the mic: where is the university heading

Antecedents

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

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

Taagepera, Rein. 2008. *Making Social Sciences More Scientific: The Need for Predictive Models*. Oxford University Press

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

[okay, so maybe we are doomed...]

The Joys of Frequentism

Characterizations of frequentist significance testing (from Gill, 1999)

- “ a strangle-hold" (Rozenboom 1960)
- "an instance of the kind of essential mindlessness in the conduct of research" (Bakan 1960),
- "a terrible mistake, basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (Meehl 1978)
- "deeply flawed or else ill-used by researchers" (Serlin and Lapsley 1993)
- "badly misused for a long time" (Cohen 1994)
- "systematically retarded the growth of cumulative knowledge" (Schmidt 1996)
- "The significance test as it is currently used in the social sciences just does not work" (Hunter 1997)

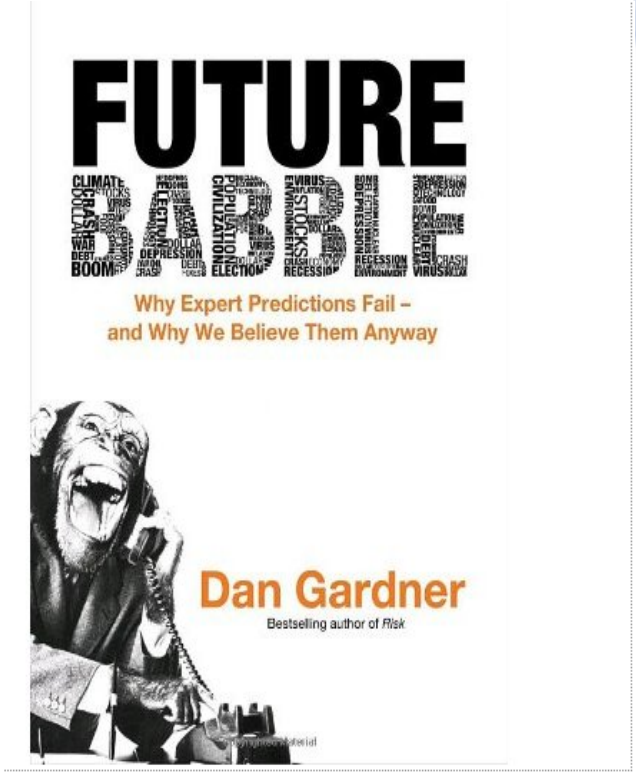
In the popular press

Of Beauty, Sex and Power

Too little attention has been paid to the statistical challenges in estimating small effects

Andrew Gelman and David Weakliem

A reprint from
American Scientist
the magazine of Sigma Xi, The Scientific Research Society



January 5, 2011 The New York Times Reprints

Journal's Paper on ESP Expected to Prompt Outrage

By BENEDICT CAREY

One of psychology's most respected journals has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events.

In the popular press

September 18, 2010

New Drugs Stir Debate on Rules of Clinical Trials

By AMY HARMON

Growing up in California's rural Central Valley, the two cousins spent summers racing dirt bikes and Christmases at their grandmother's on the coast. Endowed with a similar brash charm, they bought each other matching hardhats and sought iron-working jobs together.

Why Almost Everything You Hear About Medicine Is Wrong

Newsweek

Search Newsweek

by **Sharon Begley**

January 24, 2011

Illustration by Jacob Thomas

If you follow the news about health research, you risk whiplash. First garlic lowers bad cholesterol, then—after more study—it doesn't. Hormone replacement reduces the risk of heart disease in postmenopausal women, until a huge study finds that it doesn't (and that it raises the risk of breast cancer to boot). Eating a big breakfast cuts your total daily calories, or not—as a study released last week finds. Yet even if biomedical research can be a fickle guide, we rely on it.

In the popular press

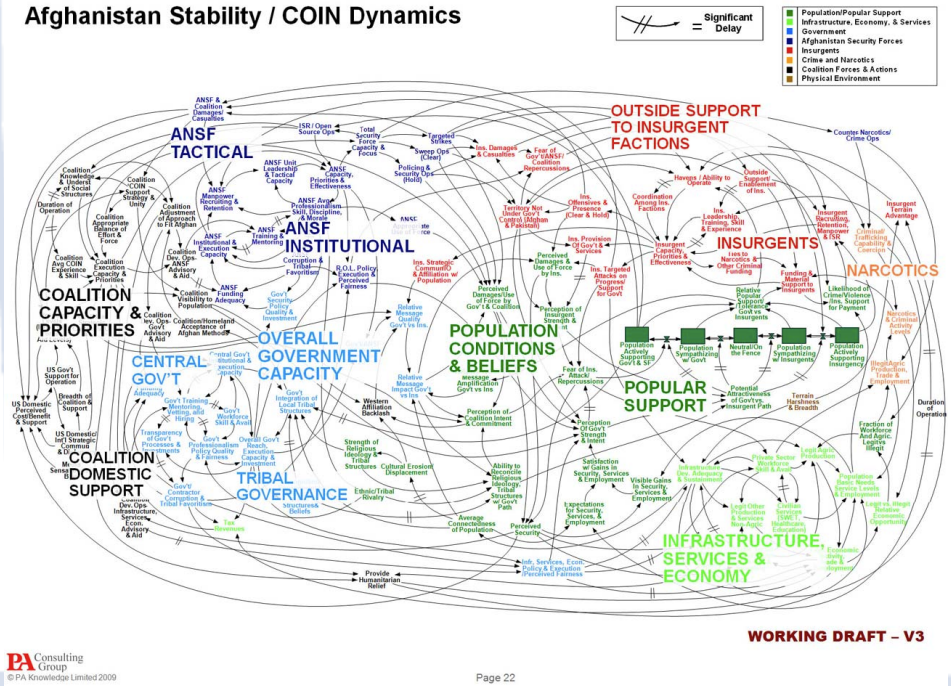
Supreme Court Rules Against Zicam Maker

By ADAM LIPTAK
Published: March 22, 2011

WASHINGTON — The Supreme Court unanimously ruled on Tuesday that investors suing a drug company for securities fraud may rely on its failure to disclose scattered reports of adverse affects from an over-the-counter cold remedy that fell short of statistical significance.

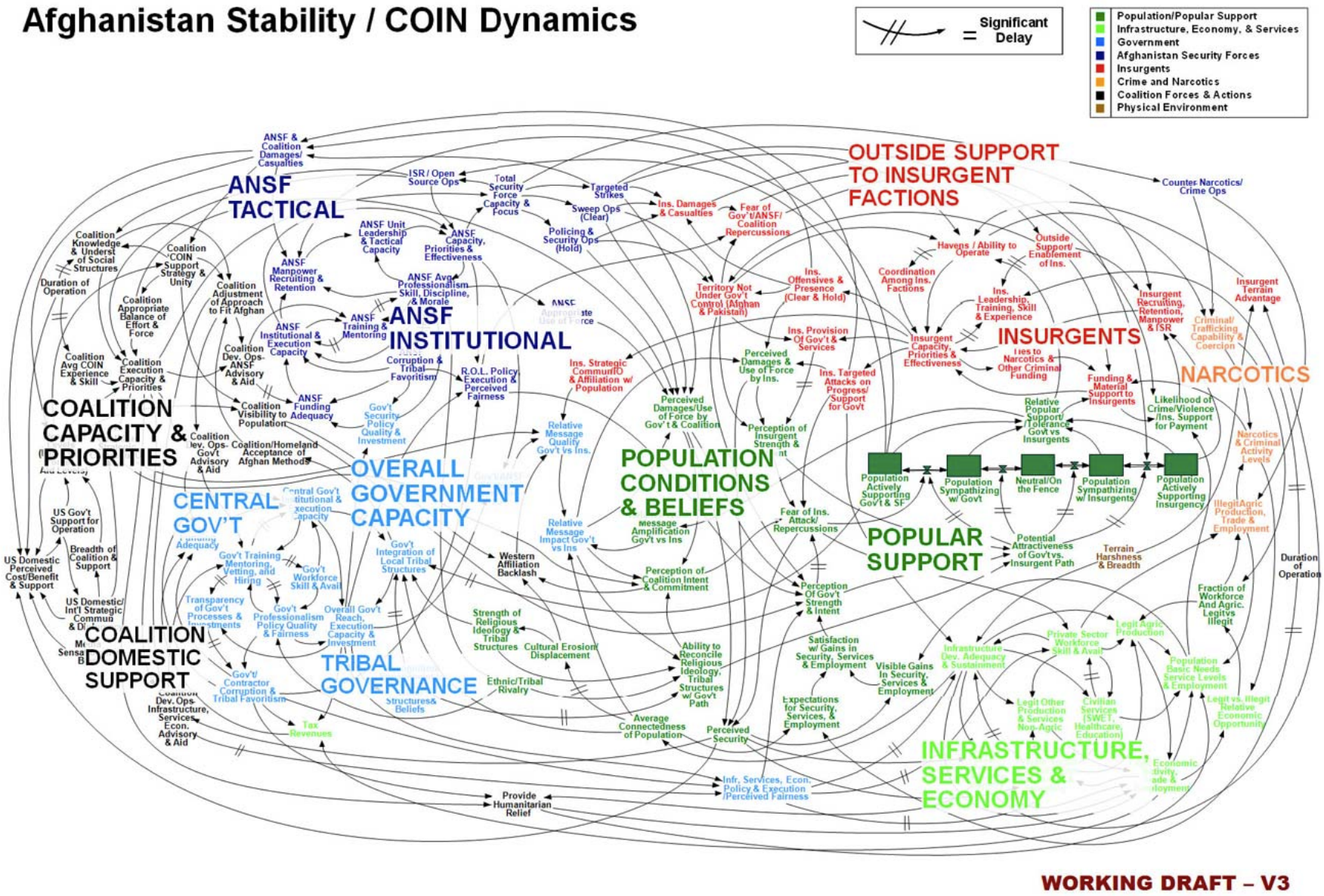
Eric Shelton/Associated Press

Use of Zicam was linked to a loss of smell, a condition known as anosmia.

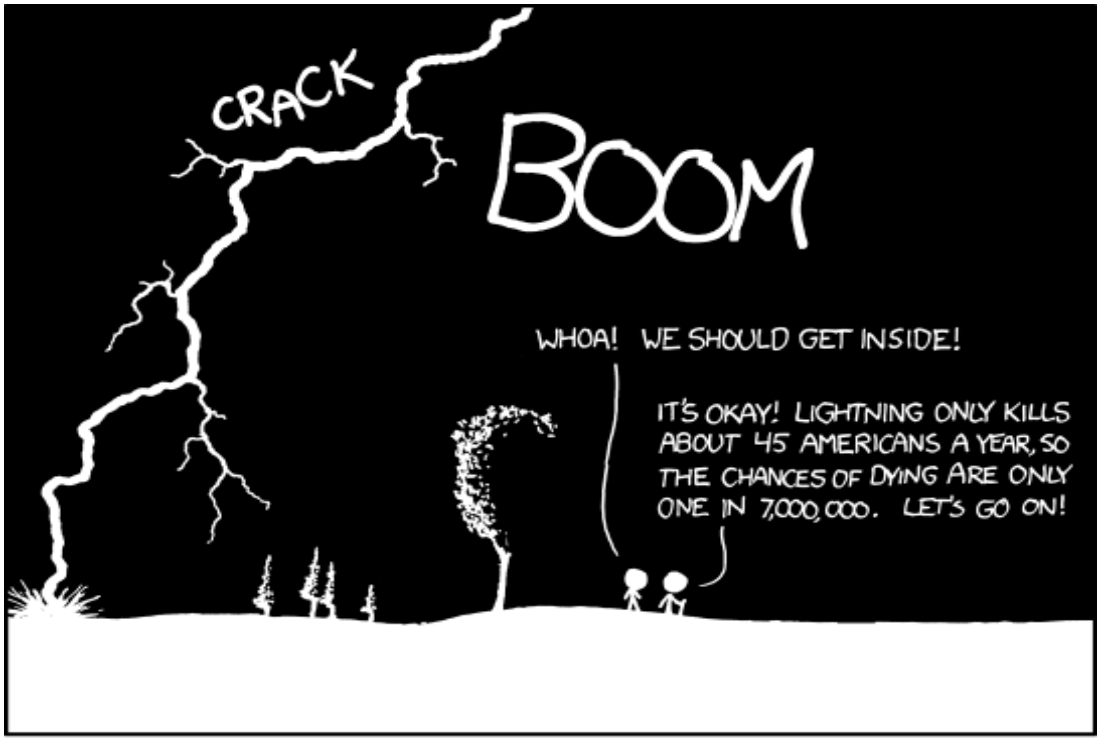


"McChrystal's Hairball"

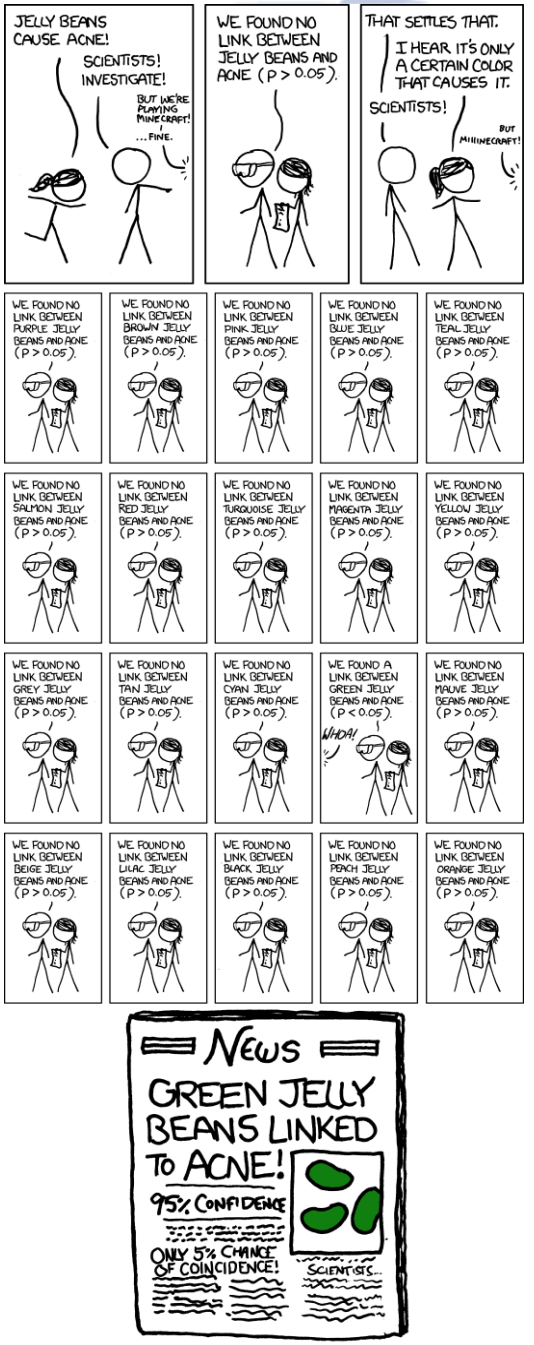
McChrystal's Hairball



Elite commentary



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.



Frequentism is okay provided...

- The null hypothesis is meaningful
  - In the original work in industrial/agricultural stats, it usually is
  - In electoral studies, it often is
- The power of the test is reasonably high
  - 1 – Pr(Type II error) ≈ 0 does not count as “reasonable”...
  - Case in point: augmented Dickey-Fuller test for cointegration
- The test is not repeated excessively
  - Estimate: Oneal-Russett data set has been analyzed at least 3000 times to produce 113 articles
- You are looking at confidence intervals, not knife-edge tests
  - Contrary to the ubiquitous Mystical Cult of the Stars and P-Values
- You remember the correct incantations to caveat a frequentist analysis and can convey them to your audience

Frequentism vs.
The Hypothetico-Deductive Method

Frequentism vs. The Hypothetico-Deductive Method

The hypothetico-deductive—a.k.a. “theory driven”—approach was formalized as such in the mid-19th century

- Gather data (observations about something that is unknown, unexplained, or new)
- Hypothesize an explanation for those observations.
- Deduce a consequence of that explanation (a prediction). Formulate an experiment to see if the predicted consequence is observed.
- Wait for corroboration. If there is corroboration, go to step 3. If not, the hypothesis is falsified. Go to step 2.

Source: Wikipedia

Problems introduced by HDM

- Conventionally, the hypothesis should be a plausible one derived from a theory
  - theory is what keeps *parakeets_per_capita* out of our models. Well, most models.
- Frequentist tests, however, are entirely dependent on the assumptions about the *null hypothesis*, which generally is not plausible
- If the original theory was plausible, the variables in the model probably have a true effect that is not exactly equal to zero, and hence the null hypothesis should always be rejected for a suitably large sample
  - The dependence of the power of tests on the sample size means the conclusions are dependent on an atheoretical feature of the method of observation

Other issues

- Note the usually unappreciated differences between the p-value approach of Fisher and the Type I/II error approach of Neyman and Pearson. (see Gill 1999)
  - These have been ignored since the ABBA—"anything but Bayesian analysis"—compromise of the 1920s
  - In political science, we've been moving away from Neyman-Pearson and towards Fisher, the opposite of what we should be doing (IMHO...)
- It is nearly impossible to explain to a non-expert how and why the conventional approach actually works
  - Even for confidence intervals, which are usually interpreted as the reverse of what they actually say

Bayesian alternative

- You already have some idea about the effects of the variables in your model
- You collect some data
- Your adjust those beliefs based on the data

Why aren't we all Bayesians?

- At an intuitive level, we generally are
  - Babies are Bayesians
  - Even crows and ravens are Bayesians
- Technical implementations of Bayesian estimation remain very difficult
- Bayesian analysis does not hide uncertainty and requires an understanding of probability distributions
  - People are uncomfortable with uncertainly and seem to prefer precise answers, even when those are wrong [Tetlock, Kahneman, sort of]

Prediction

Early technical forecasting models

- Divination model of sheep liver
- Babylonia, ca. 600 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”

Prediction

Against

- Peace Science Society

Prediction

Against

- Peace Science Society

In Favor

- Bruce Bueno de Mesquita, LLC
  - History Channel
  - Economist
- Taagepera [2008]
- Ward, Greenhill and Bakke [2010]
- Bacon

Prediction

Against

- Peace Science Society

In Favor

- Bruce Bueno de Mesquita, LLC
  - History Channel
  - Economist
- Taagepera [2008]
- Ward, Greenhill and Bakke [2010]
- Francis Bacon and 400 years of philosophy of science

Bacon

- Science (Bacon, Descartes):
  experiment >> theory >> authority
- Scholasticism (don't go there, EITM...) :
  authority >> theory >> experiment
- Bacon's Effect on the academic establishment: “crickets”
  - Newton was rejected by the Scholastics because he didn't have a good enough theory compared to Aristotle
- Scholastics: NON OPUS HABENT TETRI PRAEDICTIONES

Bacon

- Science (Bacon, Descartes):
  experiment >> theory >> authority
- Scholasticism (don't go there, EITM...) :
  authority >> theory >> experiment
- Bacon's Effect on the academic establishment: “crickets”
  - Newton was rejected by the Scholastics because he didn't have a good enough theory compared to Aristotle
- Scholastics: NON OPUS HABENT TETRI PRAEDICTIONES
  - “We don't need no stinking predictions”
- Scientific method was not accepted in academic circles until the late 19th century
  - Be afraid, be very afraid
  - But if you drink the Kool-Aid, you'll probably get a job

Prediction

Against

- Peace Science Society

In Favor

- Political Instability Task Force: $2M/yr for 20 years
- Integrated Conflict Early Warning System: $40M
- IARPA ACE and OSI projects: $50M

Applied Prediction Projects in IR

- USAID Famine Early Warning System, early 1990s
- State Failures Project 1994-2001
- Joint Warfare Analysis Center 1997
- FEWER [Davies and Gurr 1998]
- Various UN and EU forecasting projects
- Center for Army Analysis 2002-2005
- Swiss Peace Foundation FAST 2000-2006
- Political Instability Task Force 2002-present
- DARPA ICEWS 2007-present
- IARPA ACE and OSI: 2010-present

Dare you suggest that I adjust my philo-methodo-ontological approach due to the availability of filthy lucre???

Dare you suggest that I adjust my philo-methodo-ontological approach due to the availability of filthy lucre???

US Govt

PSSI

Maybe, maybe not...

Role of prediction for logical positivists

- "Explanation" in the absence of prediction is "prescientific"
- Critical case: astrology vs astronomy
  - More generally, mythological accounts provide "explanation" [Quine]
- Prediction was simply assumed to be a defining characteristic of a good theory until relatively recently
  - Arguably, no philosopher of science prior to the mid-20th century would find the frequentist-based "explanation" emphasized in contemporary political science even remotely justified
- Ward, Greenhill and Bakke (2010): models selected on the basis of significant coefficients are generally miserable at prediction
- Why bother?: Tetlock shows human expert accuracy in political forecasting is 50%-60%

Determinism:
The Pioneer spacecraft anomaly

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

Source: Wikipedia again...

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: 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

What do we predict?

- Discrete outcomes at a fixed time
  - Experiments
  - Elections
- Probabilities of events (or combinations of events) over time
- Hazard rates
- Trends
- Counter-factuals (most difficult, and depends on accurate causal relations)

What is the intrinsic unpredictability in political behavior?

- Statistical political conflict studies: consistently around 20%
- The R^2 *is* an important measure because

$$R^2 = 1 - \frac{Var(e)}{Var(y)}$$

  - Yes, Gary King (1986) is wrong...
- Measures
  - Accuracy/precision/sensitivity
  - Classification/confusion tables
  - ROC/AUC

Methodological monoculture

What's wrong with this picture?

- Correlated variables (aren't they all?) can cause coefficients to take a sign opposite their actual effect and create standard errors the width of Wyoming
- The explanatory power of missing variables (aren't they always?) is distributed to the coefficients of variables that happen to be in the equation
- The (inevitable) presence of anomalous sub-populations and outliers has a disproportionate effect on the coefficient values
- Times series and cross-sectional tests cannot distinguish between [the inevitable combination of] autocorrelated dependent variables and autocorrelated errors
- Standard tests provide no diagnostics for any of these effects since they do not occur under the null hypothesis

But wait...there's more!

- No systematic way of dealing with missing data: cases must be dropped
- Qualitative variables can be handled only with crude numerical hacks
  - Pretty much the same can be said for interaction effects
- Number of variables needs to be substantially less than the number of cases
  - which is not the case in qualitative inference

Alternatives to the linear model

- principal components
- correspondence analysis
- support vector machines
- classification trees: ID3, C4.5, CHAID, random forests
- neural networks
- Fourier analysis
- hidden Markov models
- sequential, functional, topological and hierarchical clustering algorithms
- latent variable models
- genetic algorithms and simulated annealing methods

See (Duda, Hart and Stork 2001, Bishop 2006, Theodoridis and Koutroumbas 2009),

Some improvement...

The individual characteristics differ, but various of these methods allow for

- A wide assortment of nonlinear and hierarchical classification structures
- Systematic reduction of dimensionality for sets of variables that are correlated
- Either robust against missing values or actually can extract information for non-random missing values
  - "missing-at-random" rarely applies in social science data
- Accommodates situations where the number of variables is greater than the number of cases
- Subsets or ignores the effects of outliers

Those were the days...

- “...no single researcher could deal with all the variables in the model and expect to complete more than a very few comparative studies in his [sic] lifetime”
  Herbert McCloskey, *World Politics*, Jan 1956

Those were the days...

- "...no single researcher could deal with all the variables in the model and expect to complete more than a very few comparative studies in his [sic] lifetime"
  Herbert McCloskey, *World Politics*, Jan 1956

- 2011 rendition: "...expect to complete all possible variations of the model in about the time it takes to get a cup of coffee."

Deus ex machina: Bayesian model averaging

- Bayesian
- Handles the zillion model variations that currently clog the journals in a single systematic analysis
- Is this the answer we've been waiting for?? Is this “The One”?!?
  - Seems too convenient...
  - Though it could also be the technology pushing the answer at us: when everyone and their dog can run frequentist models, you can easily automate it

Are we doomed?

Methods Training: What we are doing right

- It exists at all, and is becoming increasingly common
- Basic hypothetico-deductive framework: theory→concepts→variables→measures→tests
- Descriptive statistics and visualization
- Falsification at least in a weak form
- Data reduction and convergent measures
  - but we need more of this
- Problems with the linear model, even if we don't really have solutions (unless BMA is the solution)
- Current emphasis on strong tests of causality and alternatives to linear “controls”

Methods Training: What We Need to do Better

- Re-incorporate a contemporary philosophy of social inquiry
  - "Methodology" is not merely technique
  - Students will be consumers of the whole of social science practice, not merely interpreters of regression coefficients
  - Systematic "qualitative" methodologists—Collier, Gerring, Bennett — are doing a much better job of this than quantitative methodologists
- Balance H-D method with the importance of induction
  - Accommodate contemporary data mining methods, which are not all that different from pre-HTD scientific methods
- Thorough critique of frequentism and the introduction of Bayesian concepts
  - In frequentism, emphasize Neyman-Pearson approach rather than Fisher p-values. ROC curves are a start on this.


Science
11 February 2011 | $10
scientific
data
information
climate
research
science
new
access
researchers
analysis
visualization
many
human
challenges
sharing
need
model
change
cell
use
future
brain
scientists
public
tools
community
health
knowledge
example
AAAS

Methods Training: What We Need to do Better

- Wider variety of methods and emphasis on multiple indicators in a data-rich world
  - Non-statistical inference methods—"machine learning"—need to be accommodated
- De-emphasize Kuhn (and science/culture wars), probably de-emphasize J.S. Mill
  - Mill probably would *want* to be de-emphasized
- Skepticism towards reductionist approaches: formalism does not make something scientific
  - Again, it isn't a cheap shot: astrology and alchemy are formal, but they aren't science

Towards a 21st Century Philosophy of Social Science

- "Scientific realism"?
  - logical positivism hit a dead-end in the 1950s with the ancillary hypothesis problem, but that's not *our* dead-end
- Probabilistic, not deterministic
  - The social sciences are not high energy physics in the 1920s, or 2010s
- Bayesian, not frequentist
- Pragmatic, not ideal
- Causality in the presence of human volition and open complex systems is a [the?] central problem.
  - Again, statistical controls only work in a small number of cases, usually not the ones we are considering outside of randomized experiments

Some thoughts on the future of the university

- Inspiration 1: Steve Jobs's endlessly replayed Stanford commencement address
  - Possibly the only memorable commencement address in human history
- Inspiration 2: Reform-decay-reform cycle of monastic movements
  - Universities are 3rd most durable human institution (cities and religions are 1st and 2nd; relative ordering is contested)

North American University Reforms

- 1820
  - proliferation of liberal art colleges as part of frontier settlement
  - increase in literacy due to public support of secondary education
- 1880
  - Humboldt model replaces the Scholastic model
- 1950
  - Expansion of higher education with public funding from GI Bill, NSF and NIH
  - globalization of faculties first from Europe, then Asia

Things the Boomers actually improved

- More diverse by gender, nationality, race
  - Political Sciences lags in this regard, possibly because of effects of law school, possibly because we are jerks
- Safer environment
  - Low tolerance of sexual harassment, which was terrible earlier: (cf. *Mad Men)*
  - No more departmental drunks (cf. *Mad Men)*
- Best teaching is probably better with active learning, and the worst is probably better now as well
  - Though we may have lost a style of very good teaching involving very strict standards

Things the Boomers actually improved

- Some adaptation to technology, e.g. PPT
  - Standardization of curriculum
- Strong reputation
  - Very strong correlation with economic success, even if some of this is spurious [see Jobs, Gates, Ellison]. But it is unlikely undergraduate education subtracts values (which may occur with law school and MBA programs)
  - Extremely competitive internationally

Problems the Boomers created: Cost

- Overall cost increases since 1980 at twice the rate of inflation
- Rent-seeking bureaucracies
  - “one person's bottleneck is another person's job” [motto of Penn State administration]
- Defunding of public system of the 1950s
- Textbooks [see “journals”, except worse]
- Undergraduate experience as country club

Problems the Boomers created: Student as customer

- Which works if there is a fixed and measurable set of knowledge. Only.
  - You can teach a customer to swim
  - You can't coach a customer to the level of an Olympic athlete
- SRTEs and U.S. News rankings
  - To quote Thomas Friedman on invading Afghanistan: "Welcome, suckers..."

Problems the Boomers created: Journals are broken

- Lowest-common-denominator articles are easy to write and review, and hence become the only ones acceptable and drive out anything original
  - 90%+ of articles are never cited
- “Top Three” journal publication has become a substitute for evaluation in the tenure process
  - NSF Alternative: evaluate only your top N articles (N=10 for NSF)
- Rent-seeking
  - We give away intellectual property rights to [rational choice] articles and pay to get them back
  - Very unclear how this model can survive in an internet age, but change is fiercely resisted by the professional organizations (yes, APSA)

Problems the Boomers created: Decline of the humanities

- The post-modernists succeeded where the Puritans, the Inquisition, Hitler, Mao and Stalin all failed.
  - Nice job, dudes!
- Occupational focus at the expense of the liberal arts didn't help either.

Problems the Boomers created: Adjuncts [maybe]

- Severely weakened the tenure system and replaced professional self-governance with a tiered system
- However:
  - When done humanely, it creates a class of teaching specialists, cuts costs and allows for smaller classes
  - Idea: tenure the adjuncts—who actually need the protection—and hire research faculty on an “you eat what you kill” basis

Why the internet is so important

- It is pure information
  - at least some of that information is knowledge; the remainder entertainment
- Marginal cost of duplication is zero
- Marginal cost of cataloging is zero
  - it is much better cataloging, and it is dynamic
- It is global
- It is current
- We don’t have to teach it