Bush 631-607: Quantitative Methods

Lecture 14 (04.26.2022): Review & Summary

Rotem Dvir

The Bush school of Government and Public Policy

Texas A&M University

Spring 2022

What is today's plan?

- Review course topics.
- Research designs.
- Predictions, probability and estimations.
- Social science and real-world politics.
- Data science in the real-world.

Course review: Causality

- Establish Cause $\rightarrow$ effect.
- Using (hypothetical) *Counterfactual*.

Establish causality

- Treatment and control groups.
- Same outcome measures.
- Gold standard $\rightarrow$ randomization.
- Calculate ATE over group of respondents.
- Difference in means estimator.

$$SATE = \frac{1}{n} * \sum_{i=1}^{n} * Y_i(1) - Y_i(0)$$

Establish causality

Observational research design

- Using observed data to assess causality.

- Good for generalizing results.

- Not as good for randomization.

- Problem of pre-treatment variables (confounders).

Assessing research designs

Strengths and Weaknesses

- Internal validity:
    - How does the design helps answering the research Q?
    - Experiments → strong (randomization).
    - Observational → weak (confounders).

- External validity:
    - Can we generalize the results from sample?
    - Experiments → weak (hypothetical).
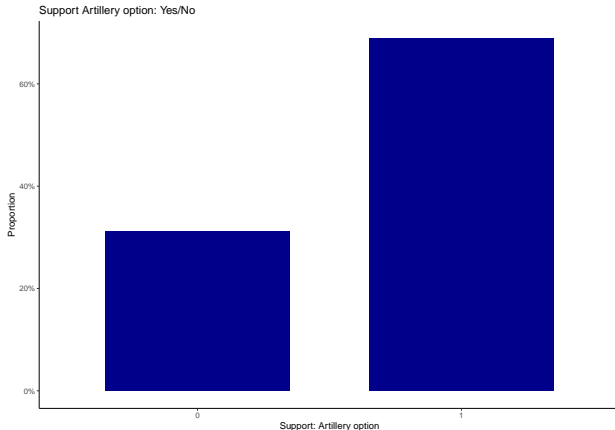   - Observational → strong (real-world, cross-national).

Course review: Measurement

- ▶ Apply quant methods for social science.

- ▶ Measures → the context of concepts.

- ▶ Challenge: latent factors
    - ▶ What is ideology? How do we measure it?
    - ▶ Terrorism?
    - ▶ Democracy - polity vs. freedom house scores.

- ▶ Challenge: missing values
    - ▶ No responses (refuse to answer)
    - ▶ Don't know / not sure
    - ▶ NA in our data

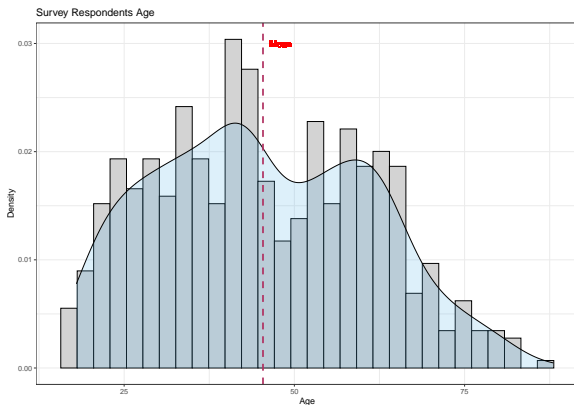# Visuals

▶ Barplot: counts/proportions for categories

```
ggplot(wardata, aes(x=factor(prefer_artillery_dummy))) +
  geom_bar(aes(y = (..count..)/sum(..count..)), width = 0.7, fill = "darkblue")
  xlab("Support: Artillery option") + ylab("Proportion") +
  scale_y_continuous(labels=scales::percent) + ggtitle("Support Artillery optio
  theme_classic()
```

Visuals

- Histogram: distribution of numerical variable

```
ggplot(wardata, aes(x=age)) +
  geom_histogram(aes(y=..density..), colour="black", fill="lightgrey")+
  geom_density(alpha=.2, fill="#56B4E9") +
  xlab("Age") + ylab("Density") + theme_bw() + ggtitle("Survey Respondents Age") +
  geom_vline(aes(xintercept=mean(age)),
             color="maroon", linetype="dashed", size=1) +
  geom_text(x = 48, y = 0.03, label = "Mean", col = "red")
```
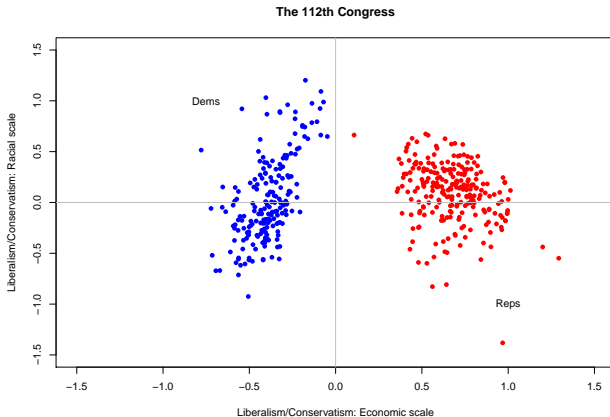
Visuals

▶ Boxplot: Compare single variable distribution

```
ggplot(wardata, aes(x=factor(agegroup), y = artillery_approve,
                    color = factor(agegroup))) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  xlab("Age groups") + ylab("Support for Artillery option") +
  theme_classic() + theme(legend.position = "none")
```
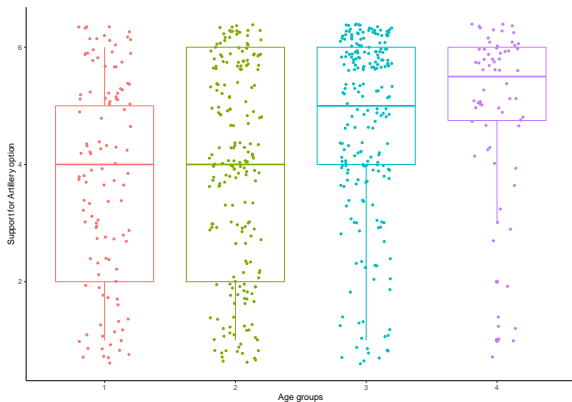
Visuals

- Scatterplot: Visualize bivariate relationship

Visuals

► Plot time trends (UN Voting data 1946-2012)

Predictions

- ▶ Predict with sample mean: using loops.

- ▶ Prediction error = actual outcome - predicted outcome.

- ▶ RMSE: average magnitude of prediction error.

- ▶ Correlations:
  - ▶ Summary of bivariate relationship.
  - ▶ How two factors 'move together' on average.
  - ▶ Always relative to mean value.

```
# Voting with US
cor(unvoting$idealpoint, unvoting$PctAgreeUS, use = "pairwise")
```
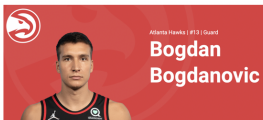
```
## [1] 0.7498446
```

Predictions

- The linear model: $Y = \alpha + \beta * X_i + \epsilon$

- Model elements:
  - Intercept (α): the average value of Y when X is zero.
  - Slope (β): the average change in Y when X increases by 1 unit.
  - Error/disturbance term (ϵ): the deviation of an observation from a perfect linear relationship.

- Least squared:
  - How to estimate the regression line.
  - 'Select' $\hat{\alpha}, \hat{\beta}$ to minimize SSR.
  - R syntax: lm(y $\sim$ x, data = mydata)

Regression to the mean

- High (low) observations are followed by low (high) observations.

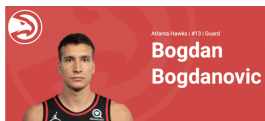- Observations 'regress' towards the average value of the data.
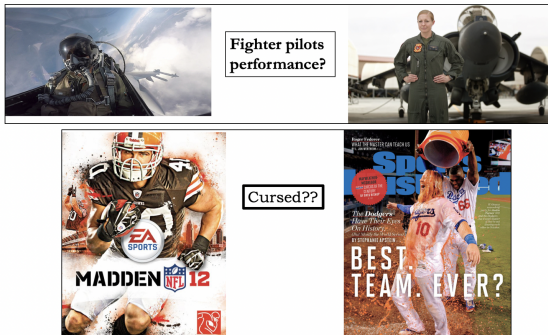
Regression to the mean

Game 1 (04.18.2022)	
Points	6
Rebounds	3
Assists	2
FG %	0-8 (0%)
3-PT %	0-4 (0%)

Game 2 (04.20.2022)	
Points	29
Rebounds	4
Assists	3
FG %	12-18 (66%)
3-PT %	5-10 (50%)

Average over 2 games	
Points	17.5
Rebounds	3.5
Assists	2.5
FG %	46%
3-PT %	35.7%

Regression to the mean

Season Average (63 games)

Points	15.1
Rebounds	4
Assists	3
FG %	43.1%
3-PT %	36.8%

Game 1 (04.18.2022)

Points	6
Rebounds	3
Assists	2
FG %	0-8 (0%)
3-PT %	0-4 (0%)

Game 2 (04.20.2022)

Points	29
Rebounds	4
Assists	3
FG %	12-18 (66%)
3-PT %	5-10 (50%)

Average over 2 games

Points	17.5
Rebounds	3.5
Assists	2.5
FG %	46%
3-PT %	35.7%

Probability

- Quantify uncertainty: step 1.
- Sample space, events (cards, coin toss)

- Probability: $P(A) = \frac{Elements(A)}{Elements(\Omega)}$
- Conditional probability $= P(A|B) = \frac{P(A\&B)}{P(B)}$
- Uncertainty of sample means or sums.
- Expectations of r.v. $\rightarrow$ population value.
- Variance of r.v. $\rightarrow$ 'spread' of distribution.
- CLT and large samples.

Uncertainty and estimation

- Estimate based on r.v.s.

- Quantity of interest: *point estimate* (mean / diff-in-means)

- How to learn of estimator distribution? simulations. . .

- Calculate SD, or SE in single sample.

- Construct 95% CIs - how to interpret?

Hypothesis tests

- Estimators: sample means / diff-in-means

- *Proof by contradiction.*

- Steps for testing:

  1. Define null and alternative hyps (H_0; H_1).
  2. Select *test statistic* and level of test (α).
  3. Derive reference distribution.
  4. Calculate p-values.
  5. Make a decision: reject/retain.

- **Decision rule**:

  - **Reject null** if p-value is *below* $\alpha = 0.05$
  - Otherwise **retain the null** or **fail to reject**.

Hypothesis test

- Run *Two-sample t-test* with t.test()

```
t.test(exp.dat$cont_cor1[exp.dat$trt1 == 0],
       exp.dat$cont_cor1[exp.dat$trt1 == 1])
```

```
##
## 	Welch Two Sample t-test
##
## data:  exp.dat$cont_cor1[exp.dat$trt1 == 0] and exp.dat$cont_cor1[exp.dat$tr
## t = -13.697, df = 993.53, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.59653 -17.68267
## sample estimates:
## mean of x mean of y
##  1489.333  1509.973
```

Putting everything together

- Hypotheses:
  - $H_0 : \beta_1 = 0$
  - $H_a : \beta_1 \neq 0$

- Our estimators: $\hat{\beta}_0, \hat{\beta}_1$

- SE and CIs:
  - $\hat{\beta}_0 \pm 1.96 * \hat{SE}(\hat{\beta}_0)$
  - $\hat{\beta}_1 \pm 1.96 * \hat{SE}(\hat{\beta}_1)$

- Hypotheses test:
  - Test statistic: $\frac{\hat{\beta}_1 - \hat{\beta}_1^*}{\hat{SE}(\hat{\beta}_1)} \sim \mathsf{N}(0,1)$
 - $\hat{\beta}_1$ is **statistically significant** if $p < 0.05$.

Rebels and Nukes (2015)

- Multivariate regression: account for confounders

```
summary(lm(pursuit ~ rebel + milservice + polity2, data = nukes))
```

```
##
## Call:
## lm(formula = pursuit ~ rebel + milservice + polity2, data = nukes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06587 -0.04408 -0.02544 -0.01020  0.99682
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0073899  0.0027782   2.660  0.00783 **
## rebel       0.0320096  0.0044238   7.236 5.08e-13 ***
## milservice  0.0217914  0.0045106   4.831 1.38e-06 ***
## polity2     0.0004679  0.0002801   1.670  0.09489 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1672 on 7684 degrees of freedom
##   (1164 observations deleted due to missingness)
## Multiple R-squared:  0.01596,    Adjusted R-squared:  0.01558
## F-statistic: 41.54 on 3 and 7684 DF,  p-value: < 2.2e-16
```
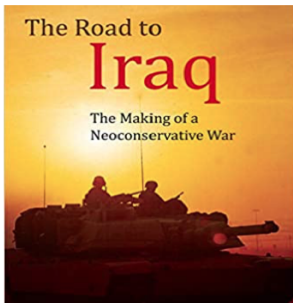
OLS coefficient interpretation

- ▶ Rebel experience and nuclear technology (2015)
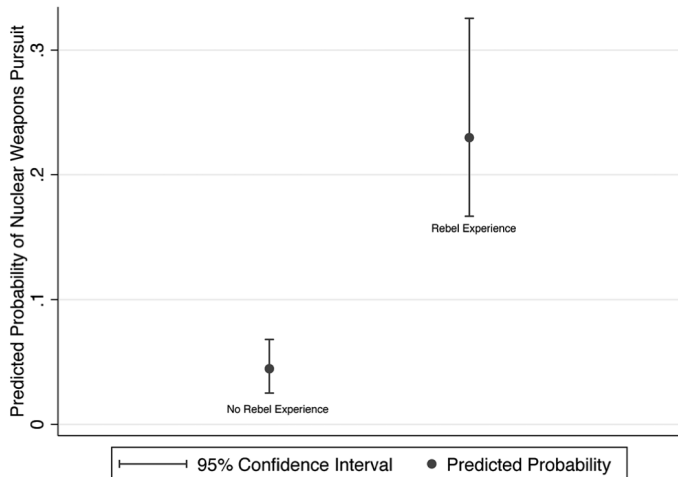
How to use our research?

Applying theories and IR research

Applying IR research in global affairs

- ▶ The motivation:
  - ▶ We study IR or social science dynamics.
  - ▶ Do policymakers use? or even care about this knowledge?

The New York Times | http://nyti.ms/1f4Huyy

SundayReview | OP-ED COLUMNIST

Professors, We Need You!

Nicholas Kristof   FEB. 15, 2014

SOME of the smartest thinkers on problems at home and around the world are university professors, but most of them just don't matter in today's great debates.

The most stinging dismissal of a point is to say: "That's academic." In other words, to be a scholar is, often, to be irrelevant.

David Rothkopf @
@djrothkopf

Kristof gets why we at FP are dialing back academic contributions--too many are opaque, abstract, incremental, dull. nyti.ms/1fpsd84

6:30 AM · Feb 16, 2014 · Twitter for Websites

50 Retweets   **31** Likes

Applying IR research

- Main challenge → **Time** and reading research article.

- Scholars adapt:
  - Joint forums: *bridging the gap*.
  - Policy-focused writing (Lawfare blog, War on the rocks, Monkey cage).

- Is it working? Ask policymakers. . .

- Previous work (2014):
 - Not really.
  - Academic work not aimed to 'close the gap'.

Applying IR research

- Recent evidence (2021) → replicate 2014 survey.

- Innovations for better insights:
  - Diverse sample - more areas FP.
  - Both high and low-ranking officials.
  - Embedded experiment for direct effects.
  - Broader conception for engagement (social media).

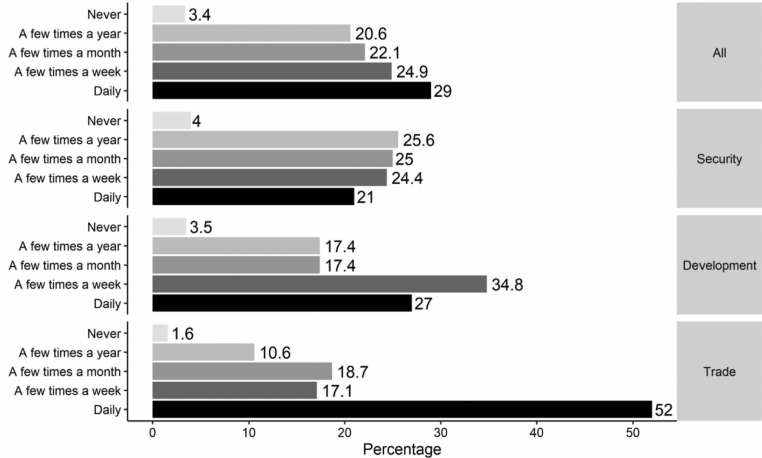- Sample: 616 officials (Clinton, Bush, Obama administration).

Useful political research?
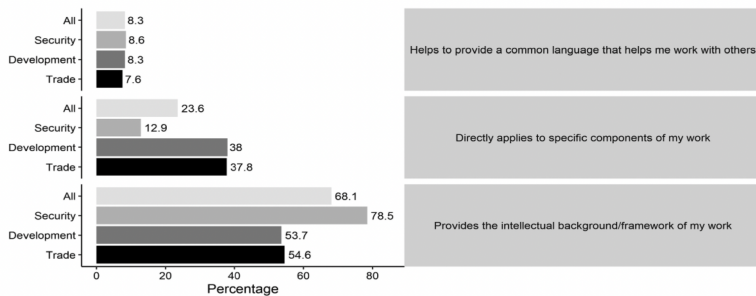
▶ Research consensus and updating policy views.
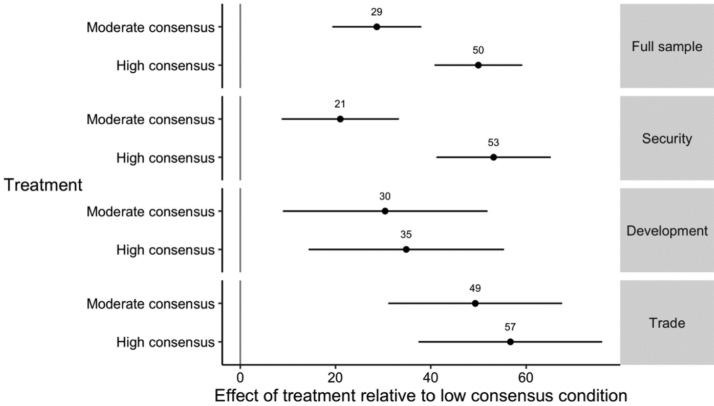
Useful political research?
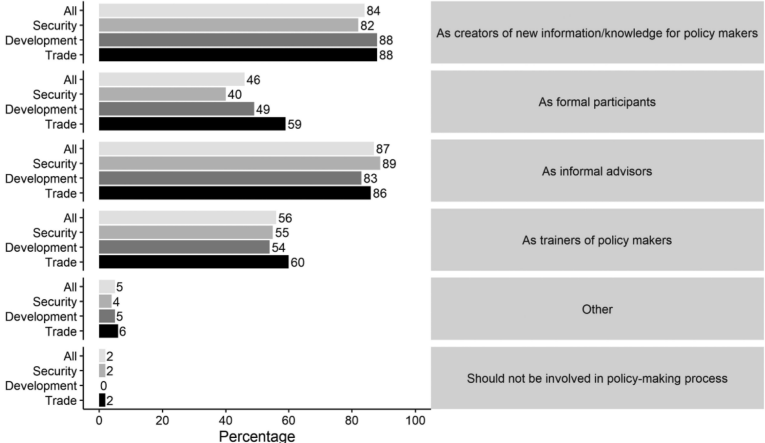
- Frequency of using research into government work.

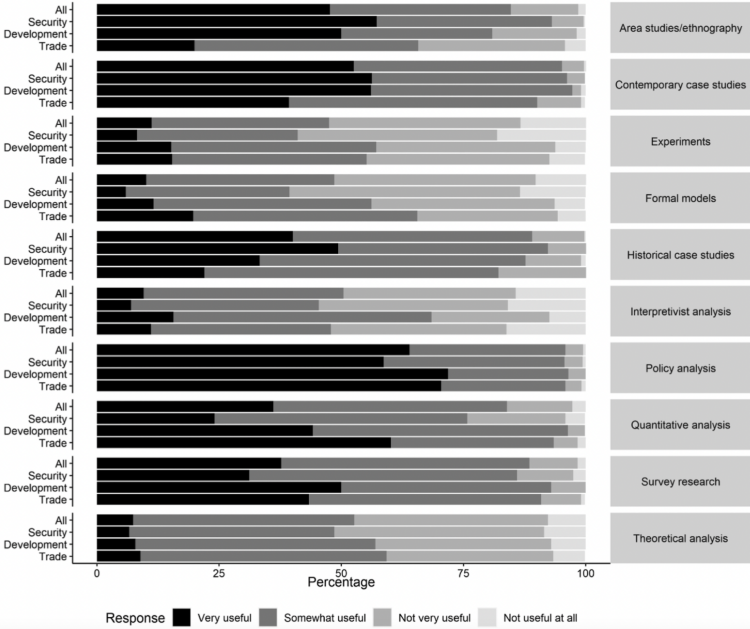# Useful political research?
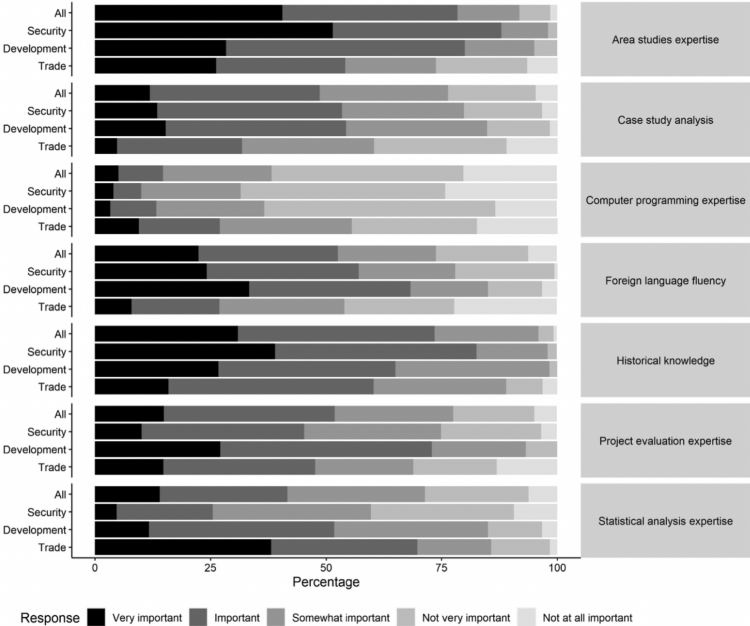
▶ How do you use academic research?

Useful researchers

► How can researchers contribute the most?

Useful skills. . .
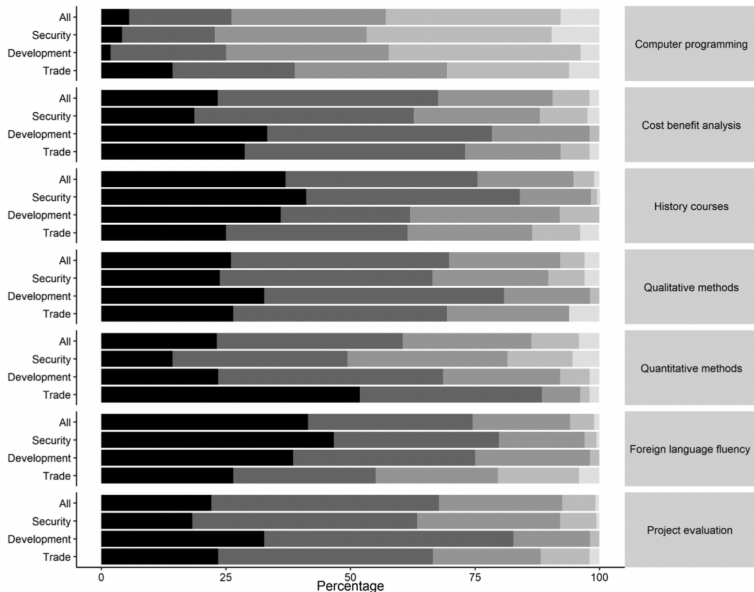
Useful skills. . .

Working for the government...

Using political research

- How can research be useful (Byman & Kroenig 2016)
  - Practical and useful recommendations.
  - Focus - clarify complex situations.
  - Time relevant research.

- Scenarios for applying academic insights:
  - Challenge existing government knowledge - shocks (9.11, Soviet collapse).
  - Policy failures (Iraq insurgency outbreak).
  - Missing baseline knowledge (Somalia intervention).

Becoming useful political advisor

- Concrete steps:
  - Networking and personal connections.
  - 'Inject' research into bureaucracy.
  - Concise and clear reports in nonacademic outlets.

- Tamper expectations:
  - What is being relevant?
  - Not likely to drastically shape policy.
  - Influence the **deliberation** process.

- What's in it for policymakers?
  - Offer *contrarian arguments* to accepted view.

Data science in the real world

Prediction by the numbers (2019)

- ▶ Data analysis → set of tools to understand the world.
- ▶ The core role of probability.
- ▶ Apply complex concepts like repeated sampling.
- ▶ Bayesian logic and saving lives.

A wrap on Spring 2022

Part of the journey is the end.