CSSS 569: Visualizing Data

Visual Displays in Quantitative Social Science

Christopher Adolph

Department of Political Science
and
Center for Statistics and the Social Sciences
University of Washington, Seattle

Why teach a social science statistics course about the Visual Display of Quantitative Information (VDQI)?

Visual displays are woven throughout the sciences & esp. statistics

Like literacy and numeracy, visual communication takes practice to do well

Social scientists in particular lack this training, and under-utilize visuals

Yet visuals are essential to navigate the complex, multidimensional data and models facing social scientists

Good visuals *work*. They convey the result and make it memorable.

Scholars don't just write papers—they present them.

Visuals make or break talks—including *job* talks

Three Examples of Visual Quantitative Analysis

- Success! Stopping infectious disease
- Failure. The *Challenger* disaster
- Confusion? Sorting through complex models of policymaking.

(Consider as well three uses of visuals:

how we use visuals to explore *data*

how we use visuals to understand *model implications*

how we use visuals to test *model fit*)

John Snow stops the Cholera epidemic

Cholera outbreaks were common in 19th century London; 10,000s of deaths

Popular theory: Cholera caused by "miasma" in the air coming from swamps

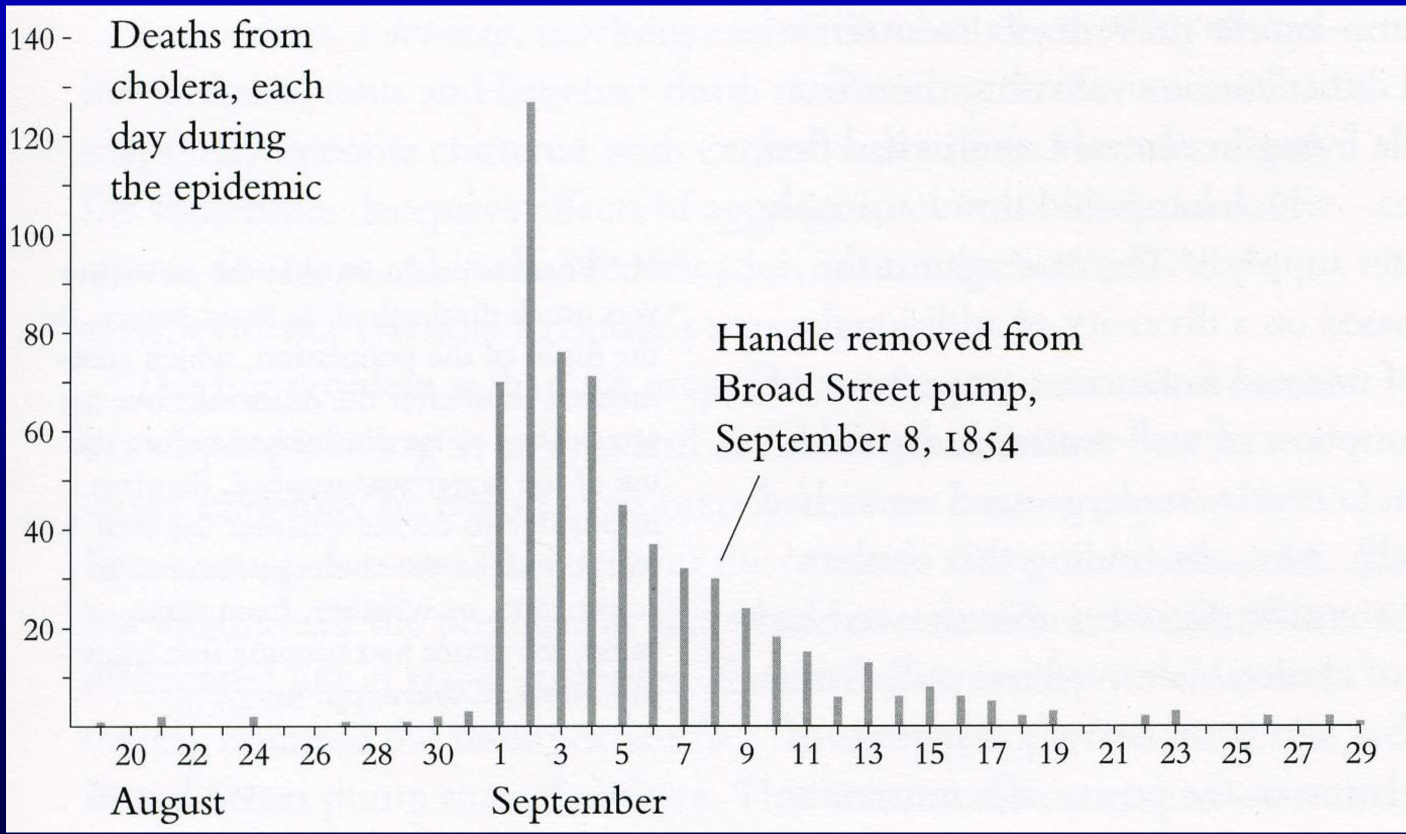
Or perhaps a "poison" that slowly loses strength as it passes from victim to victim

London doctor John Snow believed cholera actually caused by contaminated water

Outbreak in 1854: 500 deaths in 10 days in Soho

John Snow stops the Cholera epidemic

The best VDQI we could expect a modern newspaper?



(plot from Tufte, *Visual Explanations*)

Overwhelming tendency to view time series data this way
Doesn't help us make inferences about the data

The data aren't being compared to any covariates:
time series plots are usually boring models

John Snow stops the Cholera epidemic

In 1954, London water was provided by competing private firms

Residents would walk to the nearest street pump for water

Snow recorded the location—house by house—of each death through the outbreak.

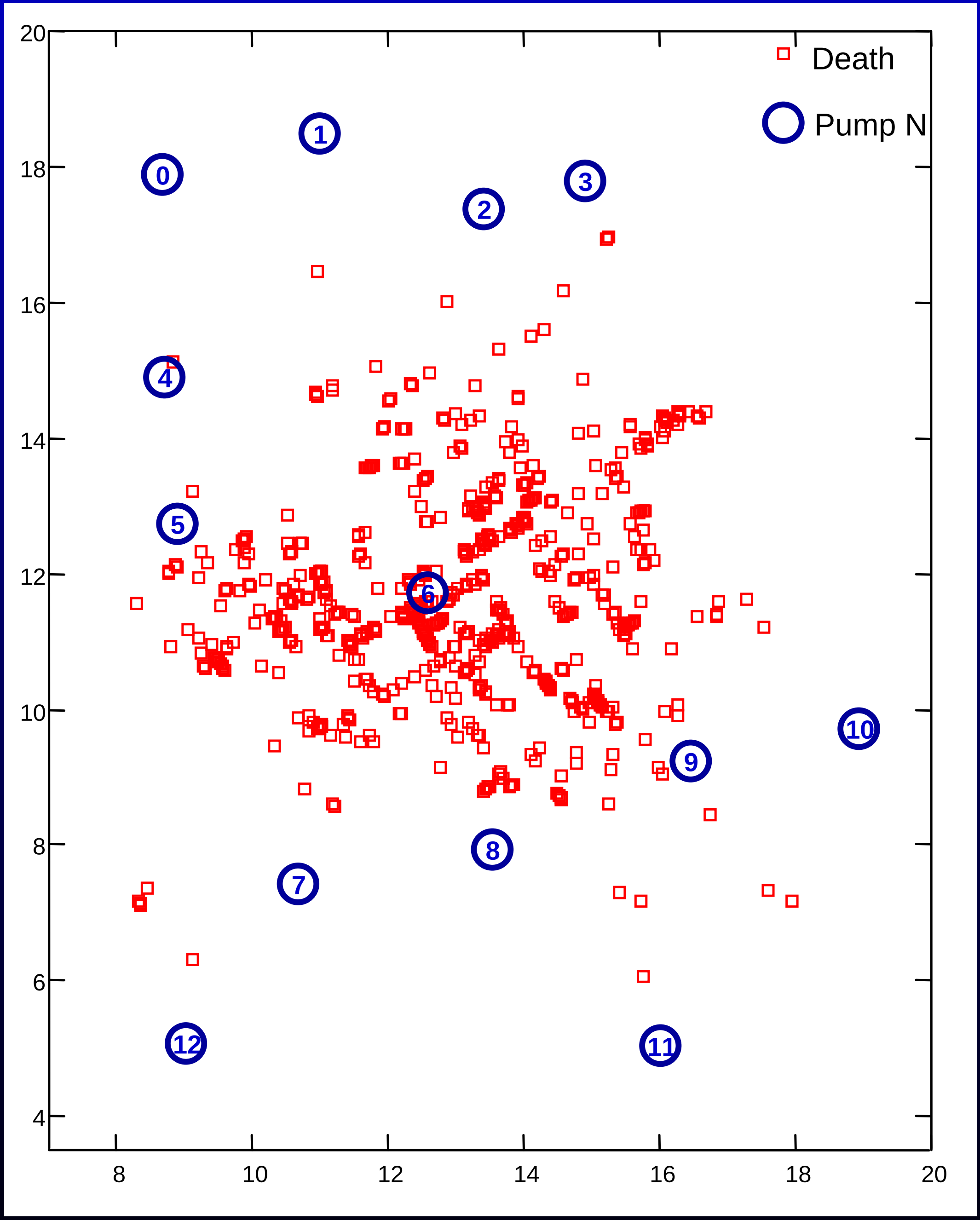
Placed these spatial data on a map, along with the *water pumps*

Was one “pump”, from a particular company, contaminated with cholera?

John Snow stops the Cholera epidemic

Reproduced from *Visual and Statistical Thinking*, ©E.R. Tufte 1997, based on Snow's drawing .

John Snow stops the Cholera epidemic



John Snow stops the Cholera epidemic

(The second version, better suited to slide display, is by Alyssa Goodman)

Deaths are concentrated around Broad Street pump, not others

Was it the source of the epidemic?

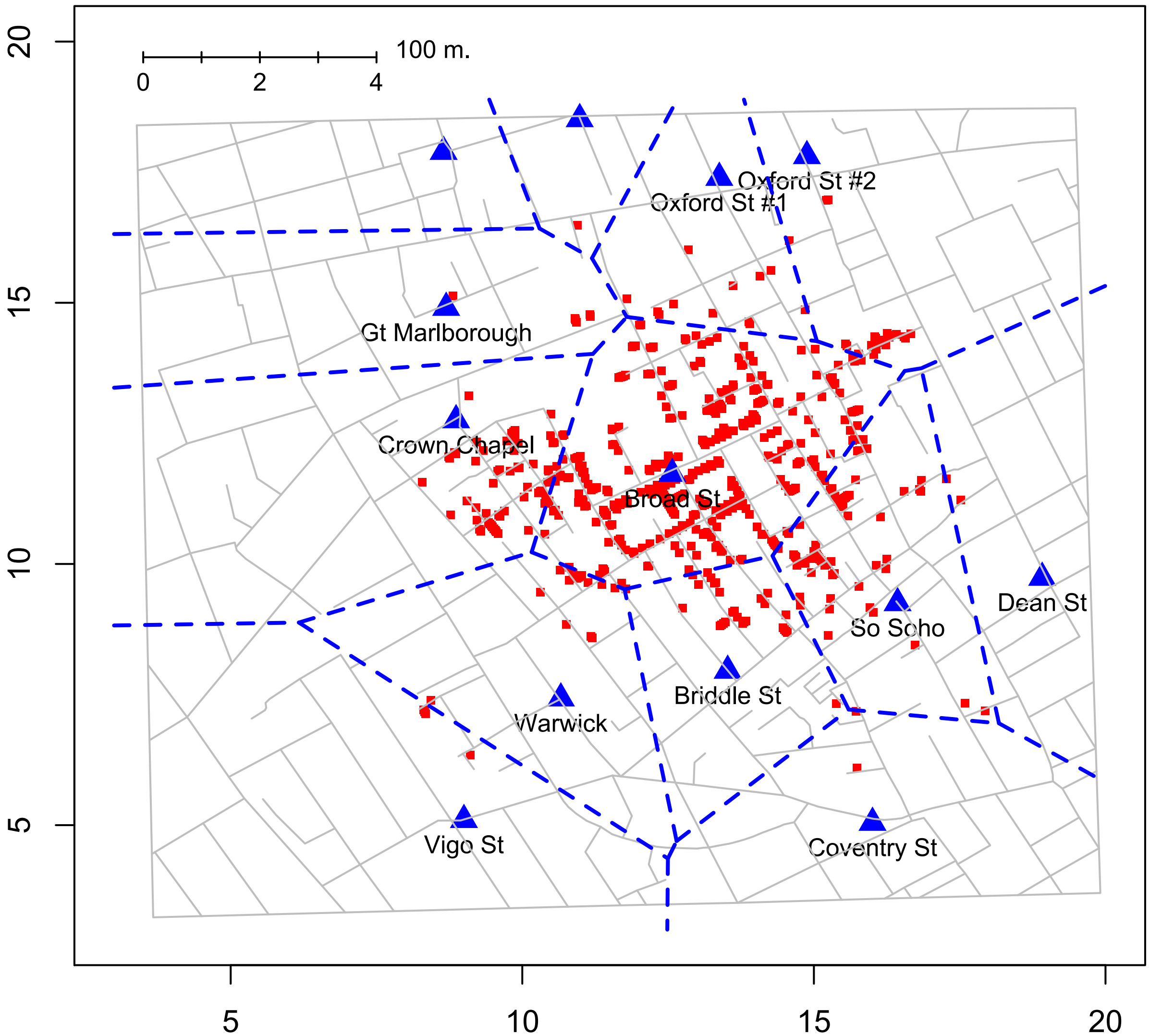
Or could this evidence be consistent with a different story?

How might the map be misleading?

What does the map hide?

What other evidence would you like to have?

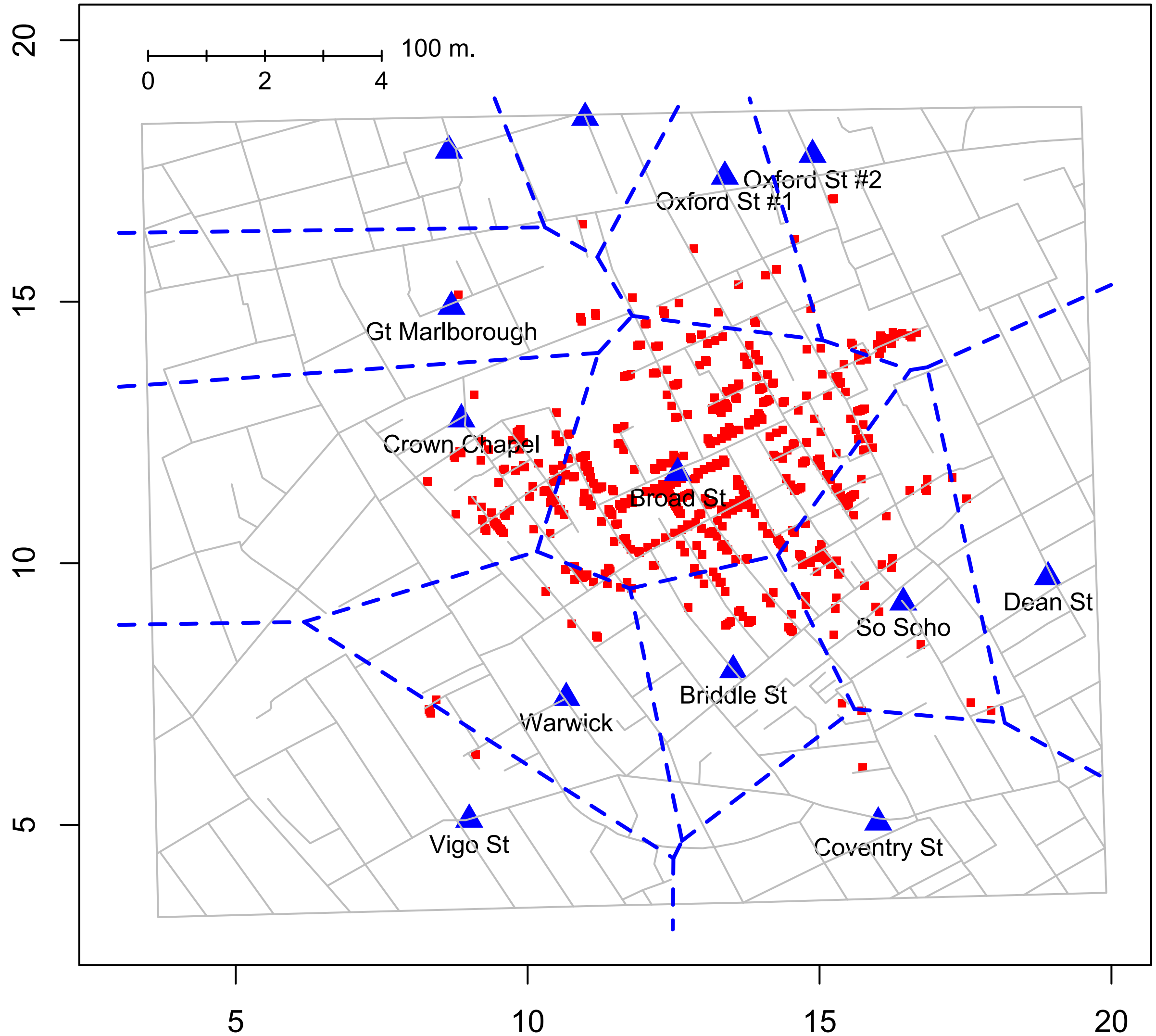
A map with a model



The blue dashed lines mark Voroni cells, or regions closest to a single pump.

Model: people go to the nearest pump $\Rightarrow$ everyone in a cell uses the same pump

Exceptions that prove the rule . . .



What explains the outliers in the map?

Exceptions that prove the rule . . .

Reproduced from *Visual and Statistical Thinking*, ©E.R. Tufte 1997, based on Snow's drawing .

Exceptions that prove the rule . . .

Three cases:

- A prison (work house) with its own well.
- A brewery with its own water source. Saved by the beer.
- Some distant deaths attributable to preference for Broad St. water.

John Snow stops the Cholera epidemic?

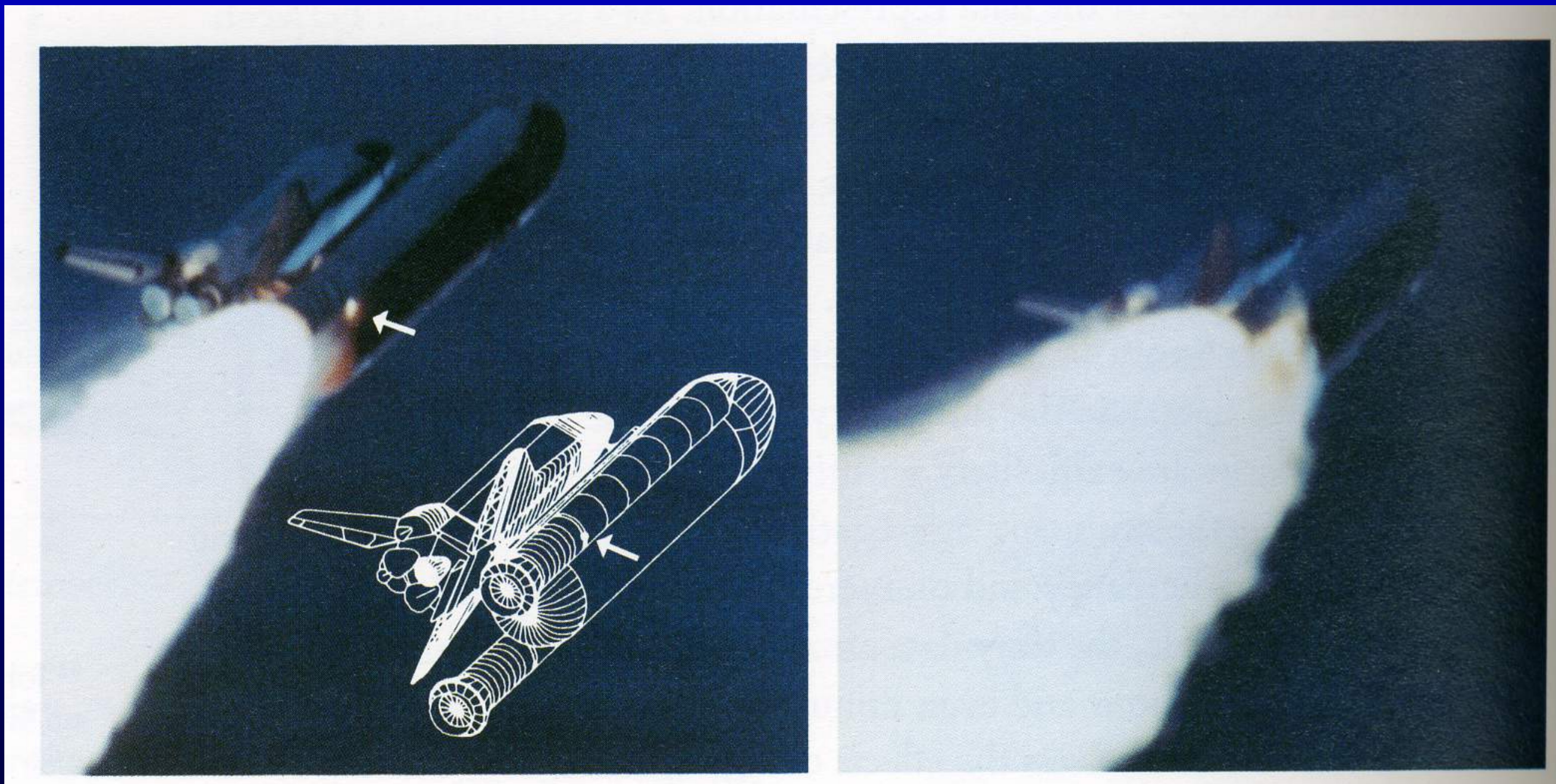
Snow used his data and map to convince officials
to remove the handle from the Broad Street pump.

Credited with stopping the outbreak and
providing the first experimental evidence for germs

Some questions to consider later:

- Did the Broad Street Pump really cause the cholera outbreak?
- Did removing the handle stop it? (this is a *separate* question)
- How could Snow's map be improved as a VDQI?

The *Challenger* launch decision



In 1986, the Challenger space shuttle exploded moments after liftoff

Decision to launch one of the most scrutinized in history

Failure of O-rings in the solid-fuel rocket boosters blamed for explosion

Could this failure have been forseen? Using statistics?

The *Challenger* launch decision

Here is the data on O-ring failures at different launch temperatures

Flights with O-ring damage	
Flt Number	Temp (F)
2	70
41b	57
41c	63
41d	70
51c	53
61a	79
61c	58

Morton-Thiokol engineers made this table & worried about launching below 53° (Why?)

O-ring would erode or have "blow-by" (2 ways to fail) in cold temp

Failed to convince administrators there was a danger

(Counter-argument: "damages at low and high temps")

Are there problems with this presentation? with the use of data?

The *Challenger* launch decision

Engineerrs did not consider successes, only failures;
selection on the dependent variable

All flights, chronological order			
Damage?	Temp (F)	Damage?	Temp (F)
No	66	No	78
Yes	70	No	67
No	69	**Yes**	53
No	68	No	67
No	67	No	75
No	72	No	70
No	73	No	81
No	70	No	76
Yes	57	**Yes**	79
Yes	63	No	76
Yes	70	**Yes**	58

Other problems?

The *Challenger* launch decision

Engineerrs did not consider successes, only failures;
selection on the dependent variable

All flights, chronological order			
Damage?	Temp (F)	Damage?	Temp (F)
No	66	No	78
Yes	70	No	67
No	69	**Yes**	53
No	68	No	67
No	67	No	75
No	72	No	70
No	73	No	81
No	70	No	76
Yes	57	**Yes**	79
Yes	63	No	76
Yes	70	**Yes**	58

Other problems? Why sort by launch number?

The *Challenger* launch decision

O-ring damage pre-Challenger, by temperature at launch			
Damage?	Temp (F)	Damage?	Temp (F)
Yes	53	**Yes**	70
Yes	57	0	70
Yes	58	0	70
Yes	63	0	72
No	66	No	73
No	67	No	75
No	67	No	76
No	67	No	76
No	68	No	78
No	69	**Yes**	79
Yes	70	0	81

The evidence begins to speak for itself.

What if Morton-Thiokol engineers had made this table before the launch?

The *Challenger* launch decision

Why didn't NASA make the right decision?

Many answers in the literature:
bureaucratic politics; group think; bounded rationality, etc.

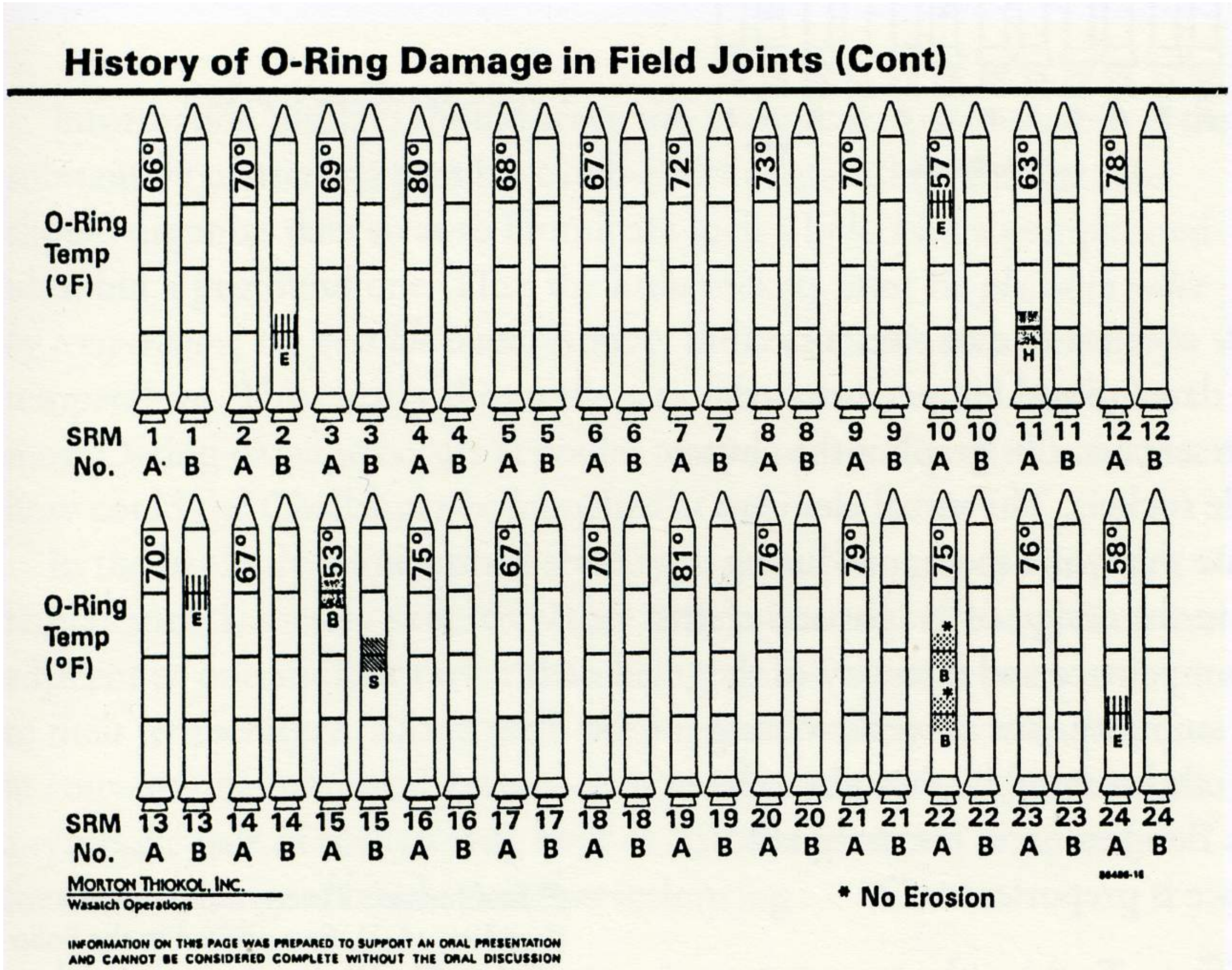
But Tufte thinks it may have been a matter of presentation & modeling:

- Never made the right tables or graphics
- Selected only failure data
- Never considered a simple statistical model

What do you think? How would you approach the data?

The *Challenger* launch decision

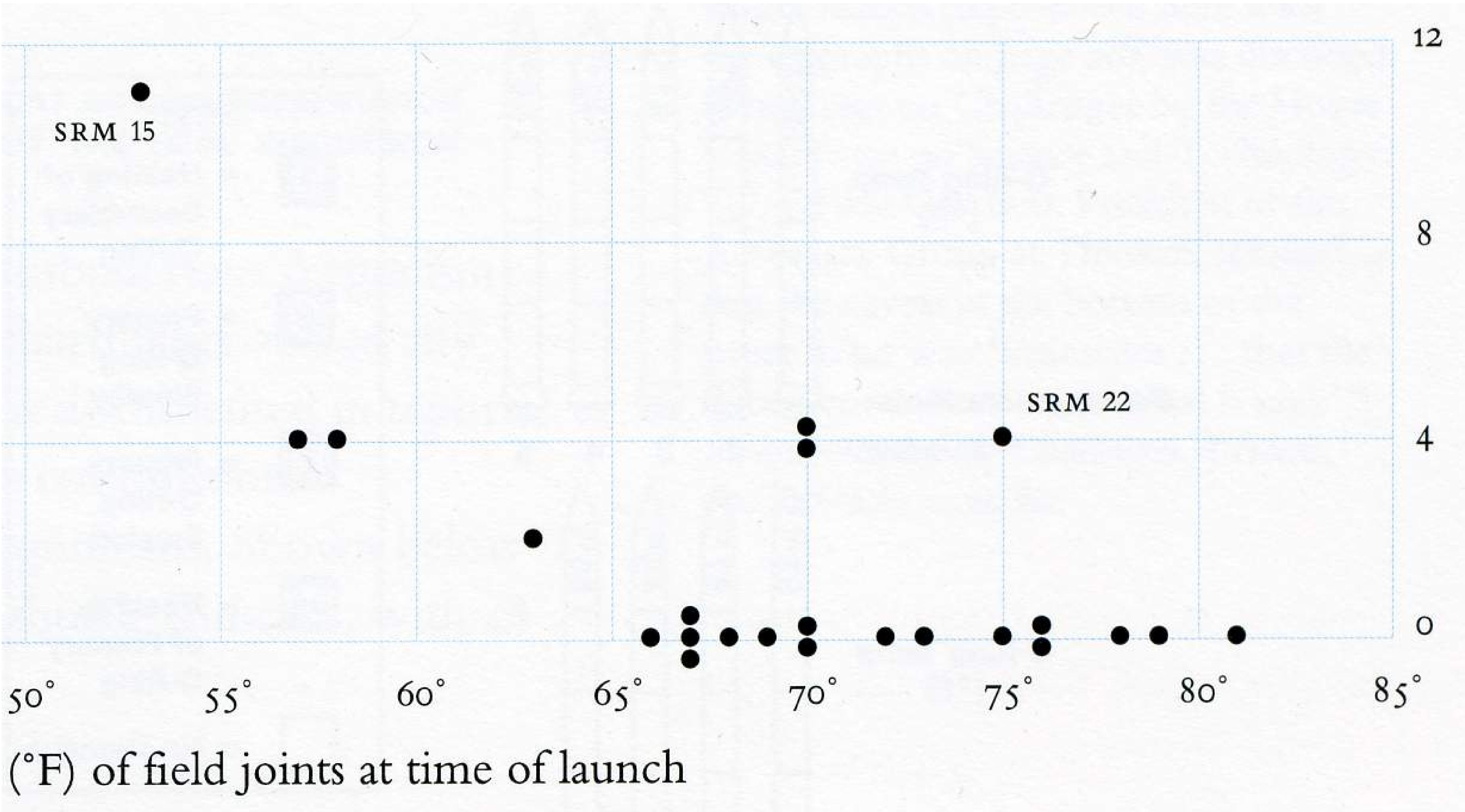
This is what Morton-Thiokol came up with to present after the disaster:



A marvel of poor design that obscures the data and makes analysis harder.

The *Challenger* launch decision

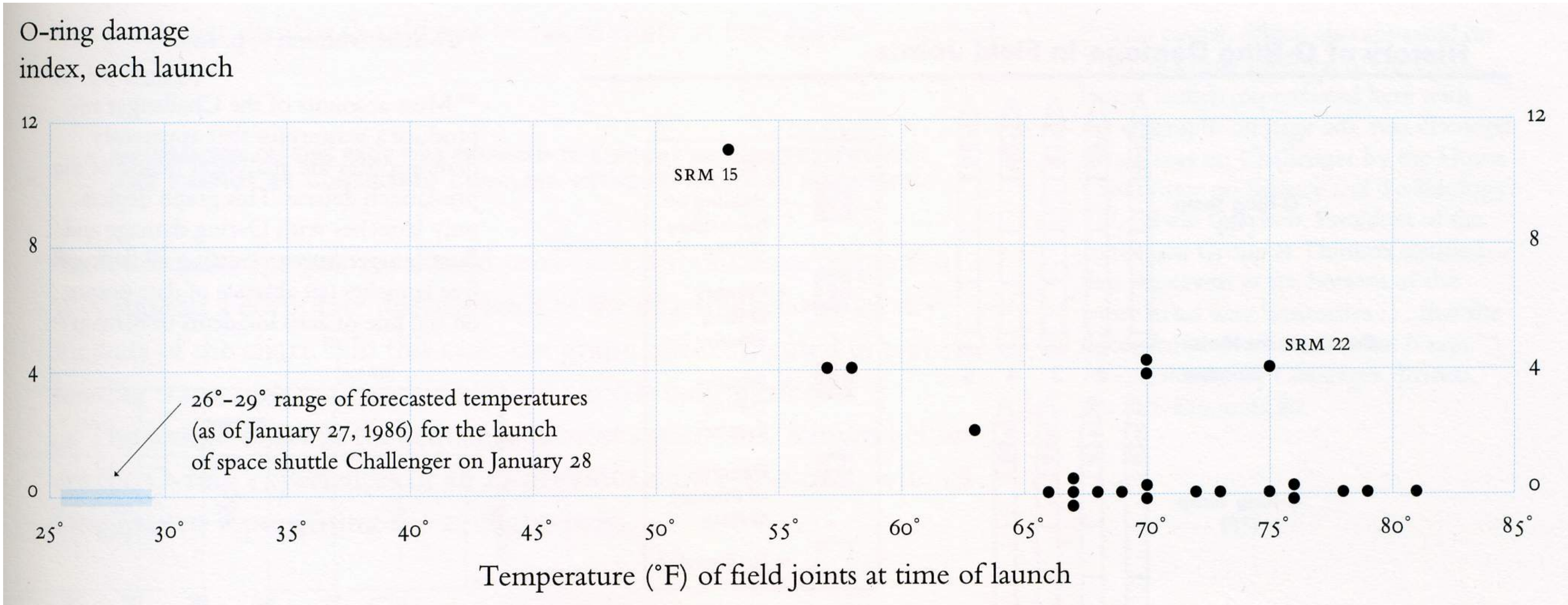
How about a scatterplot? Better for seeing relationships than a table



Suspicious. What was the forecast temperature for launch?

The *Challenger* launch decision

What was the forecast temperature for launch? 26 to 29 degrees Fahrenheit!



The *Challenger* launch decision

Imagine you are the analyst making the launch recommendation.

You've made the scatterplot above. What would you add to it?

Put another way, what do you is the first question you expect from your boss?

The *Challenger* launch decision

Imagine you are the analyst making the launch recommendation.

You've made the scatterplot above. What would you add to it?

Put another way, what do you is the first question you expect from your boss?

"What's the chance of failure at 26°?"

The scatterplot suggests the answer is "high", but that's vague.

But what if the next launch is at 58°? Or 67°?

Clearly, we want a more precise way to state the probability of failure

We need a *model*, and a way to convey that model to the public.

The *Challenger* launch decision

A simple exercise is to model the probability of O-ring damage as a function of temperature

We'll run a simple logit: $\Pr(\text{Damage}) = 1/(1 + \exp(-\text{Temp} \times \beta))$

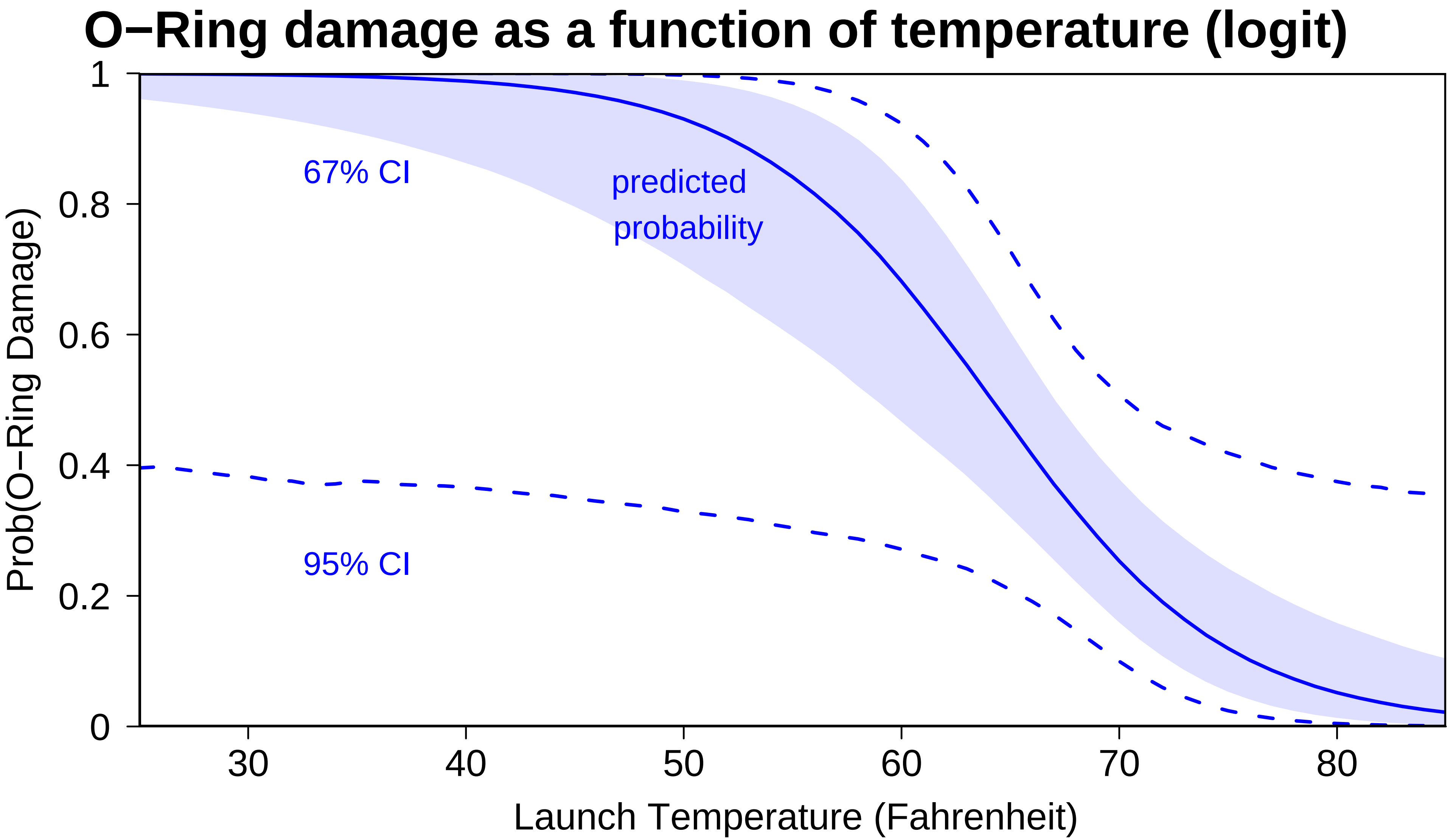
R gives us this lovely logit output. . .

Variable	est.	s.e.	p
Temperature (F)	-0.18	0.09	0.047
Constant	11.9	6.34	0.062
N	22		
log-likelihood	-10.9		

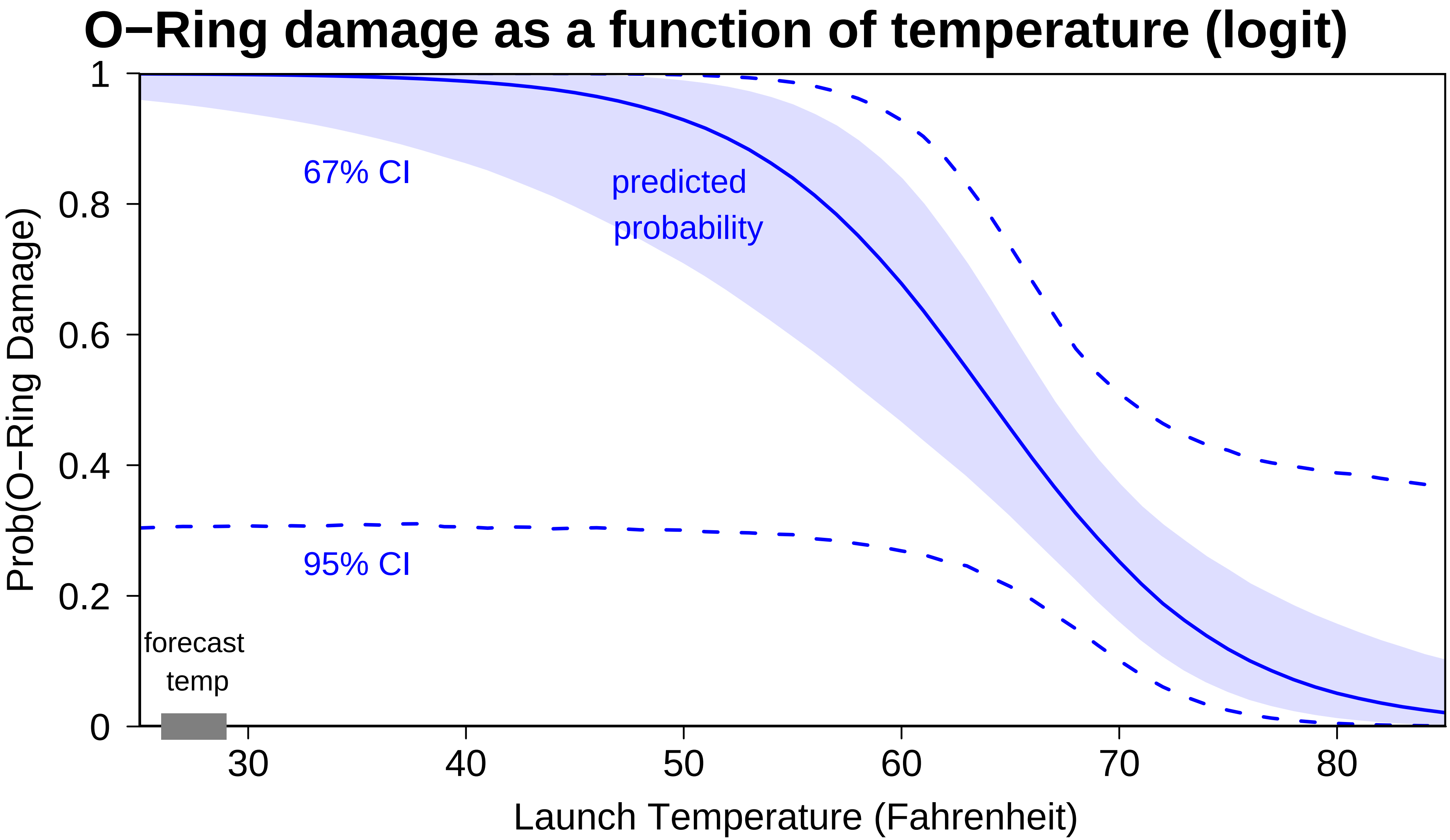
which most social scientists read as "a statistically significant negative relationship b/w temperature and probability of damage"

But that's pretty vague too.

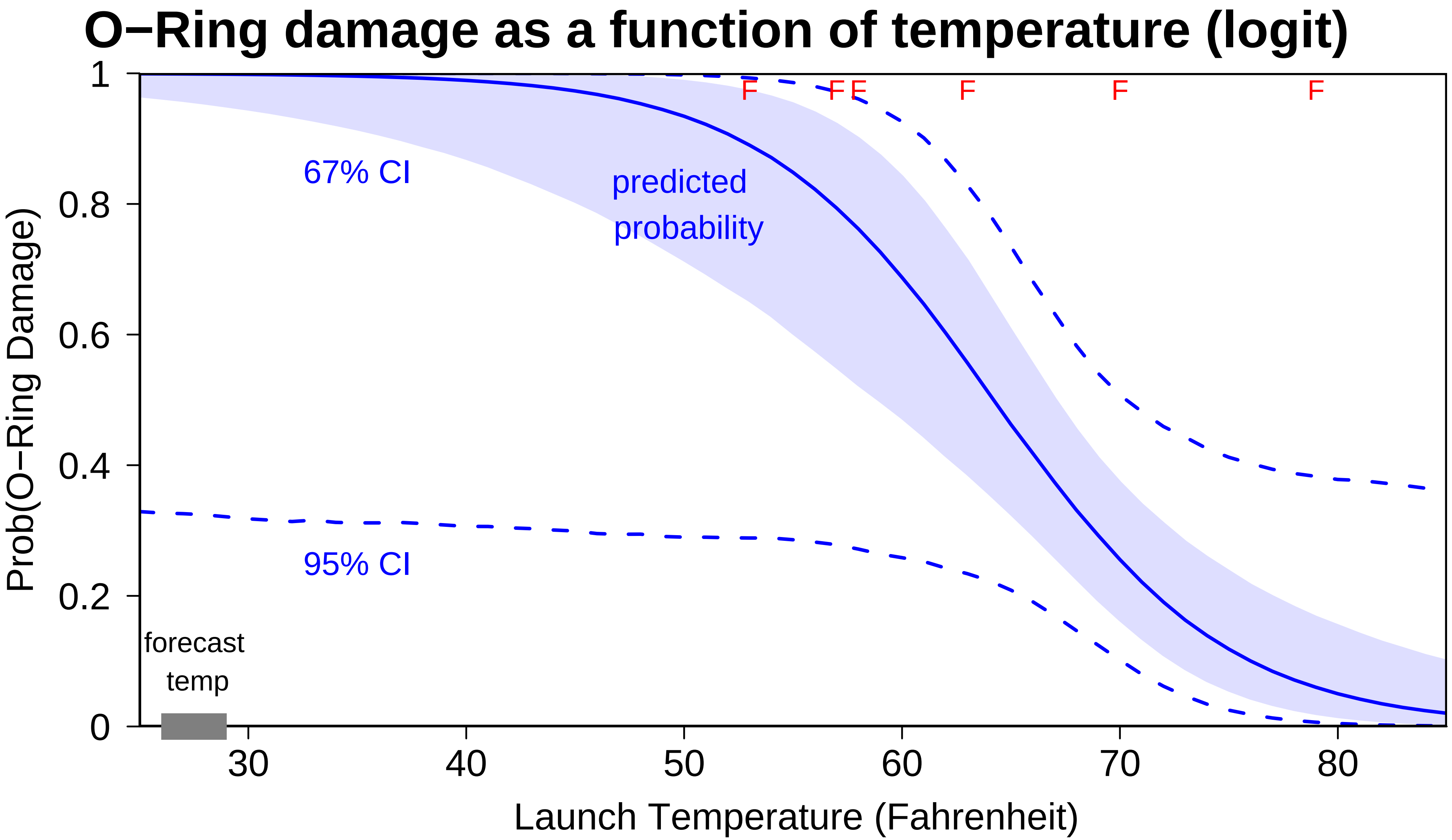
Is there a more persuasive/clear/useful way to present these results?



A picture shows model predictions *and* uncertainty

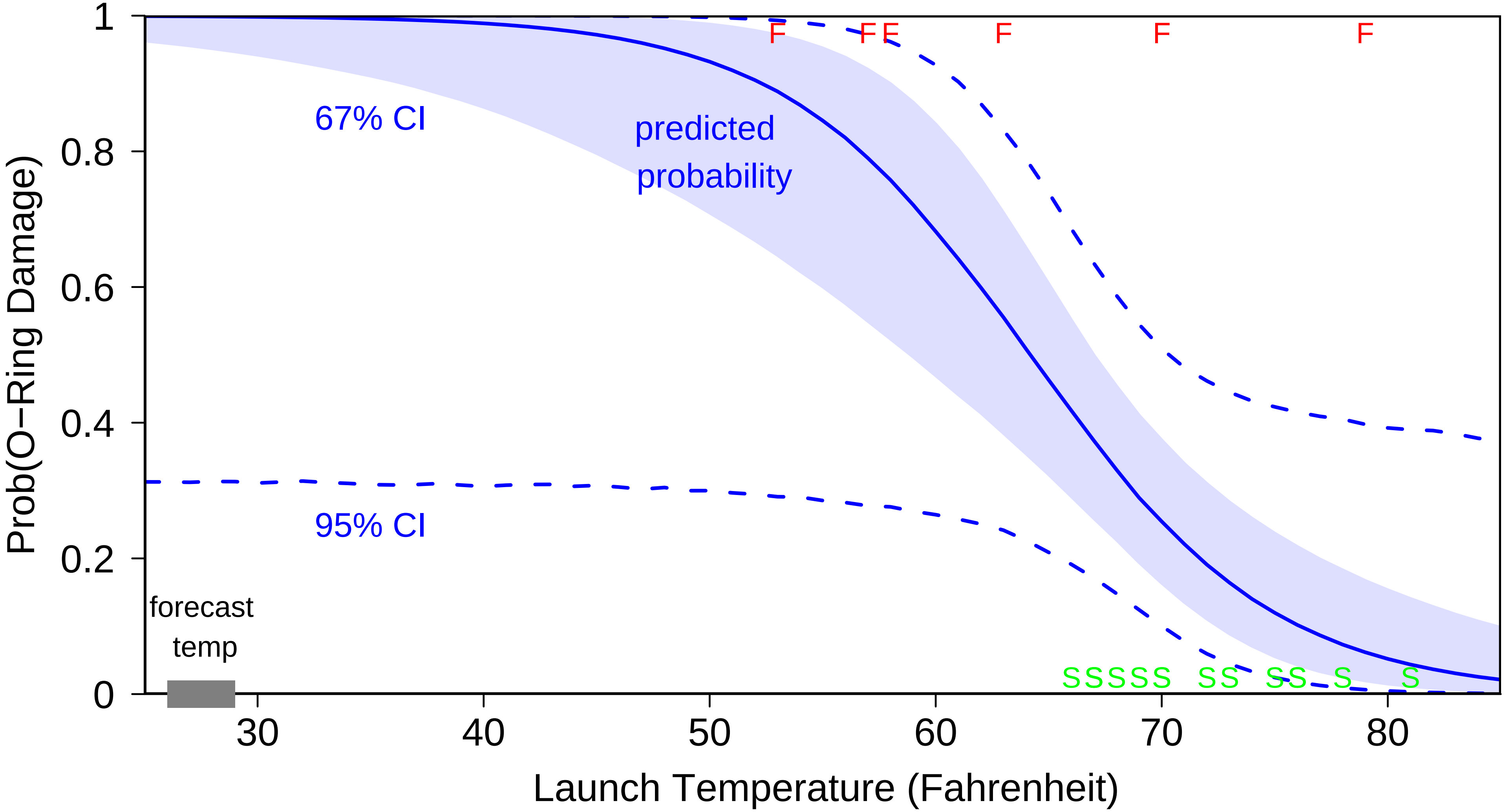


And gives a more precise sense of how foolhardy launching at 29 F is.

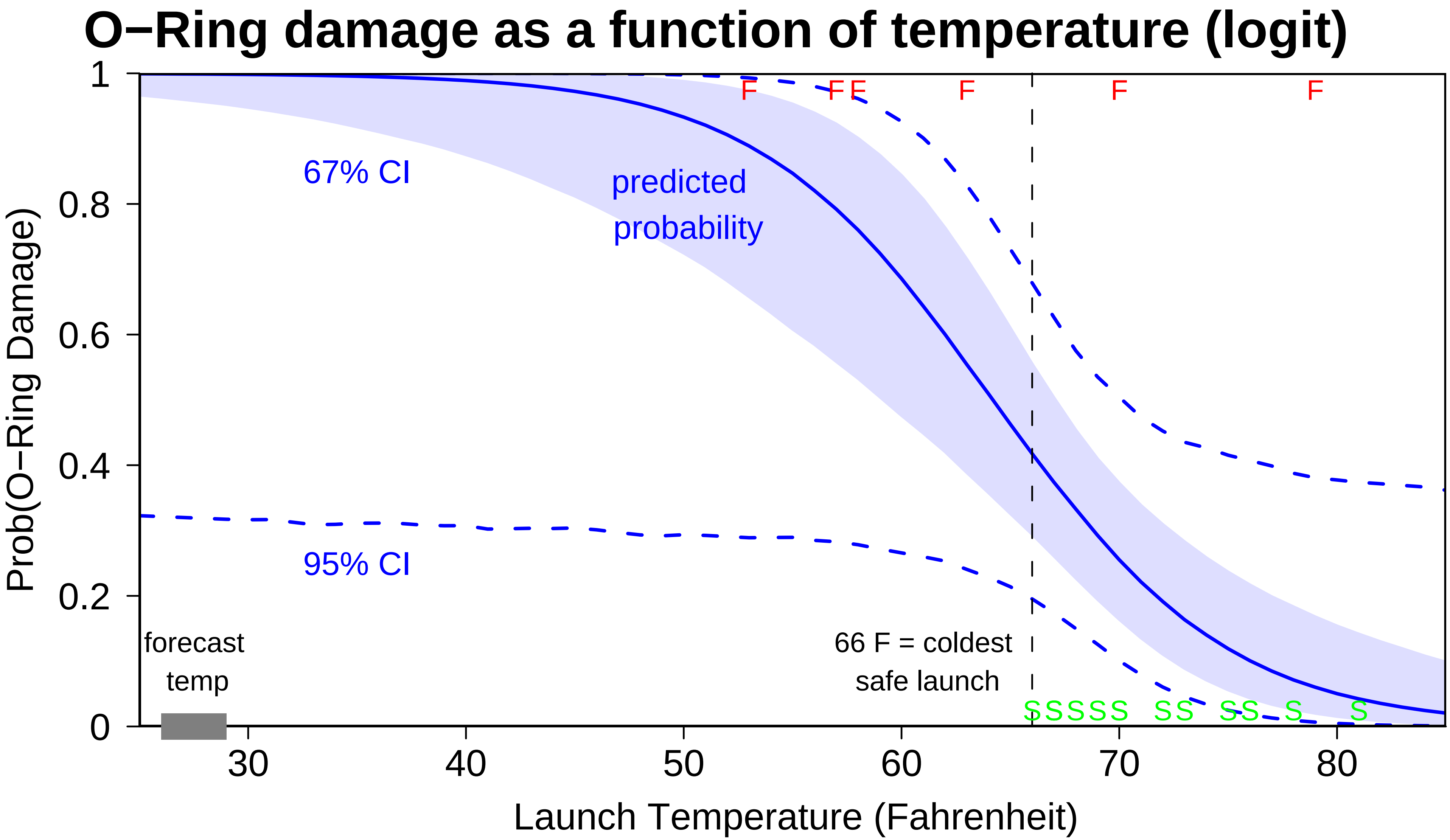


It's also good to show the data giving rise to the model.

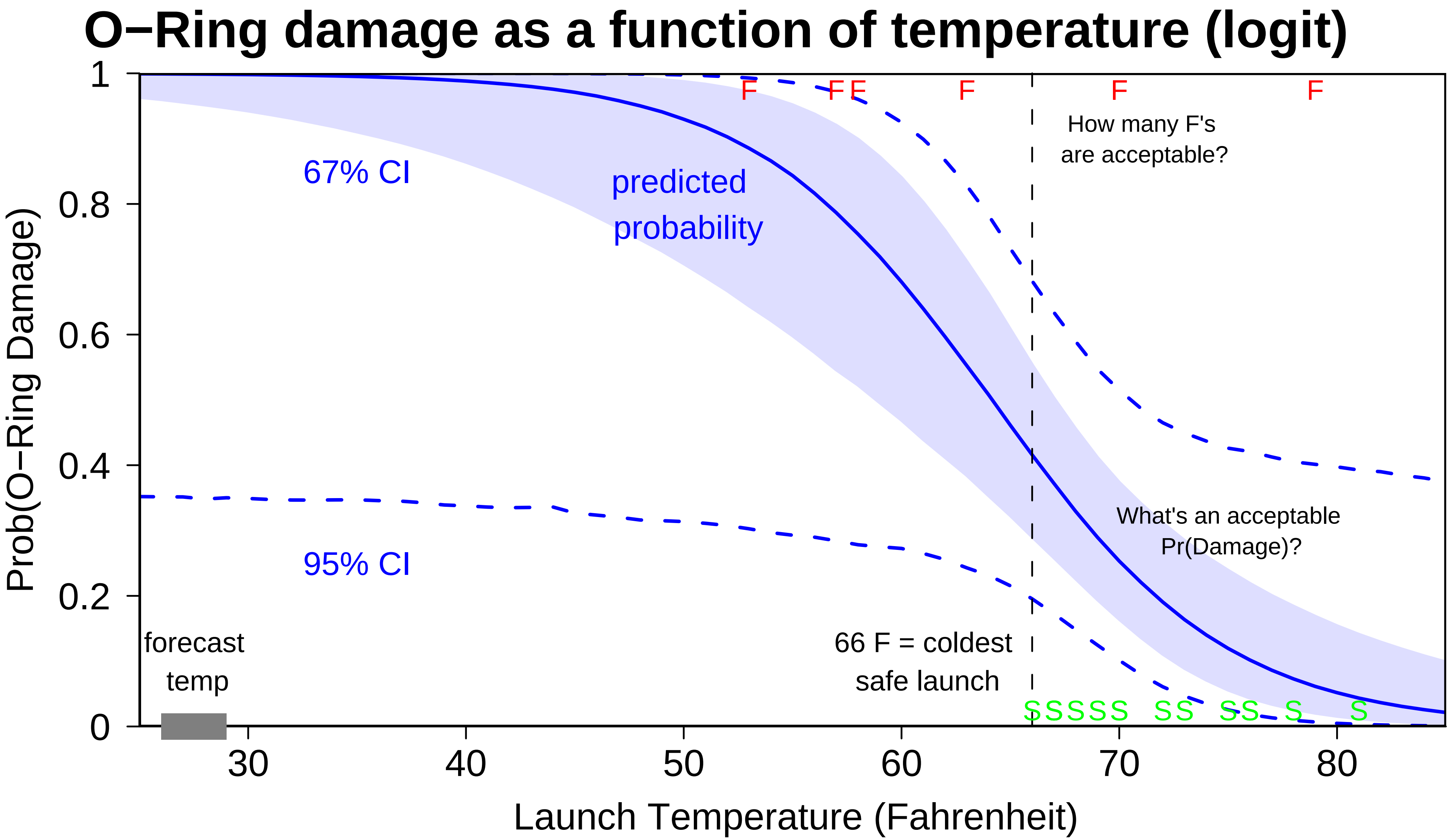
O−Ring damage as a function of temperature (logit)



Remembering that the Failures are only meaningful compared to Successes

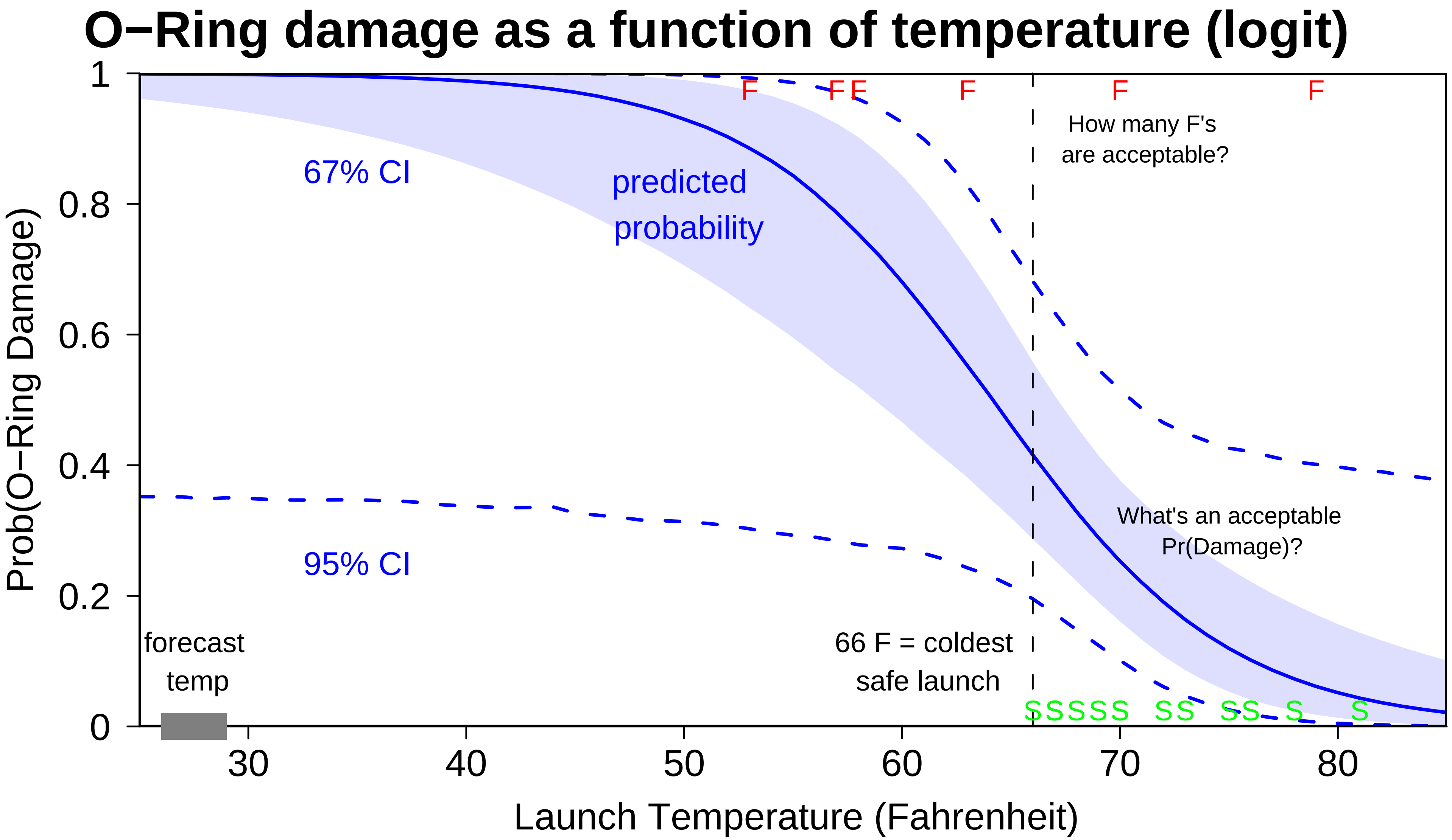


Looking just at the data tempts us to say that launches under 66° F are virtually guaranteed failures. This inference is based on an unstated model.



But the estimated logit model should give us pause.

There is a significant risk of failure across the board.



What is an acceptable risk of O-ring failure?

Was the shuttle safe at any temperature?

In a hearing, Richard Feynmann dramatically showed O-rings lose resilence when cold by dropping one in his ice water.

Experiment cut through weeks of technical gibberish concealing O-ring flaws

But it shouldn't take a Nobel laureate to explain a bivariate relationship:
any scientist with a year of statistical training could have
used the launch record to reach the same conclusion

And it would take no more than a single graphic to show the result

The *Challenger* launch decision

Lessons for social scientists:

Even relatively simple models and data are easier to understand with visuals

Tables can hide strong correlations

Imagine what might be hiding in datasets with dozens of variables?

Or in models with complex functional forms?

Visuals help make discussion more substantive

See the size of the effect, not just the sign

Make relative judgments of the importance of covariates

Make measured assesments of uncertainty (not just "accept/reject")

Enough with the pat examples!

John Snow & the *Challenger* are famous VDQIs from the natural sciences

Note they are both essentially bivariate

No confounders, no functional form issues, no interactions (well. maybe. . .)

None of the complexity that makes social science fun/frustrating

So let's look at how VDQIs improve a typical social science analysis

American interest rate policy

From my work on central banking (*The Myth of Bureaucratic Neutrality*)

Federal Reserve Open Market Committee (FOMC) sets interest rates $10\times/$year

Members of the FOMC vote on the Chair's proposed interest rate

Dissenting voters signal whether they would like a higher or lower rate

Dissents are rare, but may be symptomatic of how the actual rate gets chosen

Many factors might influence interest rate votes:

Individual	Career background Appointing party Interactions of above
Economy	Expected inflation Expected unemployment
Politics	Election cycles

American interest rate policy

My main concern is the individual determinants, esp. career background

Career background is a composite variable

Fractions of career spent in each of 5 categories:

Financial Sector	(FinExp)
Treasury Department	(FMExp)
Federal Reserve	(CBExp)
Other Government	GovExp
Academic Economics	EcoExp

These 5 categories plus an (omitted) Other must sum to 1.

American interest rate policy

Effect of the composition constraint:

If we want to consider the effects of a change in one category, we have to adjust the other categories simultaneously.

	Initial Composition		Hypothetical New Composition
FinExp	0.1	ΔFinExp	0.25
GovExp	0.3	$= 0.15$	0.25
FMExp	0.1		0.083
CBExp	0.2	$\rightarrow$	0.167
EcoExp	0.3		0.25
Sum	1.0		1.00

Above transitions (uniquely) preserve ratios among all categories except FinExp.

(See `rpcf()` in my `simcf` package for `R`)

Effect of a change in one category works thru *all* the βs for the composition

American interest rate policy

We'll fit an ordered probit model to the interest rate data:

$$\begin{aligned}
\Pr(y_i = j | \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\tau}) &= \int_{\tau_{j-1}}^{\tau_j} \mathrm{Normal}\left(\mathbf{x}_i \boldsymbol{\beta}, 1\right) \\
j &\in \{\mathrm{ease}, \mathrm{assent}, \mathrm{tighten}\}
\end{aligned}$$

Don't worry if this model is unfamiliar;
suffice it to say we have a nonlinear model and not just linear regression

American interest rate policy

Running the model yields the following estimates:

Response variable: FOMC Votes (1 = ease, 2 = accept, 3 = tighten)					
EVs	**param.**	**s.e.**	**EVs**	**param.**	**s.e.**
FinExp	−0.021	(0.146)	E(Inflation)	0.019	(0.015)
GovExp	−0.753	(0.188)	E(Unemployment)	−0.035	(0.022)
FMExp	−1.039	(0.324)	In-Party, election year	−0.182	(0.103)
CBExp	−0.142	(0.141)	Republican	−0.485	(0.102)
EcoExp × Repub	0.934	(0.281)	Constant	2.490	(0.148)
EcoExp × Dem	−0.826	(0.202)	Cutpoint (τ)	3.745	(0.067)
N	2957		ln likelihood	−871.68	

Table 1: **Problematic presentation: FOMC member dissenting votes—Ordered probit parameters.** Estimated ordered probit parameters, with standard errors in parentheses, from the regression of a $j = 3$ category variable on a set of explanatory variables (EVs). Although such nonlinear models are often summarized by tables like this one, especially in the social sciences, it is difficult to discern the effects of the EVs listed at right on the probability of each of the j outcomes. Because the career variables *XXX*Exp are logically constrained to a unit sum, even some of the signs are misleading. The usual quantities of interest for an ordered probit model are not the parameters ($\boldsymbol{\beta}$ and τ), but estimates of $\Pr(y_j|\mathbf{x}_c, \boldsymbol{\beta}, \tau)$ for hypothetical levels of the EVs $\mathbf{x}_c$, which I plot in Figure 1.

American interest rate policy

Response variable: FOMC Votes (1 = ease, 2 = accept, 3 = tighten)					
EVs	**param.**	**s.e.**	**EVs**	**param.**	**s.e.**
FinExp	−0.021	(0.146)	E(Inflation)	0.019	(0.015)
GovExp	−0.753	(0.188)	E(Unemployment)	−0.035	(0.022)
FMExp	−1.039	(0.324)	In-Party, election year	−0.182	(0.103)
CBExp	−0.142	(0.141)	Republican	−0.485	(0.102)
EcoExp × Repub	0.934	(0.281)	Constant	2.490	(0.148)
EcoExp × Dem	−0.826	(0.202)	Cutpoint (τ)	3.745	(0.067)
N	2957		ln likelihood	−871.68	

How do we interpret these results?

Because the model is non-linear, interpreting coeficients as slopes ($\partial y/\partial x$) is grossly misleading

Moreover, the compositional variables are tricky:
If one goes up, the others must go down, to keep the sum $= 1$.

Finally, can't interpret interactive coefficients separately.

American interest rate policy

Response variable: FOMC Votes (1 = ease, 2 = accept, 3 = tighten)					
EVs	**param.**	**s.e.**	**EVs**	**param.**	**s.e.**
FinExp	−0.021	(0.146)	E(Inflation)	0.019	(0.015)
GovExp	−0.753	(0.188)	E(Unemployment)	−0.035	(0.022)
FMExp	−1.039	(0.324)	In-Party, election year	−0.182	(0.103)
CBExp	−0.142	(0.141)	Republican	−0.485	(0.102)
EcoExp × Repub	0.934	(0.281)	Constant	2.490	(0.148)
EcoExp × Dem	−0.826	(0.202)	Cutpoint (τ)	3.745	(0.067)
N	2957		ln likelihood	−871.68	

Looking at this table, two obvious question arise:

Tell me the effect of each covariate on the probability of each kind of vote

And give me confidence intervals or standard errors for those effects

American interest rate policy

Response variable: FOMC Votes (1 = ease, 2 = accept, 3 = tighten)					
EVs	**param.**	**s.e.**	**EVs**	**param.**	**s.e.**
FinExp	−0.021	(0.146)	E(Inflation)	0.019	(0.015)
GovExp	−0.753	(0.188)	E(Unemployment)	−0.035	(0.022)
FMExp	−1.039	(0.324)	In-Party, election year	−0.182	(0.103)
CBExp	−0.142	(0.141)	Republican	−0.485	(0.102)
EcoExp × Repub	0.934	(0.281)	Constant	2.490	(0.148)
EcoExp × Dem	−0.826	(0.202)	Cutpoint (τ)	3.745	(0.067)
N	2957		ln likelihood	−871.68	

Cruel to ask of the reader: it's a lot of work to figure out.

The table above, though conventional, is an intermediate step.

Like stopping where Morton-Thiokol did, with pages of technical gibberish

The answers are there, but buried

American interest rate policy

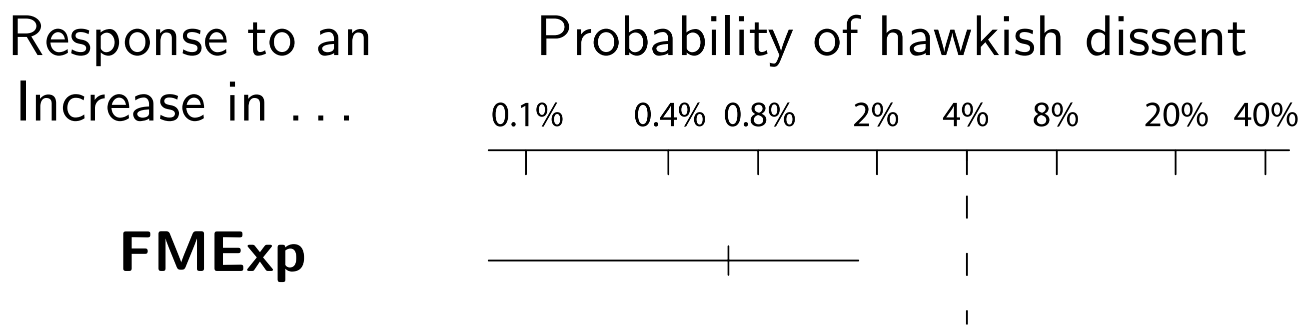
Response variable: FOMC Votes (1 = ease, 2 = accept, 3 = tighten)					
EVs	**param.**	**s.e.**	**EVs**	**param.**	**s.e.**
FinExp	−0.021	(0.146)	E(Inflation)	0.019	(0.015)
GovExp	−0.753	(0.188)	E(Unemployment)	−0.035	(0.022)
FMExp	−1.039	(0.324)	In-Party, election year	−0.182	(0.103)
CBExp	−0.142	(0.141)	Republican	−0.485	(0.102)
EcoExp × Repub	0.934	(0.281)	Constant	2.490	(0.148)
EcoExp × Dem	−0.826	(0.202)	Cutpoint (τ)	3.745	(0.067)
N	2957		ln likelihood	−871.68	

As the researcher, I should calculate the effects and uncertainty

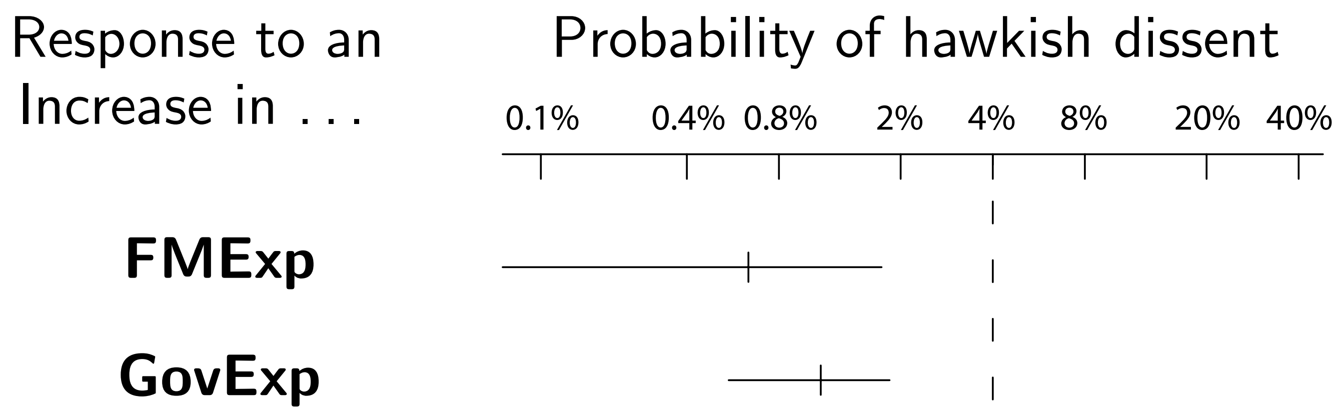
And present them in a readable way

A single graphic achieves both goals:

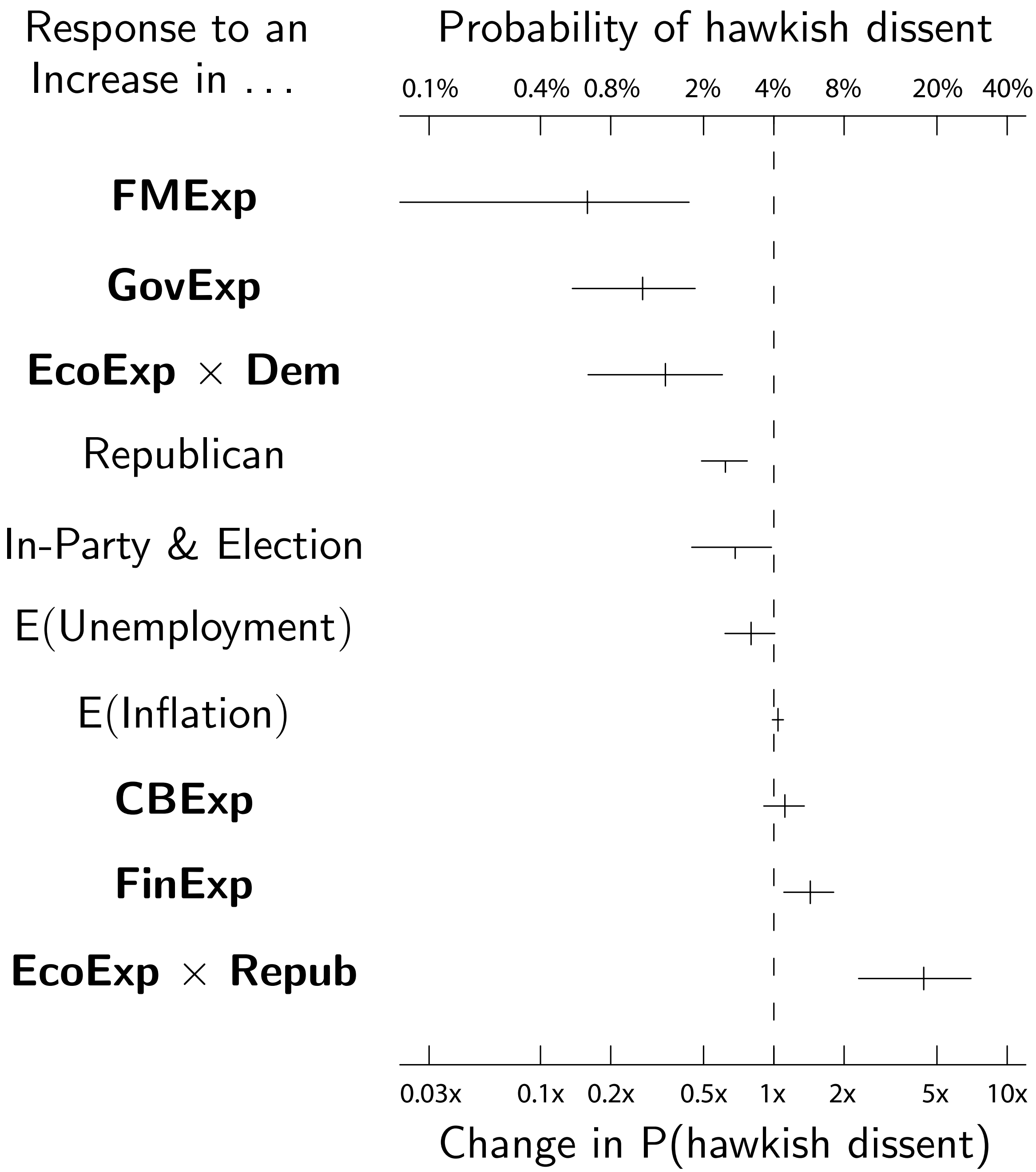
American interest rate policy



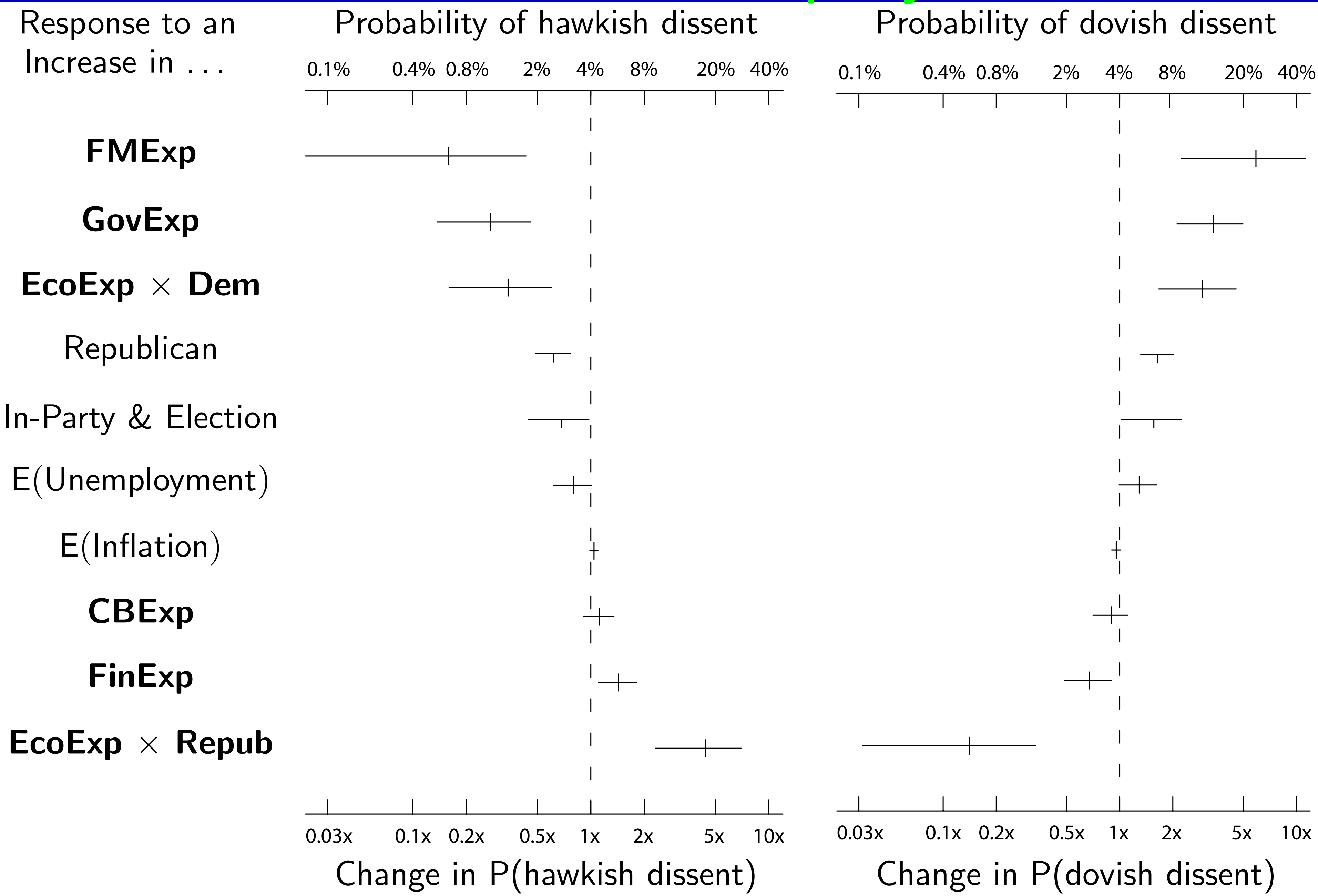
American interest rate policy



American interest rate policy



American interest rate policy



American interest rate policy

For reference, here is the caption.

For complex or unfamiliar graphics, long captions are encouraged!

Figure 1: **One solution: Probability of casting a dissenting vote on the FOMC.** First differences calculated from an ordered probit model of votes by Fed Open Market Committee members on interest rate policy ($1 =$ dissent in favor of lower interest rates, $2 =$ accept the proposed interest rate, $3 =$ dissent in favor of higher rates). For each of the explanatory variables (EVs) listed at left, plots show the effect of increasing that variable on the probability of dissenting votes favoring tightening (left plot) or easing of interest rates (right plot). The increase in the listed EV is $+1$ unit, except for career backgrounds (listed in bold), which are raised to their maximum of one. All EVs besides the listed EV are set at their means, unless this is logically impossible (e.g., other career backgrounds are set at 0 to maintain the unit sum). Probabilities under each hypothetical ($\mathrm{Pr}(y_j|\mathbf{x}_\mathrm{c}, \boldsymbol{\beta}, \tau)$) are plotted on the top scale; the relative probability compared to the scenario with all EVs set to mean levels ($\mathrm{Pr}(y_j|\mathbf{x}_\mathrm{c}, \boldsymbol{\beta}, \tau)/\mathrm{Pr}(y_j|\bar{\mathbf{x}}, \boldsymbol{\beta}, \tau)$) is shown on the bottom scale. Horizontal bars mark 90 percent confidence intervals. Scales are in $\log_{10}$ units.

A taste of things to come

No matter how complex the model, you can always summarize the relationship between x and y with pictures

Well designed VDQIs make complex models (linear or nonlinear) transparent

Just calculate $\mathrm{E}(y|\mathbf{x}_c, \hat{\boldsymbol{\beta}})$ for interesting cases $\mathbf{x}_c$, and plot them

Also allow easy presentation of uncertainty, via simulation:

- calculate $\mathrm{E}(y|\mathbf{x}_c, \tilde{\hat{\boldsymbol{\beta}}})$ for different draws of $\tilde{\hat{\boldsymbol{\beta}}}$ from the estimated model

- build up the posterior or predictive distribution of $\mathrm{E}(y|\mathbf{x}_c, \hat{\boldsymbol{\beta}})$

Any intelligent non-specialist should be able to understand your (fancy/Bayesian/dynamic/hierarchical/non-linear/interactive) model

If they can't, you're not finished writing it up,
and may be missing some implications yourself!

Scope of the class

It may sound like this course covers all of applied statistics

Because visual displays can be woven thoughout all empirical science

Course goal: *complement* your other statistical training

Start by defining Visual Displays of Quantitative Information & their uses

What is a VDQI?

Almost any representation of information is a VDQI; not just graphics:

- A plot
- A table
- A confection of plots and/or tables
- A schematic
- An equation
- A paragraph

When do we use VDQIs?

VDQIs are woven through the practice of quantitative methods:

- Exploring data
- Interpreting models
- Checking model assumptions & fit
- Persuading an audience
- Making a result memorable

How do VDQIs convey infomration?

VDQIs can present massive amounts of data for different ends:

- for lookup
- for posterity
- for gestalt impressions
- for exploration
- for rigorous comparison

The appropriate visuals vary by task

Who uses the VDQIs the researcher designs?

- The researcher herself
- The expert reader
- Decision makers
- The general public

Different VDQIs may be best suited for each audience

So how do I choose?

Some VDQIs will be more powerful than others for a particular purpose

But some VDQIs are generical well suited to some tasks

Tables are usually good for lookup, bad for gestalt impressions

Some VDQIs are inherently powerful

Scatterplots show relations b/w two continuous variables richly and simply, and will never be bettered

Some VDQIs are inherently inferior

Pie charts are inefficient, awkward, and prone to misinterpretation

For fun, type `?pie` in `R`.

But designing good visuals is more than "Pie charts bad; Dot charts good"

Course outline