Data and Uncertainty in System Dynamics

Forrester, Kalman, Markov & Bayes

Tom Fiddaman
SD Seminar
October 26, 2022

Abstract

Jay Forrester cautioned that "fitting curves to past system data can be misleading." Certainly that can be true, if the model is deficient. But we can have our cake and eat it too: a good model that passes traditional SD quality checks and fits the data can yield unique insights. With recent computing advances, it's practical to confront models with all available information, including time series data, to yield the best possible estimate of the state of a system and its uncertainty. That makes it possible to construct policies that are robust not just to a few indicator scenarios, but to a wide variety of plausible futures. This talk will discuss how calibration, Kalman filtering, Markov Chain Monte Carlo and sensitivity analysis work together, with particular attention to Bayesian inference. The emphasis will be on practical implementation with a few examples from real projects.

Agenda

- **Introduction**
- **Example – Chronic Wasting Disease Policy**
- **Methods**
  - Naïve calibration
  - Maximum likelihood
  - Kalman filtering
  - Bayesian inference
  - Markov Chain Monte Carlo (MCMC)
  - Synthetic data
- **References**

A Classic SD Perspective on Data

- **fitting curves to past system data can be misleading**
- **given a model with enough parameters to manipulate, one can cause any model to trace a set of past data curves**
- **adjusting model parameters to force a fit to history may push those parameters outside of plausible values as judged by other available information.**
- **[tracing history] does not give greater assurance that the model contains the structure that is causing behavior in the real system**
- **the particular curves of past history are only a special case**
- **Exactly matching a historical time series is a weak indicator of model usefulness.**
- **We should not want the model to exactly recreate a sample of history but rather that it exhibit the kinds of behavior being experienced in the real system.**

System Dynamics—the Next Fifty Years, Jay W. Forrester, D-4892 (2007)

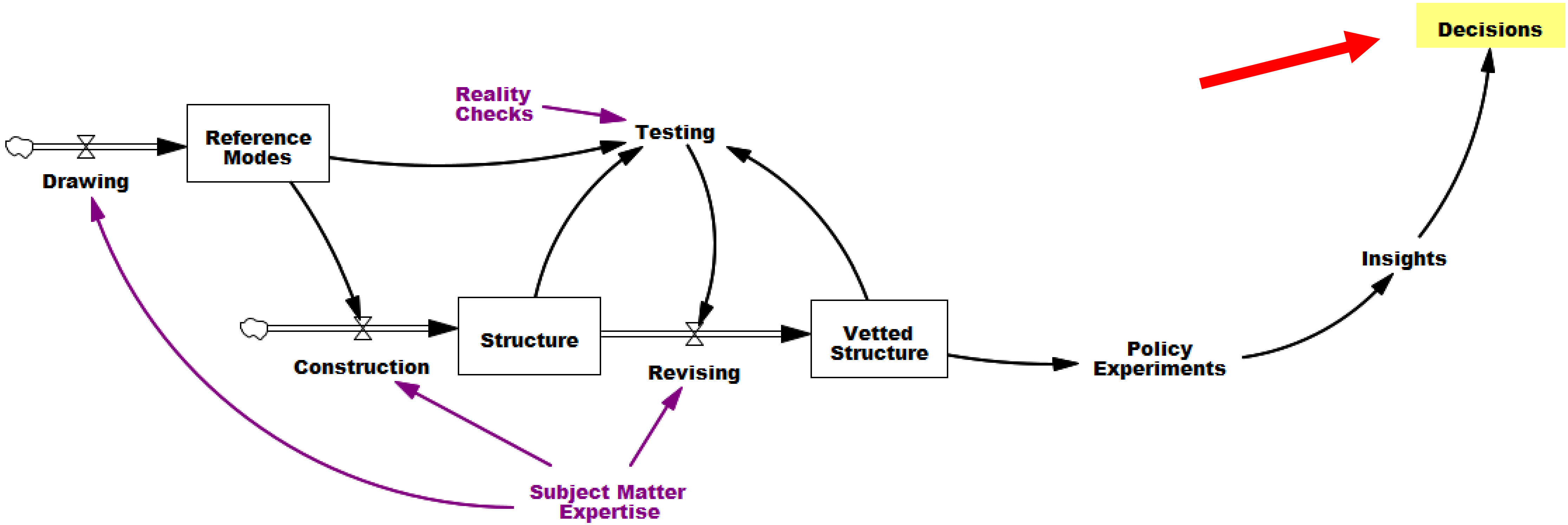
The Seven Critical Thinking Skills ... Plus a Few

Barry Richmond in the Systems Thinker

SYSTEMS THINKING SKILL	
Dynamic Thinking Framing a problem in terms of a pattern of behavior over time	Specify Problem/Issues
System-as-Cause Thinking Placing responsibility for a behavior on internal actors who manage the policies and plumbing of the system	Construct Hypothesis
Forest Thinking Believing that, to know something, you must understand the context of relationships	
Operational Thinking Concentrating on getting at causality and understanding how a behavior is actually generated	
Closed-Loop Thinking Viewing causality as an ongoing process, not a one-time event, with the "effect" feeding back to influence the causes, and the causes affecting each other	
Quantitative Thinking Accepting that you can always quantify, though you can't always measure	
Scientific Thinking Recognizing that all models are working hypotheses that always have limited applicability	Test Hypothesis

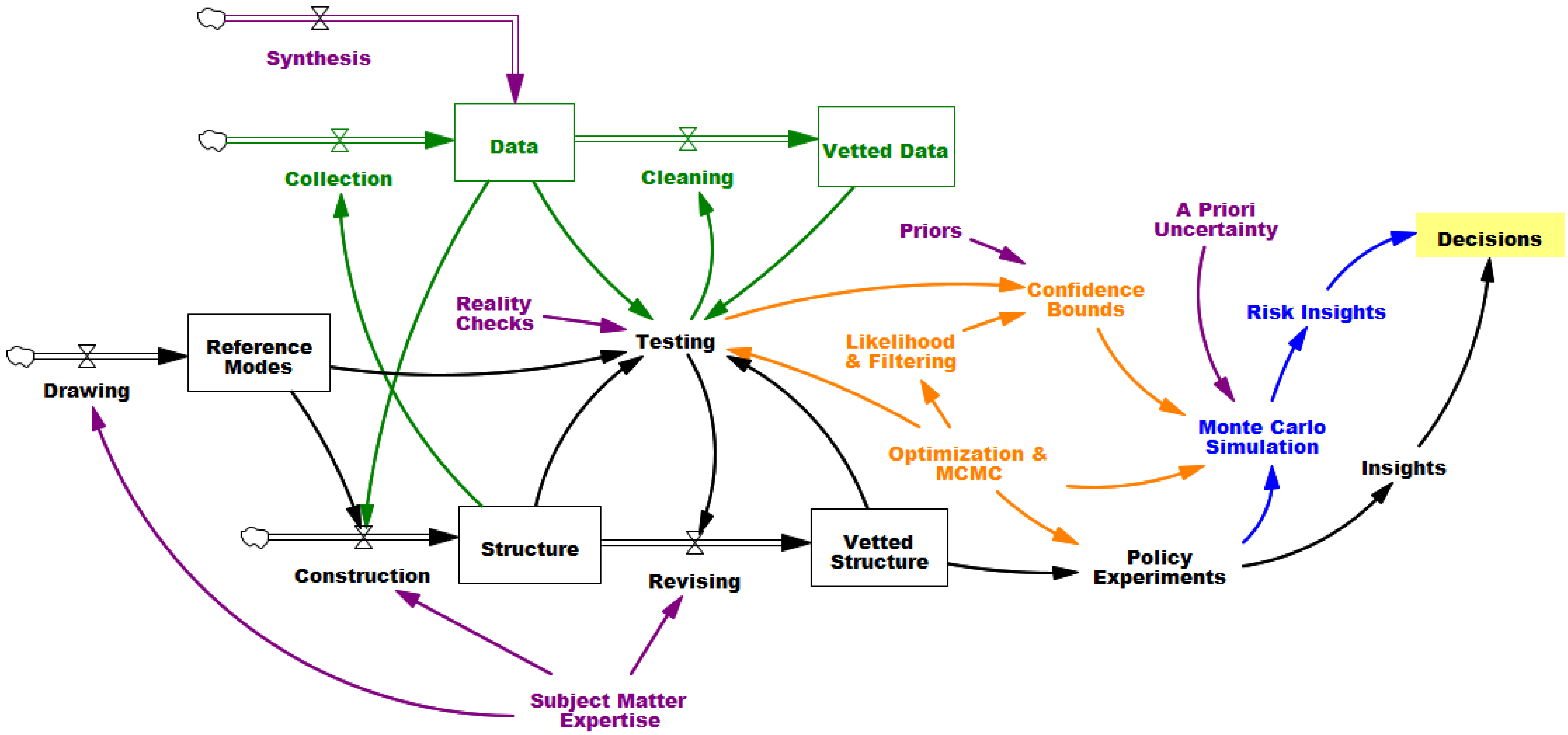
- **Statistical Thinking**
  - Understanding how noisy measurements inform and constrain your model
- **Solution Thinking**
  - Focusing on helping stakeholders to understand the problem structure so they can make better decisions
- **Behavioral Thinking**
  - Modeling decisions with available information and limited cognition
- **Complexity Thinking**
  - Recognizing that emergent behavior may involve granular detail

Classic SD

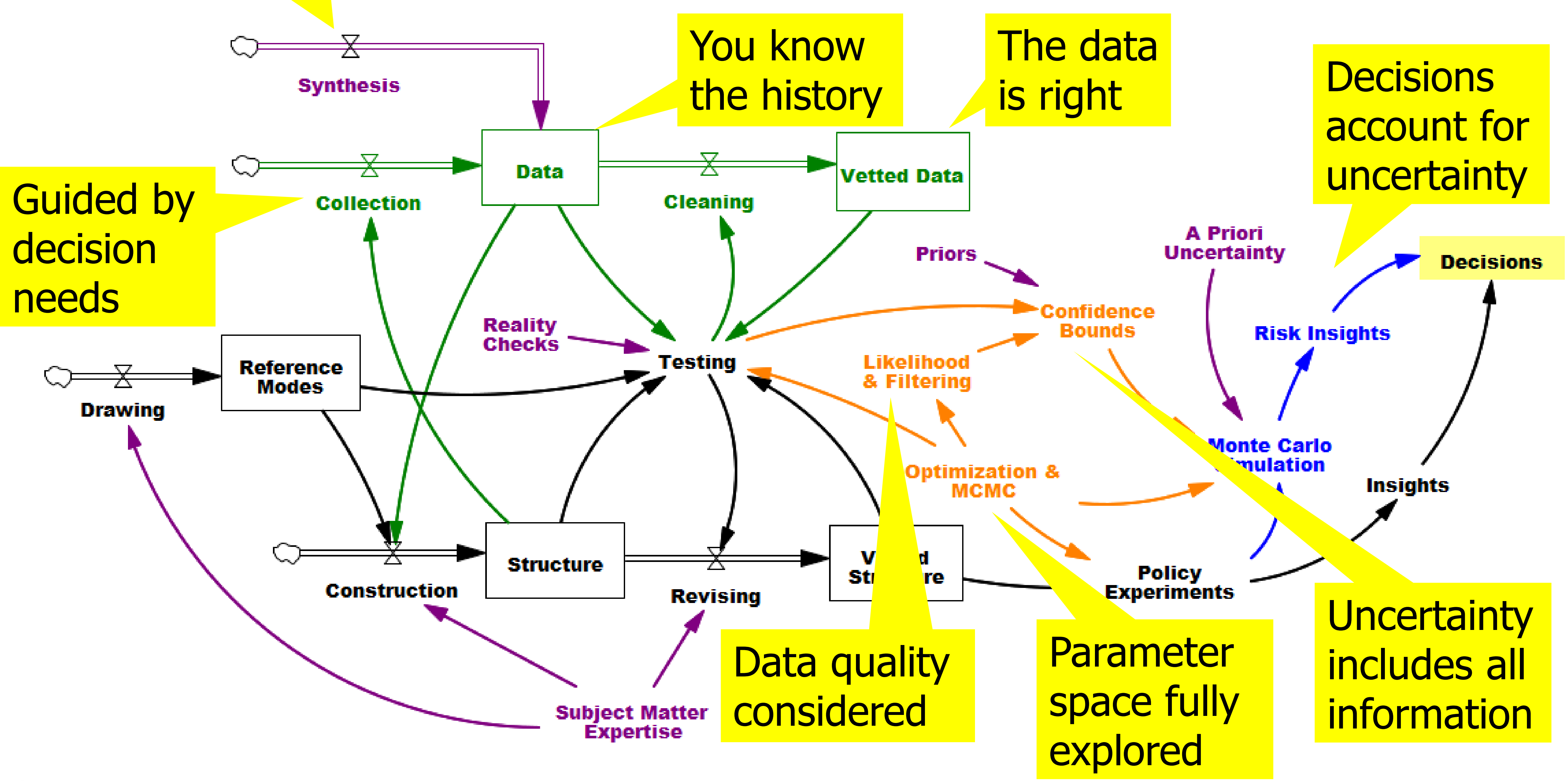


Decisions
Reality Checks
Testing
Reference Modes
Drawing
Insights
Construction
Structure
Revising
Vetted Structure
Policy Experiments
Subject Matter Expertise

Bayesian SD/Data Science



Bayesian SD/Data Science



Uses for Data

Informal - “data” construed broadly

- Informs model structure directly
- Submodel calibration informs parameter choices
- Did history happen as reference modes describe?
- Do people make decisions the way they say they do?
- Do interesting features of model behavior appear in reality?
- Convincing stories demonstrating model behavior

Formal – time series or cross sectional data

- Initialization of disaggregated states
- Verification of methods with synthetic data
- Estimation of parameters and uncertainty
- Propagation of uncertainty into outcomes
- Acceptance or rejection of candidate structures

What happens if you ignore the data?

+ **Save lots of time on collection, preprocessing and calibration**
  + Potentially reallocate to client interaction, robustness testing and scenario experimentation
+ **Less cognitive load for participants**

– **No learning about the data, or from the data directly**
– **No contribution to model quality from tests against data**
– **Hard to verify that asserted reference modes or decision structures match reality**
– **Less face validity of historical runs**
– **Difficulty understanding the gaps between a priori parameter values and most plausible values, given the model**
– **No objective basis for parameter values or confidence bounds**
– **Hard to understand the joint uncertainty of a parameter set**

Example: Chronic Wasting Disease Policy

- **Client: WI DNR via USGS NWHC**
- **Question: how best to use agency resources to reduce the prevalence and geographic spread of CWD?**
- **Stakeholders: hunters, landowners, captive cervid farmers, wildlife NGOs, waste disposal industry, tribes, other agencies …**
- **Process:**
  - Structured Decision Making (essentially, stakeholder identification of metrics of interest and ranking of alternative actions’ influence on each outcome)
  - Model informs action->metric mapping

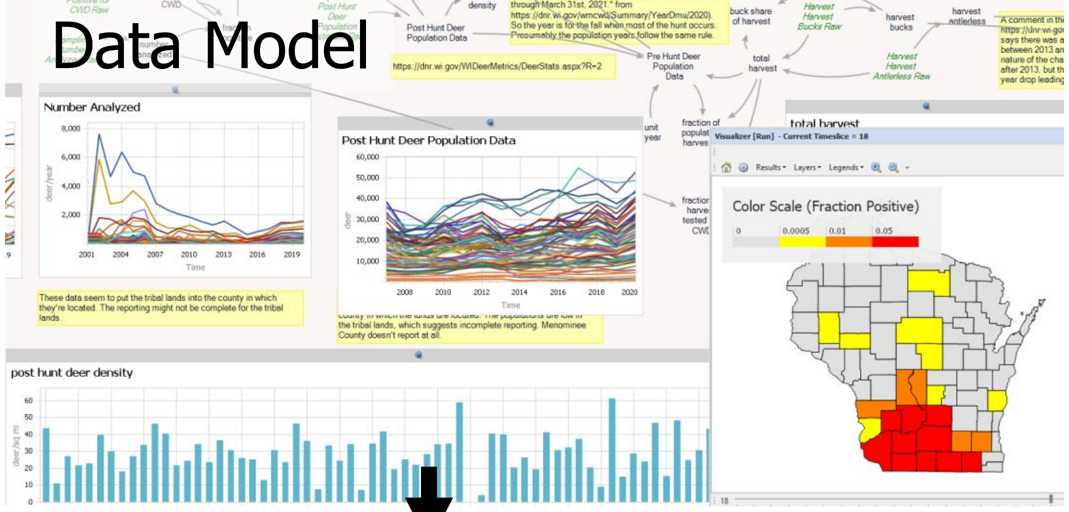
Chronic Wasting Disease

- Prion disease, like Mad Cow and scrapie
- Affects cervids (deer family)
- 100% fatal
- Long latent period, short clinical phase
- Environmental reservoir
- No human transmission …yet

VENTANA

Process

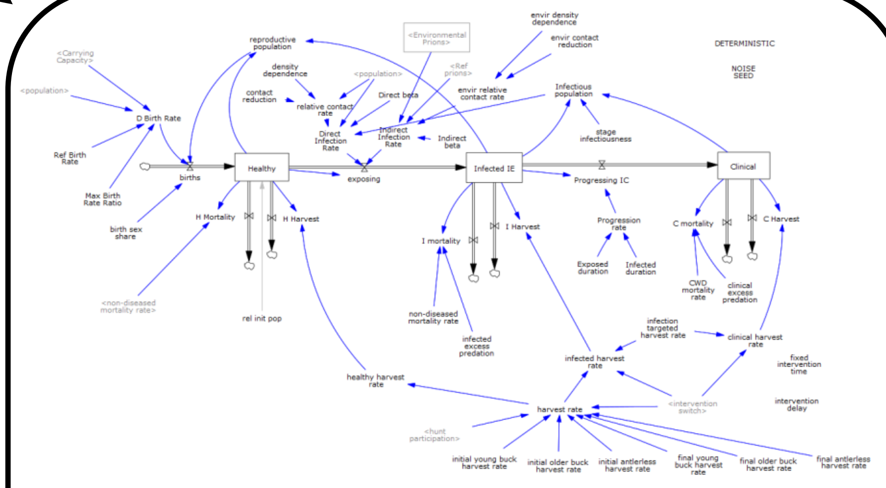
Data Model



Comparisons

Parameters
- CWD transmission
- Environmental prions
- Deer fertility & mortality

Decisions
- Hunting
- Baiting & feeding
- Surveillance
- Carcass management
- Safe practices
- Timing



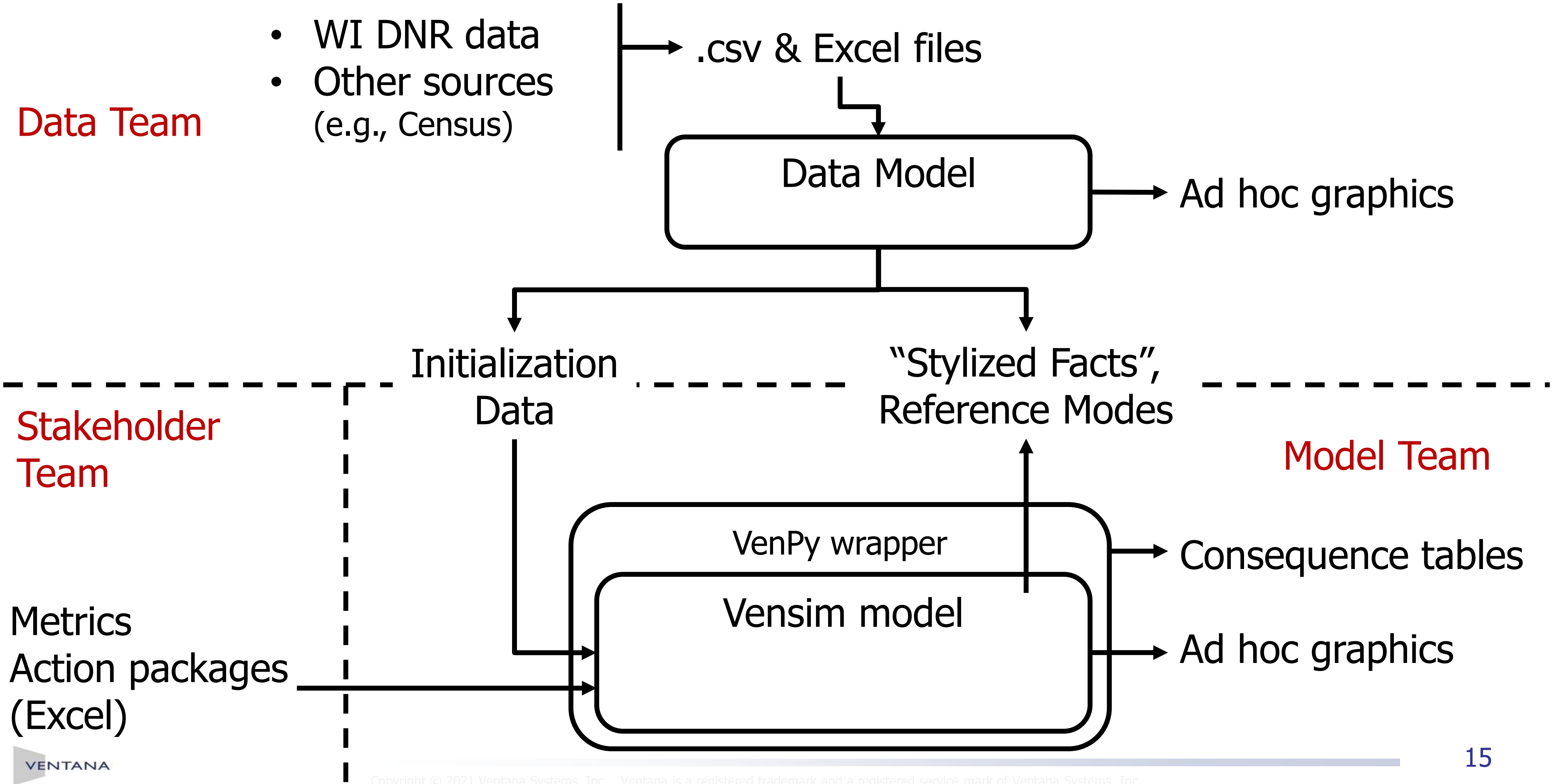
Model

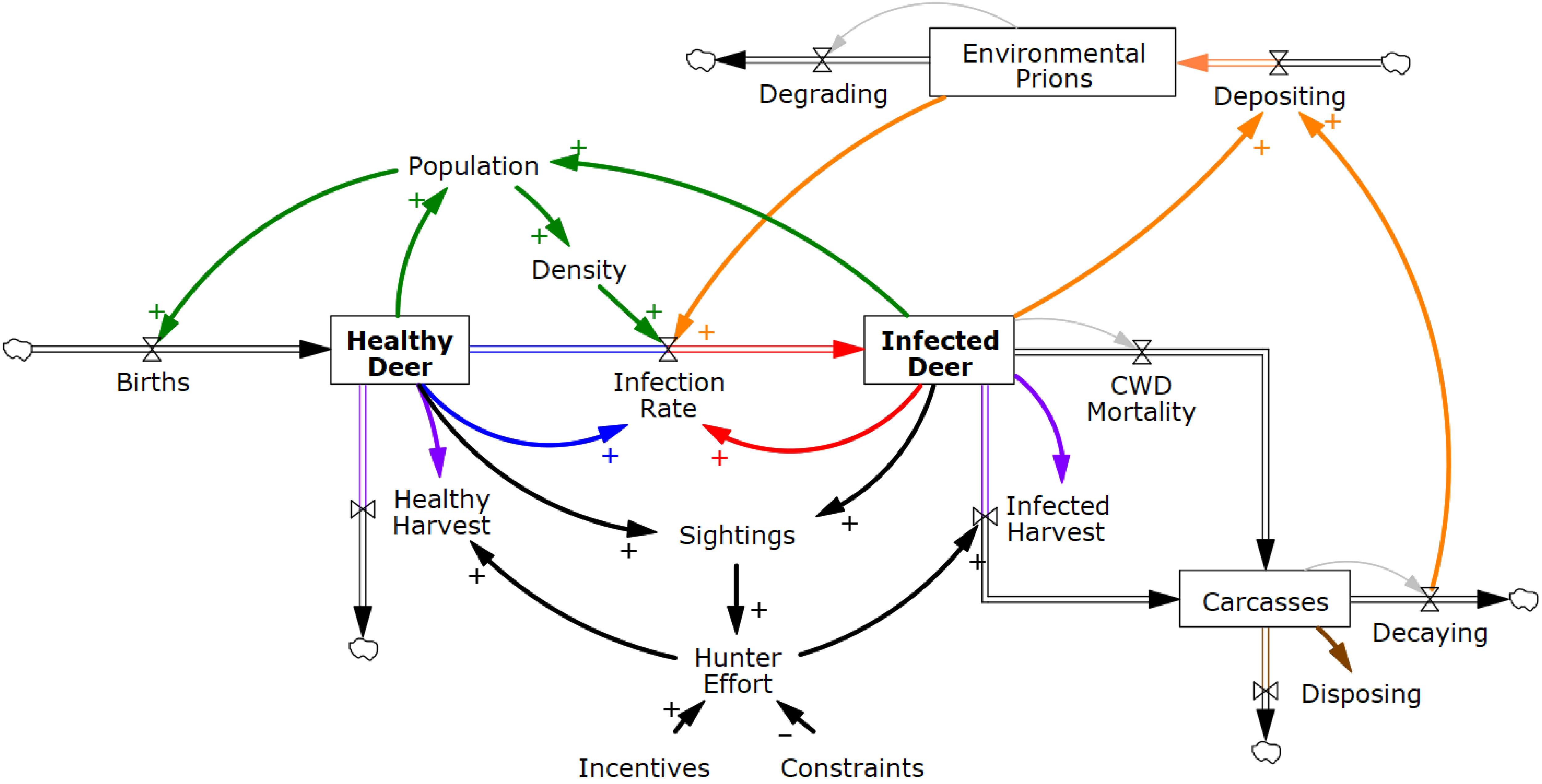
Outcomes
- Prevalence
- Fraction positive
- Population
- Age, sex structure
- Harvest
- Surveillance results
- Hunt effort
- Human exposure

Phase 1 Approach

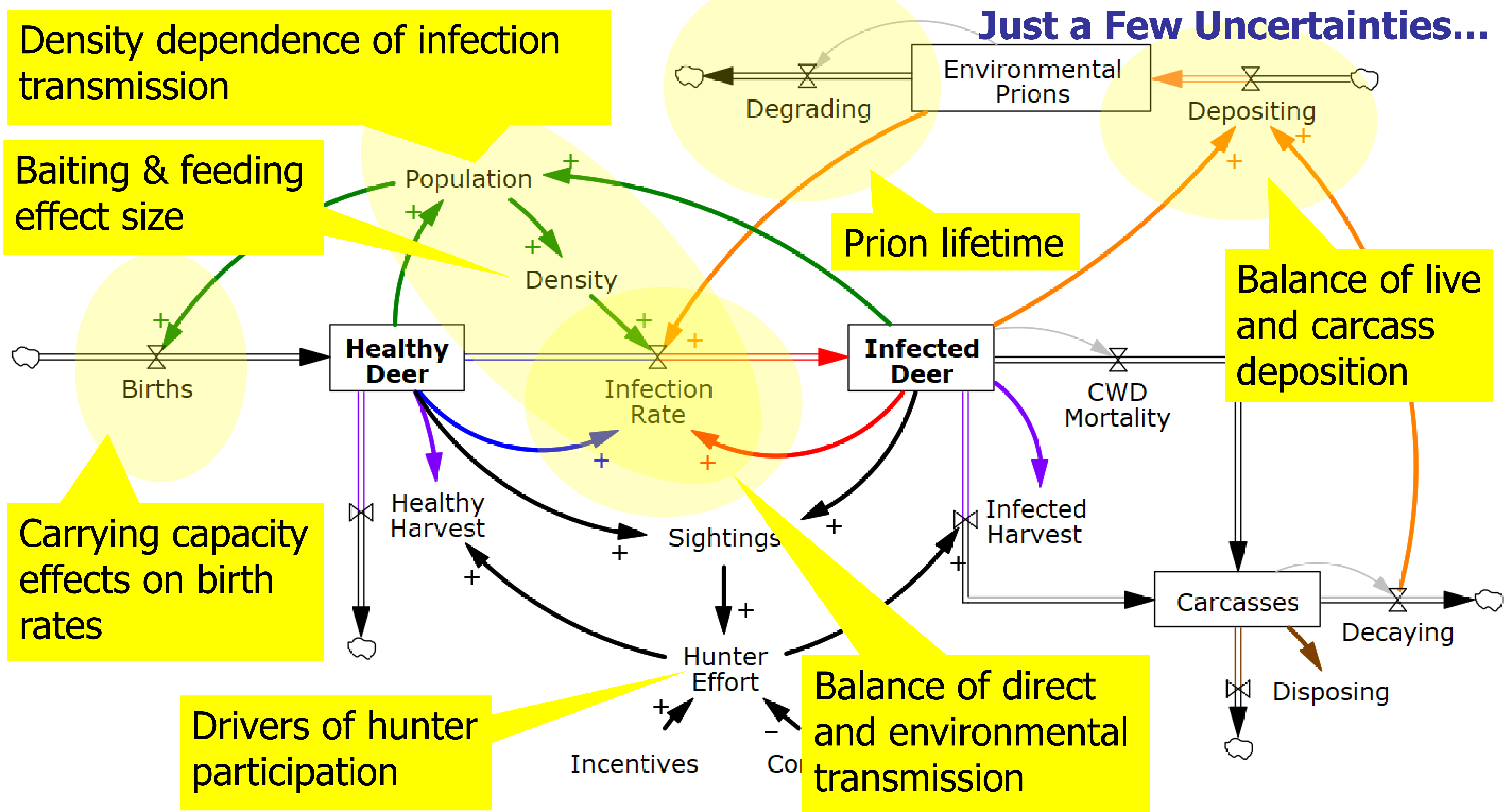
- **Situation**
  - Limited time and fixed schedule for stakeholder interactions
  - Heavy demands for scenario evaluation
  - Good precedent models in the literature
  - But large uncertainty about some features
- **Strategy**
  - No formal maximum likelihood calibration
  - Calibrate very loosely to replicate the range of disease prevalence growth rates observed in minimally-controlled situations
  - Establish parameters primarily from literature and subject matter experts
  - Develop notional uncertainties from:
    - Subject matter experts
    - Disagreement in the literature
    - Some experimentation with face validity of model results
  - Considerable use of data to describe behaviors seen in the model

CWD Project Architecture



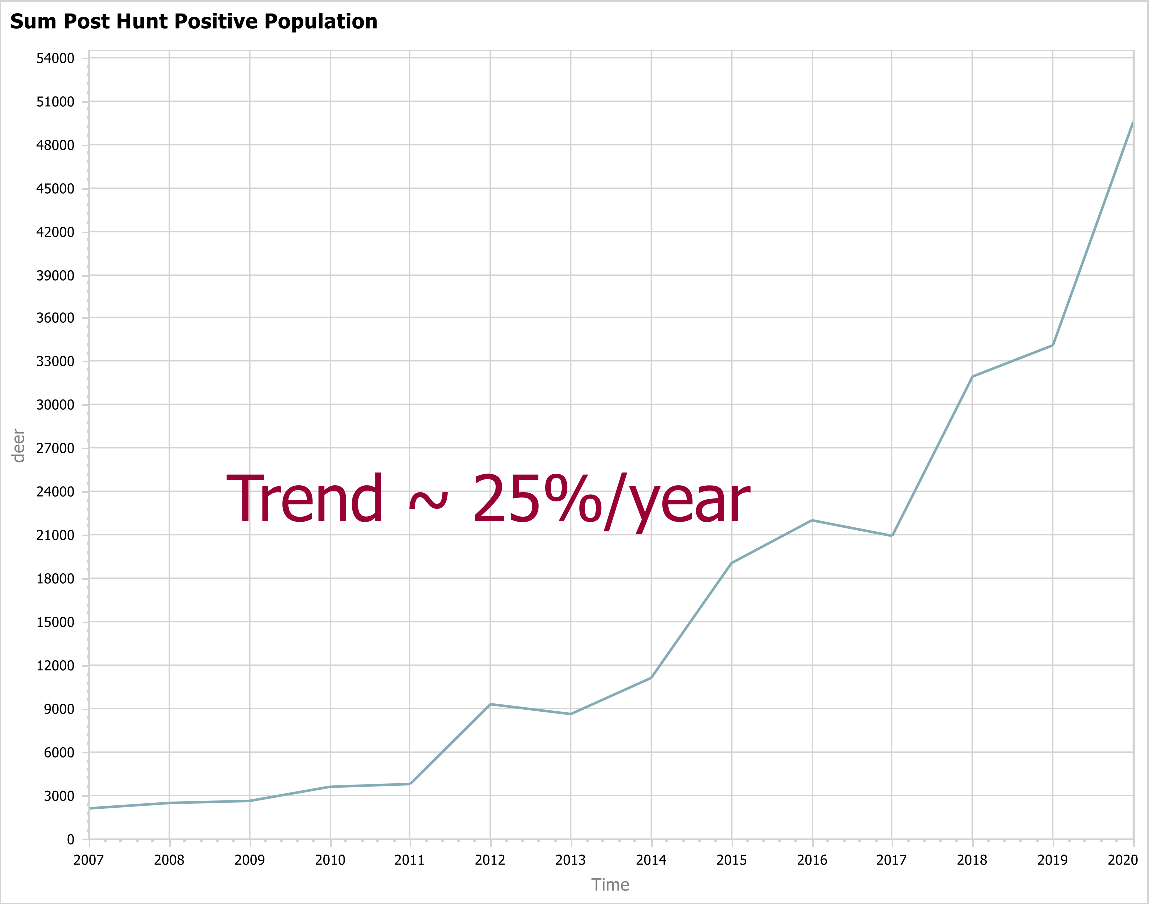

Environmental Prions
Degrading
Depositing
Population
Density
Births
Healthy Deer
Infection Rate
Infected Deer
CWD Mortality
Healthy Harvest
Sightings
Infected Harvest
Carcasses
Decaying
Hunter Effort
Disposing
Incentives
Constraints

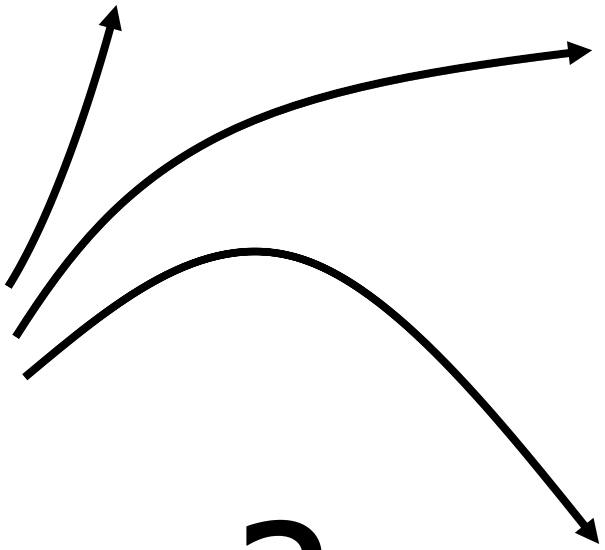
Just a Few Uncertainties...



Estimated Statewide Positive Population

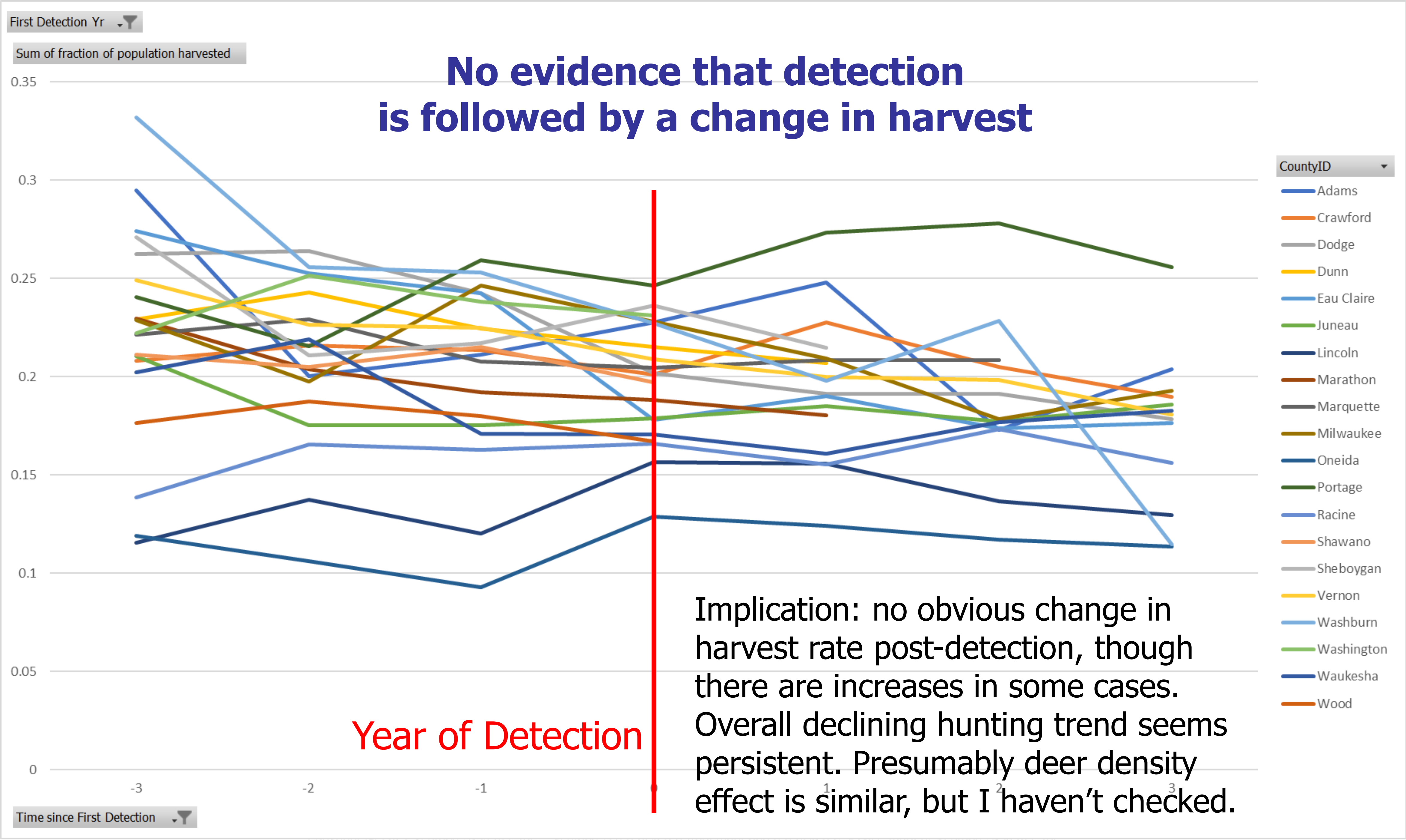
= Fraction Positive x Post Hunt Population, summed over counties





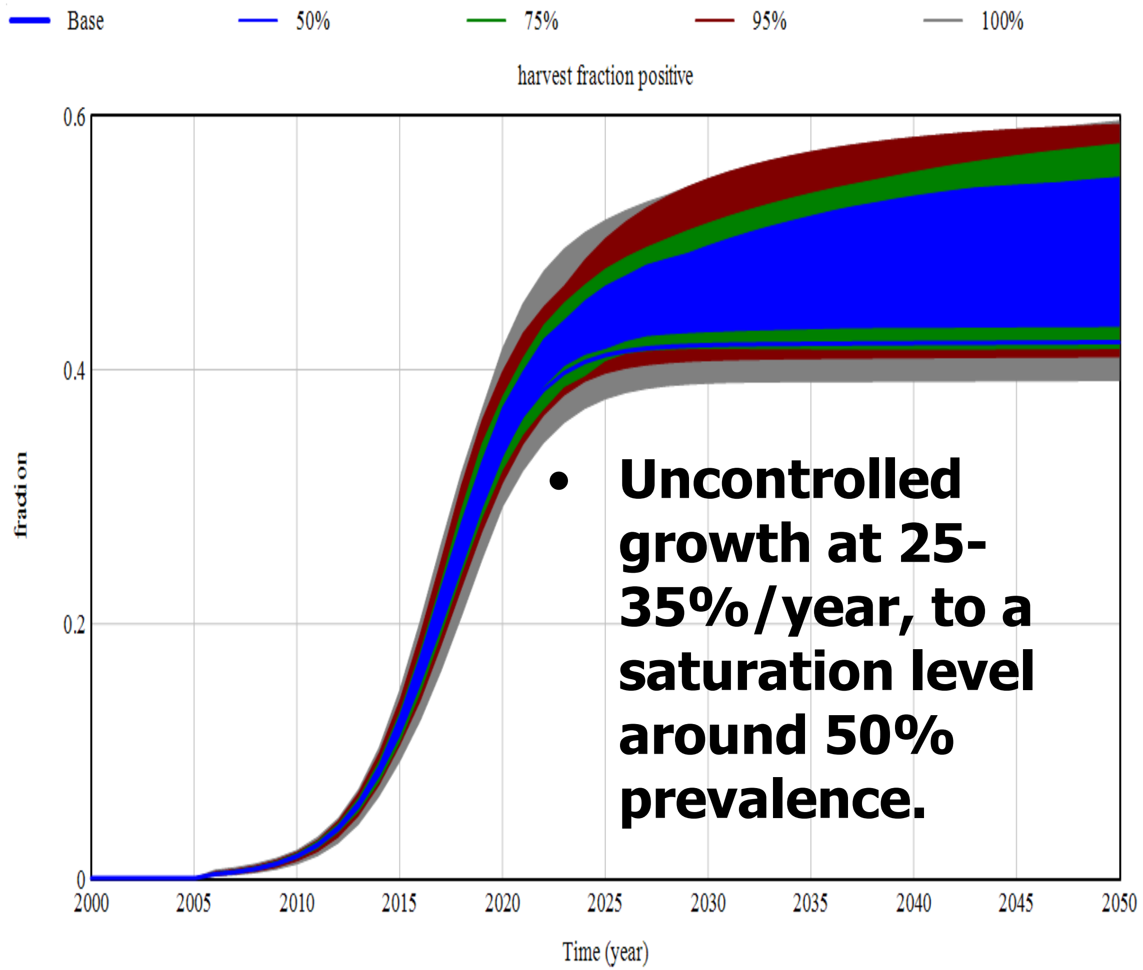
?

No evidence that detection is followed by a change in harvest

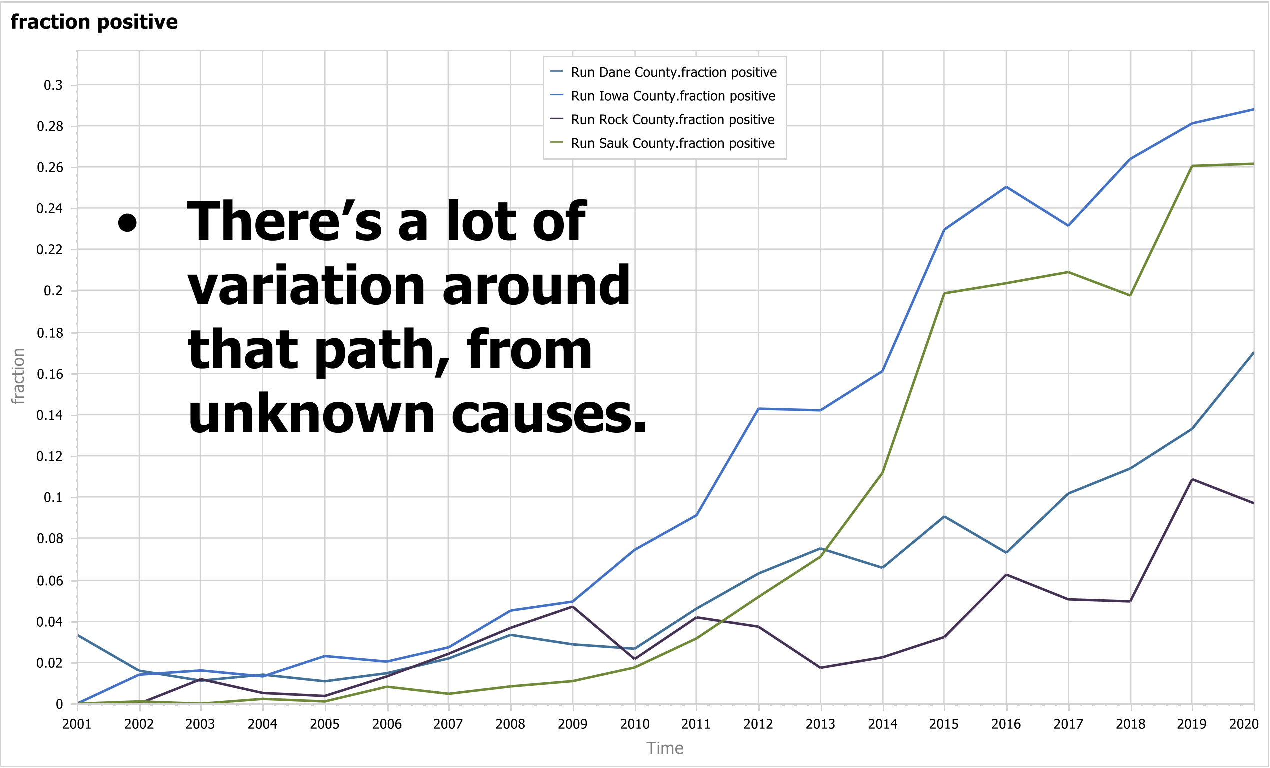


Implication: no obvious change in harvest rate post-detection, though there are increases in some cases. Overall declining hunting trend seems persistent. Presumably deer density effect is similar, but I haven't checked.

Status Quo = Growth
Highest prevalence areas are approaching saturation



- **Uncontrolled growth at 25-35%/year, to a saturation level around 50% prevalence.**



- **There's a lot of variation around that path, from unknown causes.**

SDM Consequence Table: Mean Outcomes

Automated with VenPy + Vensim DLL

	Harvest Action					
Metric	Base	Uniform	Antlerless	Older Bucks	All Bucks	Perfect Targeting
population	871	376	379	875	879	778
older buck population	154	53	86	117	98	118
healthy population	456	269	214	505	554	569
prevalence	0.48	0.29	0.44	0.42	0.37	0.27
harvest fraction positive	0.46	0.28	0.39	0.44	0.38	0.37
positive harvest consumed	74	31	36	76	74	61
clinical prevalence	0.02	0.02	0.02	0.02	0.02	0.02
total harvest	255	185	150	280	314	271
trophy harvest	46	26	26	58	49	47
relative harvest effort	0.96	1.60	1.38	1.09	1.24	1.19
Vegetation Index	1.02	1.09	1.08	1.02	1.02	1.04

Phase 1 – Good Outcomes

- **Complex discussions with stakeholders, organized around the model**
- **Systematic evaluation of policies <u>in combination</u>, not in isolation**
- **More policy experiments than discussion could possibly support**
- **Low level of conflict**

- **Significant "discoveries" for modelers and stakeholders**
  - Difficulty of achieving control with existing policies
  - Reasons for testing to overstate and understate true prevalence
  - Advantages to early intervention
  - Critical importance to follow-up on surveillance

Phase 1 – Not so good?

- **Difficult to assess: possible reticence of stakeholders who didn't buy into the modeling process**

- **Missed opportunities:**
  - Characterizing the historical trajectory – what did policies achieve?
  - Exploring counterfactual histories – where would we be with no control effort?
  - Challenging past mistakes
  - Density-dependent transmission and bathtub dynamics

- **Probably negative:**
  - The baseline, frequency dependent simulation makes control too hard
  - Little characterization of uncertainty of results

2012 Deer Trustee Report

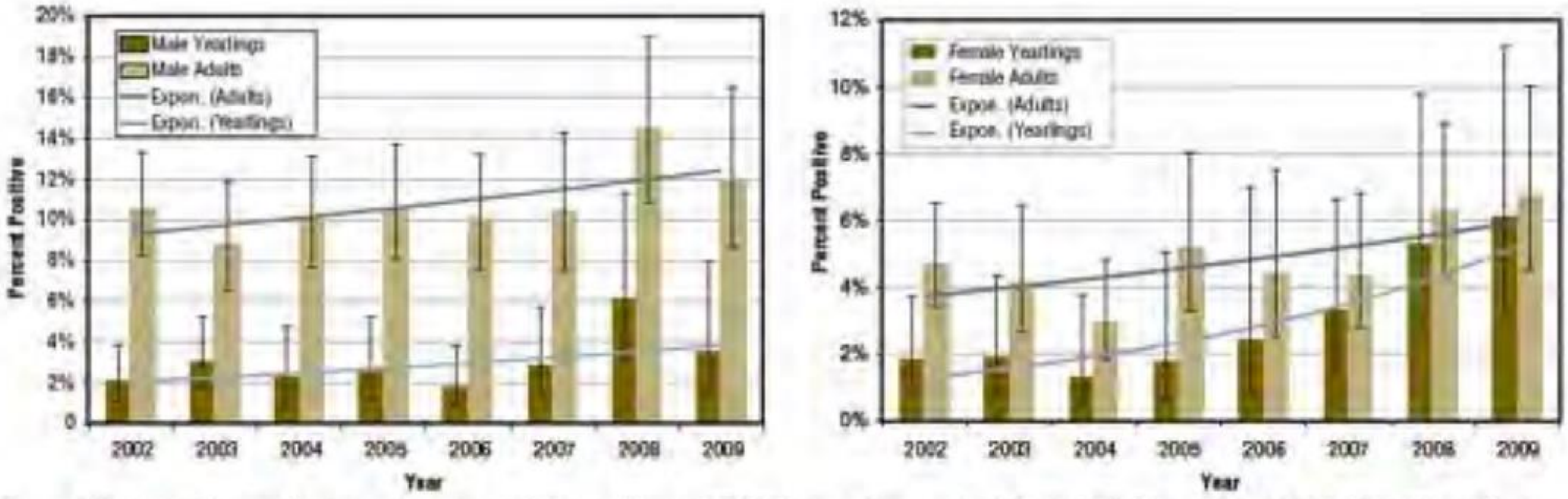


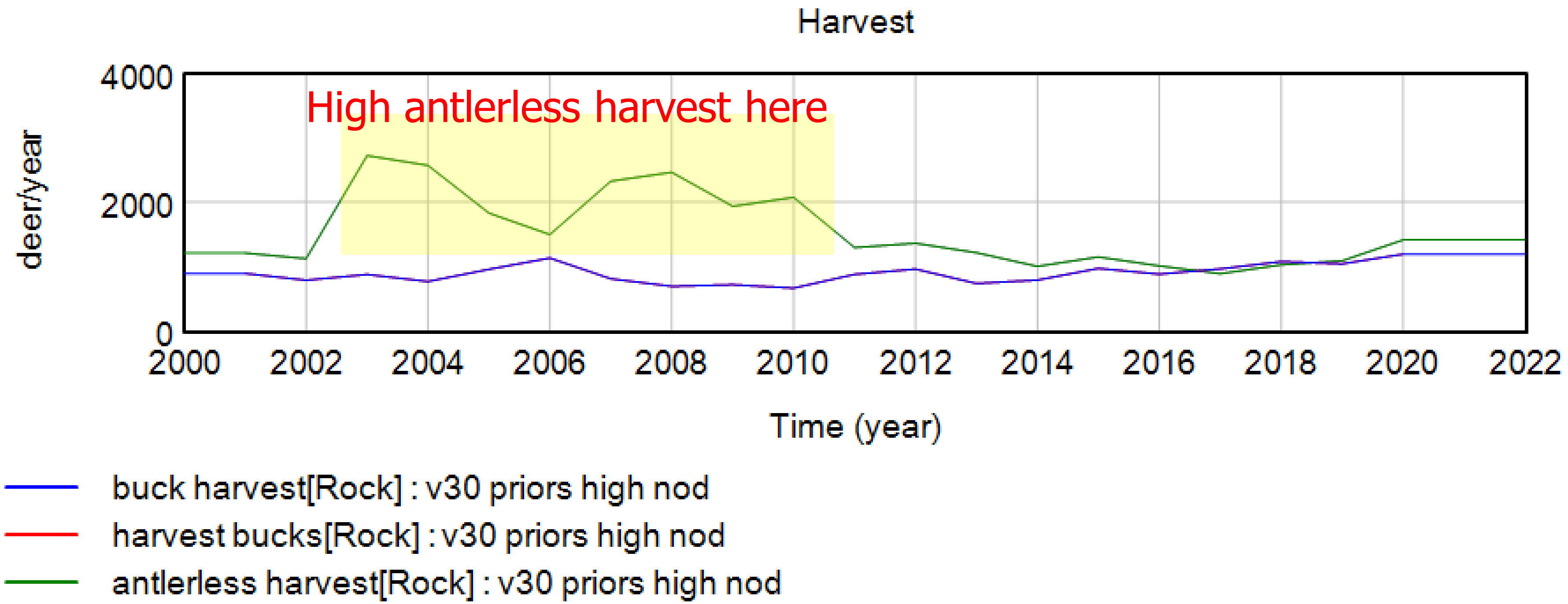
Figure 4. *Estimated prevalence and exponential trend lines of CWD in yearling and adult male (left) and female (right) white-tailed deer from the western core monitoring area, 2002-2009. Vertical lines are 95% confidence intervals.*

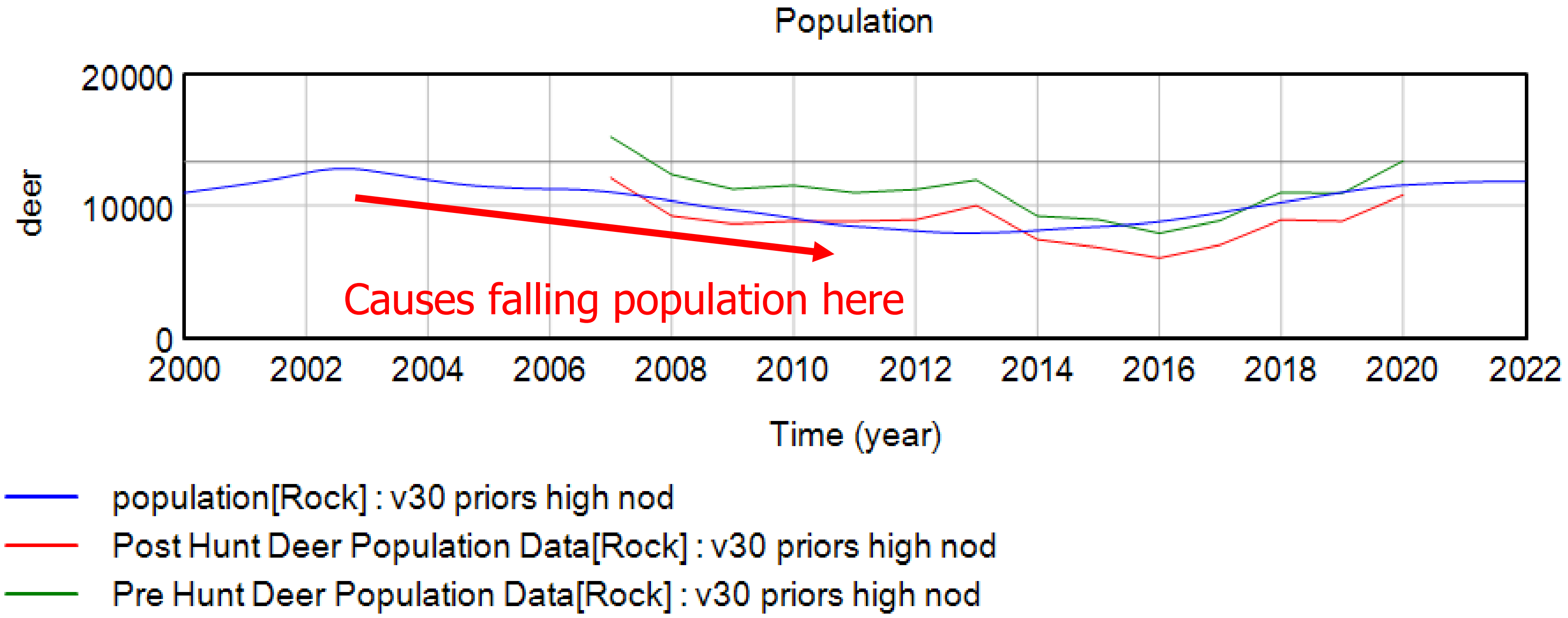
" Figure 14 presents graphs used in the planning document. The graphs imply (using fitted exponential trend lines) an upward trend in infection rates, even for yearlings. Yet, the graphs also present 95% confidence limits for each year; and, in every case these limits overlap. From a statistical standpoint, this means there were no significant differences between years! Wrong!

Also, illogical: if there's no evidence of exponential growth of a disease, that's a win for control!

VENTANA

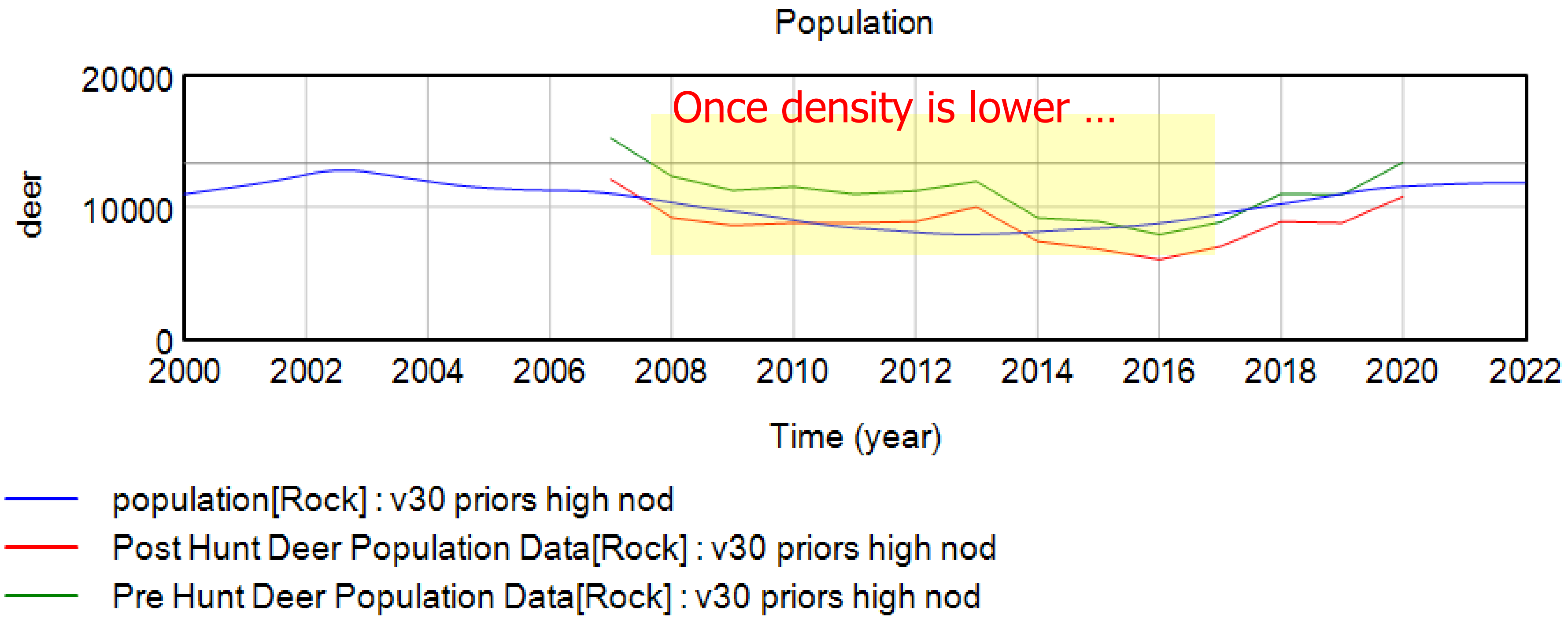
A possible high DD history

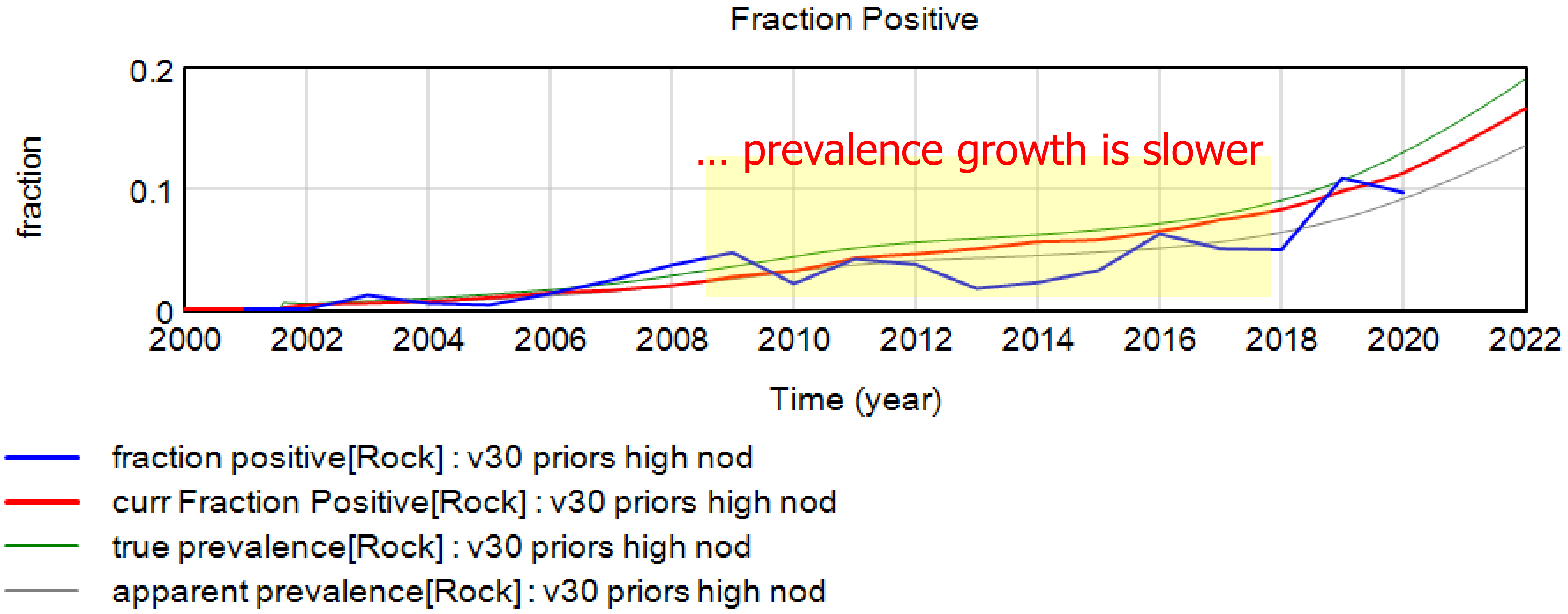




VENTANA

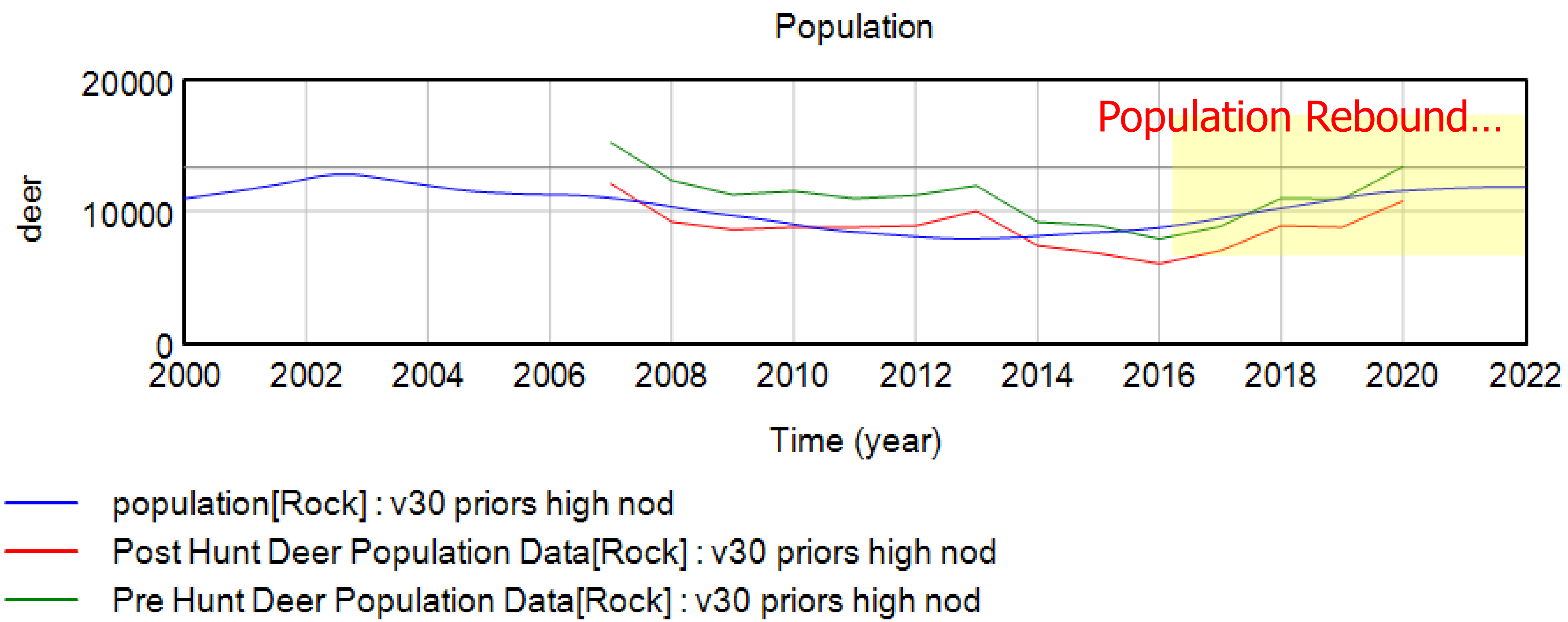
A possible high DD history

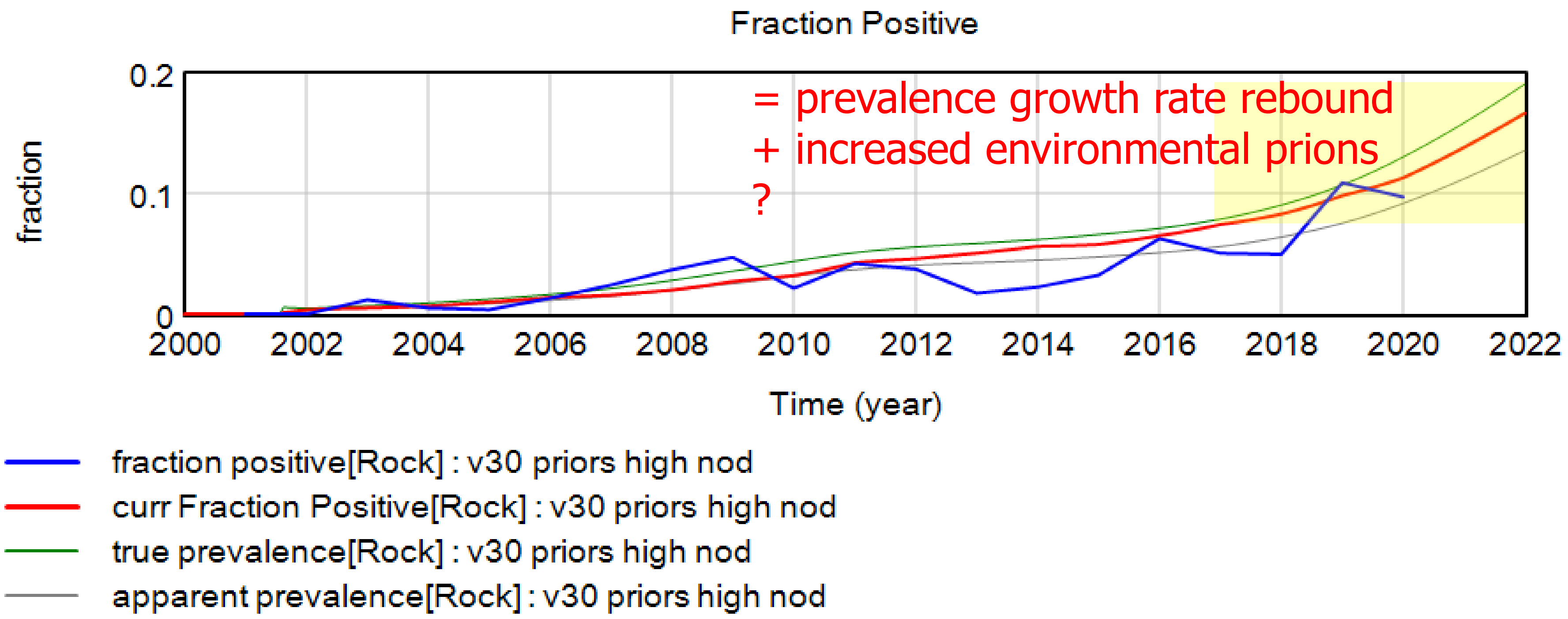




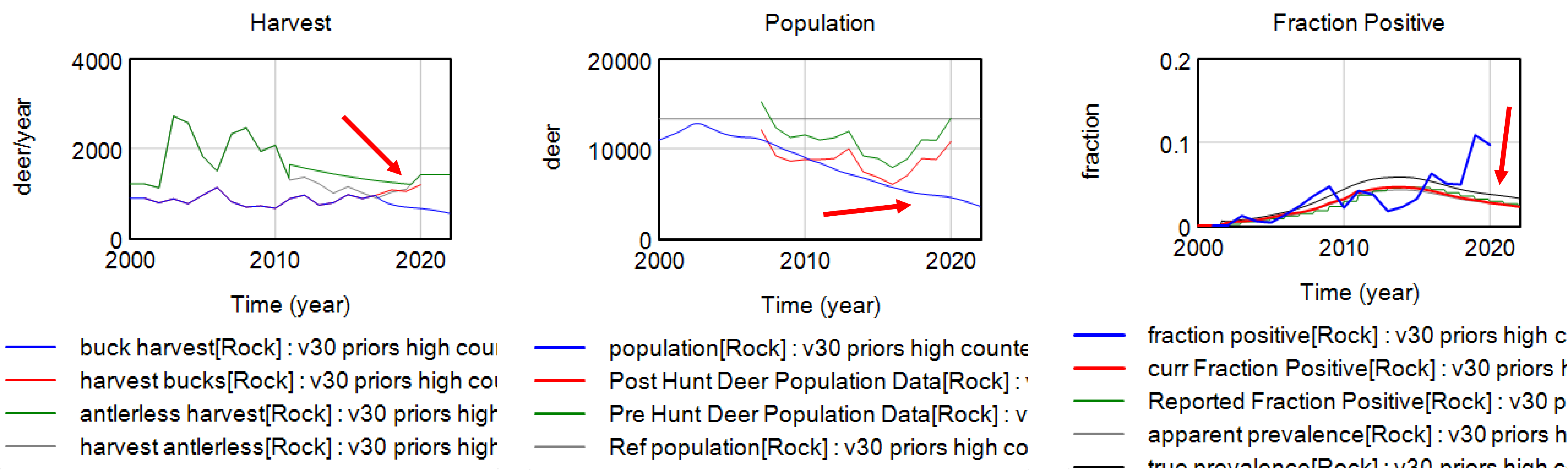
VENTANA

A possible high DD history





What if the high antlerless harvests in 2003-2010 had been sustained?

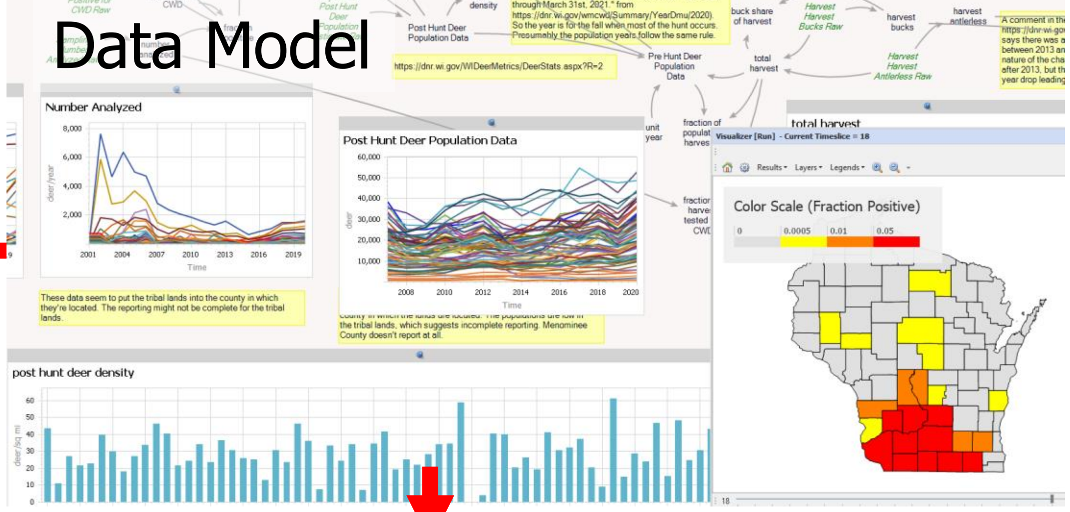


A good decision…

- **Maximizes outcomes of interest…**
- **… from the perspective of multiple stakeholders …**
- **… and is robust to uncertainty about the future and system structure**

- **Not forecasting:**
  - "If we know what will happen, we'll know what to do."
- **Rather:**
  - Anticipate the inevitable
  - Control the controllable
  - Hedge the rest

Process

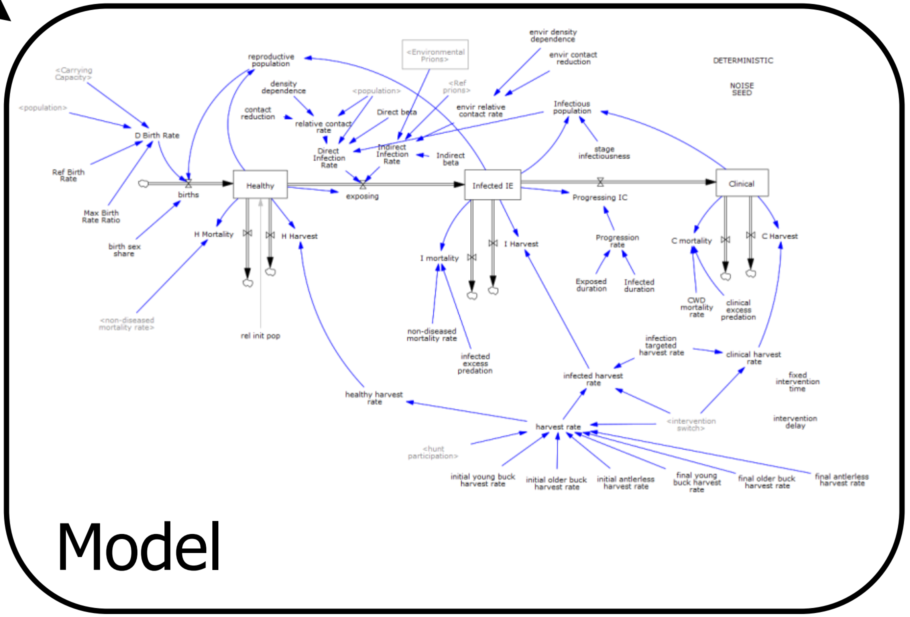


Driving Data

Parameters + Uncertainty
- CWD transmission
- Environmental prions
- Deer fertility & mortality

Decisions
- Hunting
- Baiting & feeding
- Surveillance
- Carcass management
- Safe practices
- Timing

Comparisons



Outcomes + Uncertainty
- Prevalence
- Fraction positive
- Population
- Age, sex structure
- Harvest
- Surveillance results
- Hunt effort
- Human exposure

Naïve Calibration

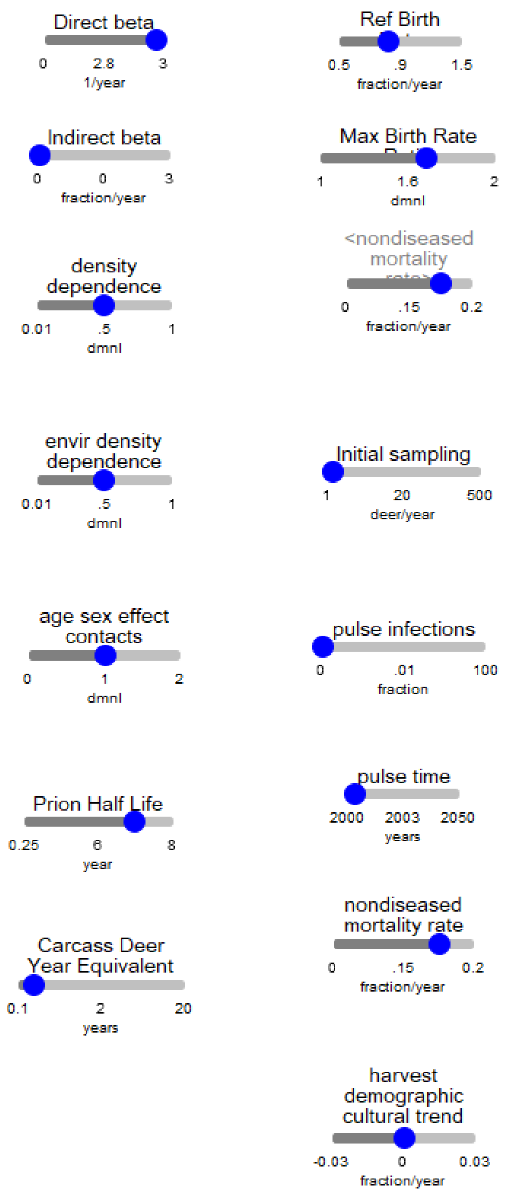
- **Build a control panel that compares model and data series**
- **Do some hand calibration to discover what parameters matter, and how much they can plausibly vary**

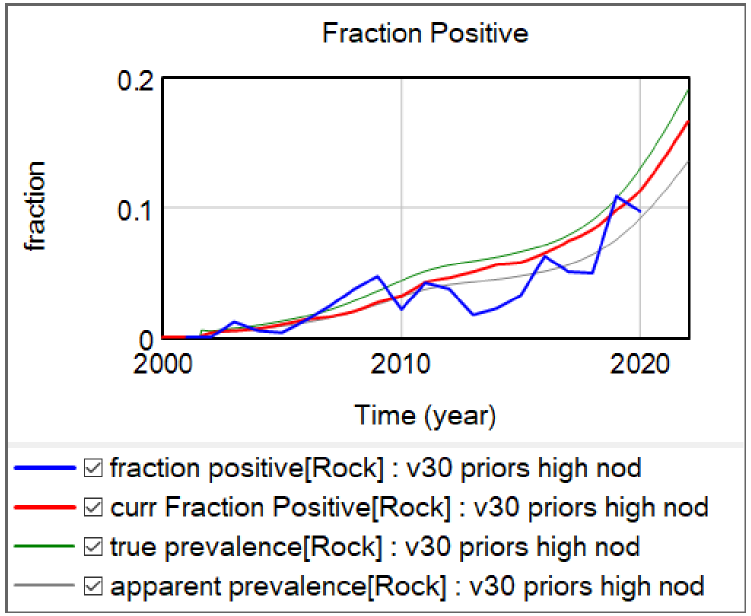
- **This may be a big share of the value!**

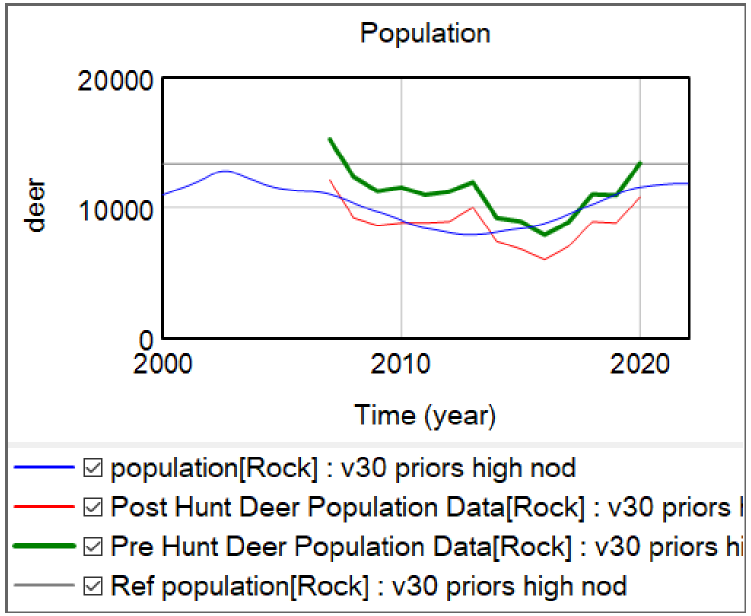
- **Then:**
  - Automate using the optimizer
  - Initially, don't worry too much about the details
  - Essence:
    - Create a payoff or objective function that characterizes the goodness of fit
    - Use some algorithm to iterate over a list of parameters to maximize the payoff

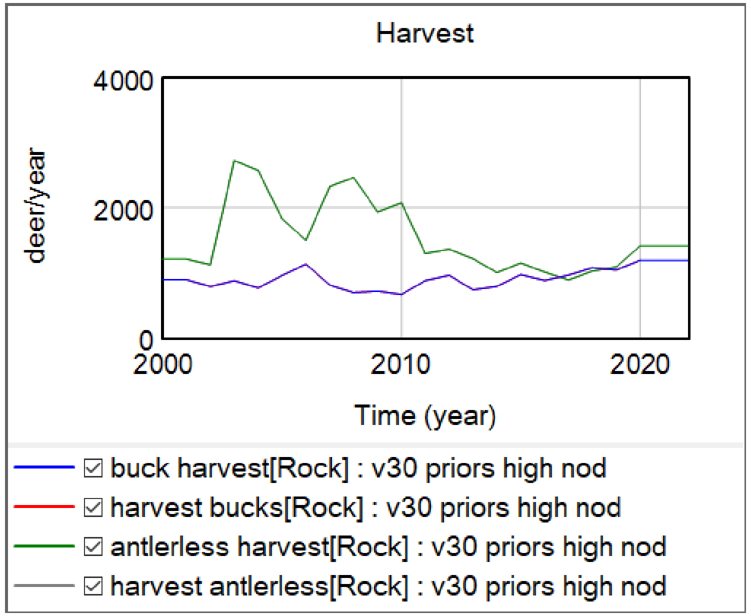
Hand Calibration Infrastructure

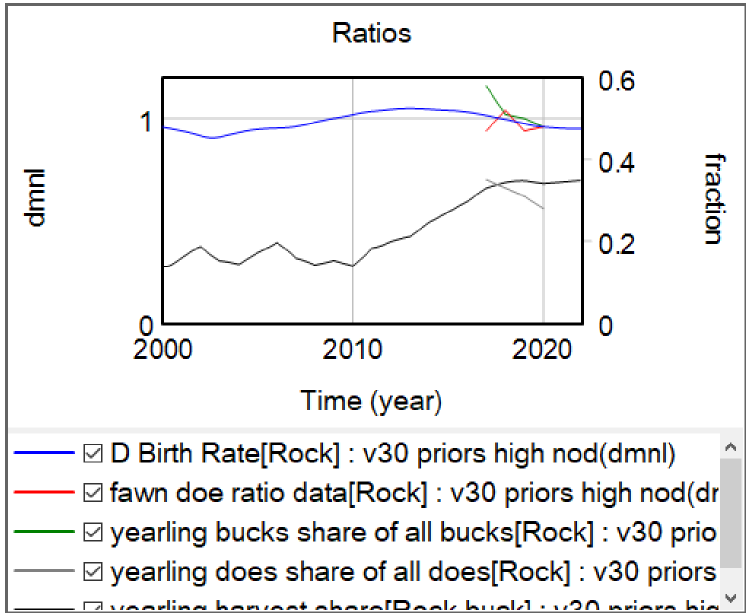
Assumptions





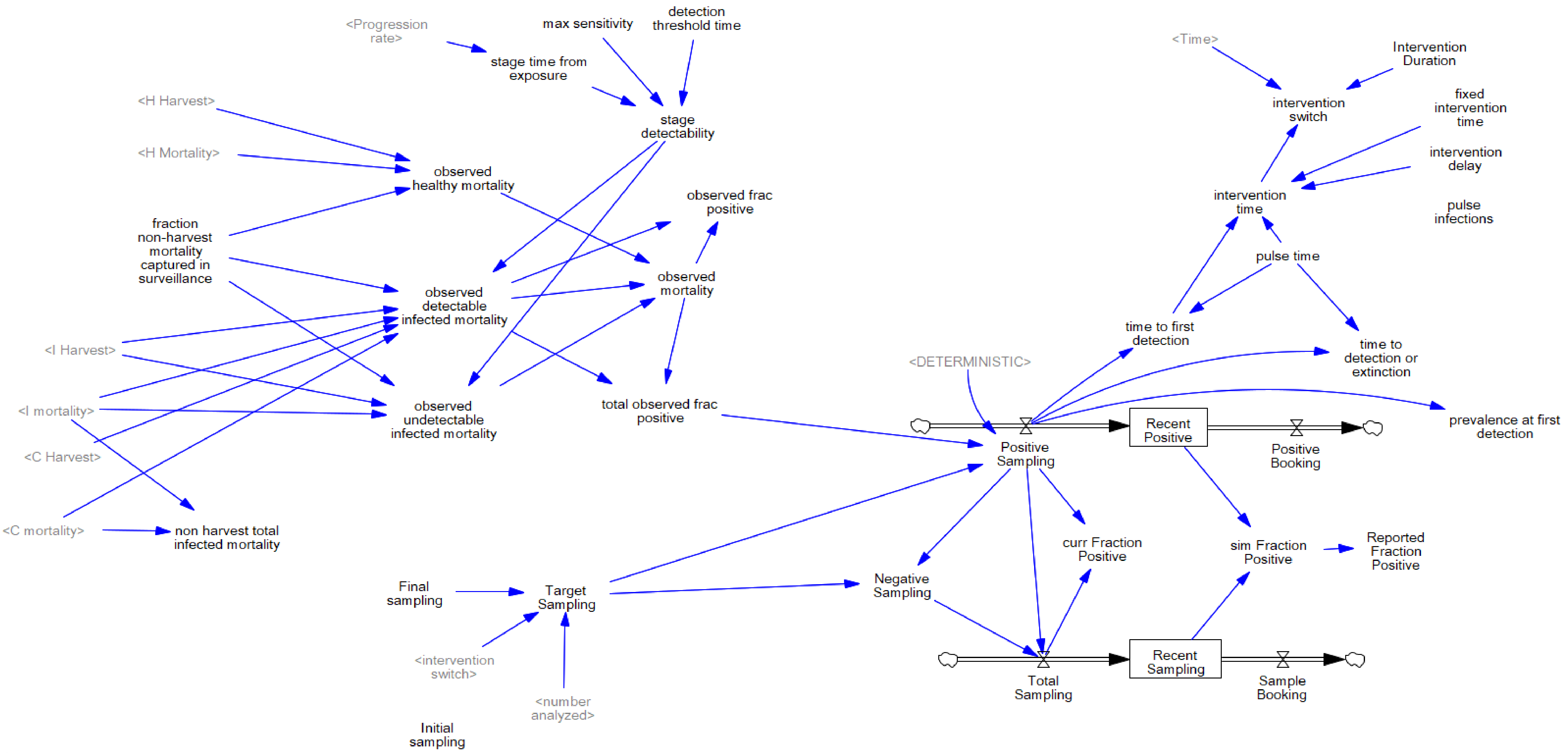






VENTANA

Modeling the Reporting Process



VENTANA

Formal Calibration
Fundamentally, what are we doing?

- **Create a model of the process that generated the data**
  - Dynamic structure of the underlying system
  - Distribution of errors, lags and other features of the measurement process
  - Distribution of disturbances to the system state
  - Priors for unknown parameters or informally characterized behaviors

- **Assuming the model is right, what parameters are most likely to have generated the data, and (maybe) are most consistent with our priors?**

- **Output**
  - Frequentist: if I keep repeating this experiment, the parameter will be in my confidence interval 95% of the time
  - Bayesian: starting from my prior knowledge, after seeing the data I'm 95% certain the parameter lies within my uncertainty interval

You're already a Bayesian

- **SD uses lots of a priori information**
  - Model structure
  - Reference modes
  - Dynamic hypotheses
  - CLDs
  - Parameters sourced from SMEs, literature, other models

- **You probably use Bayesian updates**
  - Adaptive expectations or smoothing
  - Kalman filtering

- **If you have lots of data, the answer is probably the same!**

Bayesian System Dynamics

- **Bayes Rule: P(A|B) = P(B|A)*P(A)/P(B)**
- **Conjunction Rule: P(A and B) = P(A)*P(B), assuming A, B independent**

P(Params|Data) = P(Data|Params) * P(Params) / P(Data)

Posterior — The Answer

Likelihood — "Traditional" Calibration

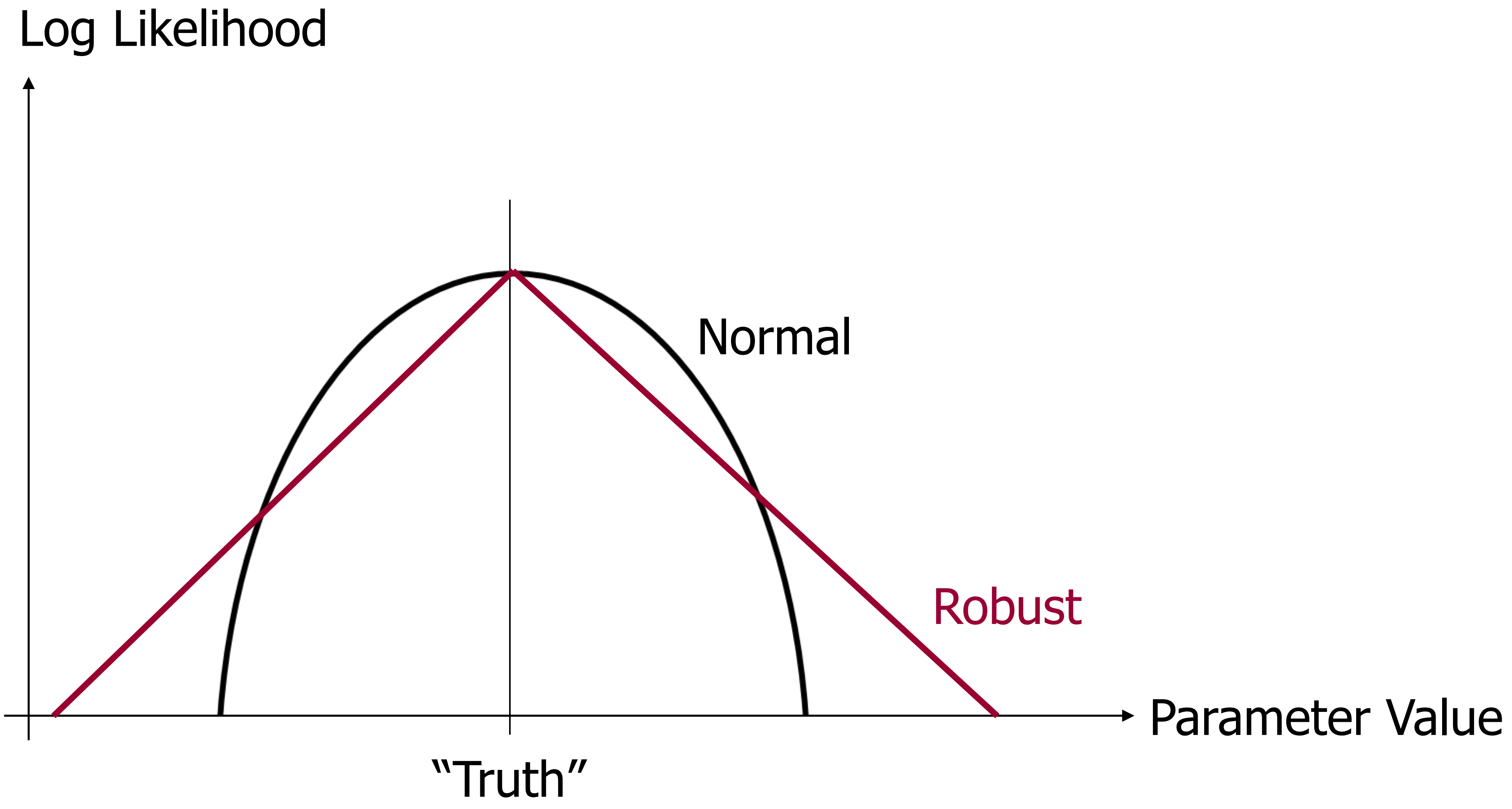
Prior — New term, expressing beliefs, which you might regard as data points from other scales or domains

Ignore — Purely a function of the data, not the parameters, so it's a constant scaling factor

Log-Likelihood
Gaussian errors

- **Likelihood =** $\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{model-data}{\sigma}\right)^2 / 2}$
  - This is the PDF of the Gaussian (Normal) distribution
  - σ represents the scale of the error associated with a data point
  - σ might vary with time, or with the scale of the data
  - You can estimate σ
- **Maximizing Log(Likelihood) is the same as maximizing Likelihood, but more convenient because Log() turns multiplication into addition**
- **Log(Likelihood) =**
  – LN(σ) - the bigger the σ, the lower the likelihood, as it's spread thinner
  – LN($\sqrt{2\pi}$) - this is a constant we can ignore
  – $\left(\frac{model-data}{\sigma}\right)^2 / 2$ - the weighted sum of squares, adjusted by the divisor /2

What does the likelihood look like?



Priors

- **A prior expresses our belief about a parameter before we see the data**
- **No priors = uniform priors**
  - It's not always a good choice, *but* if you have lots of data, it might not matter.
- **Non-informative or Maximum Entropy priors**
  - Contribute as little information as possible, i.e. assume maximum ignorance a priori
  - For a scale parameter like a time constant, this may be simple, like -LN(param) for positive parameters
  - This can be tricky to construct, and you're rarely *completely* ignorant about a parameter – after all, you thought to put it in the model to begin with
- **Informative priors**
  - If you – or experts or literature – have some opinion about a parameter, you can use a subjective probability distribution to characterize that
  - This can also be tricky if multivariate correlations are important, but we can avoid that to some extent by choosing orthogonal parameters
- **You can also use priors to characterize informal knowledge about expected behavior of the model**

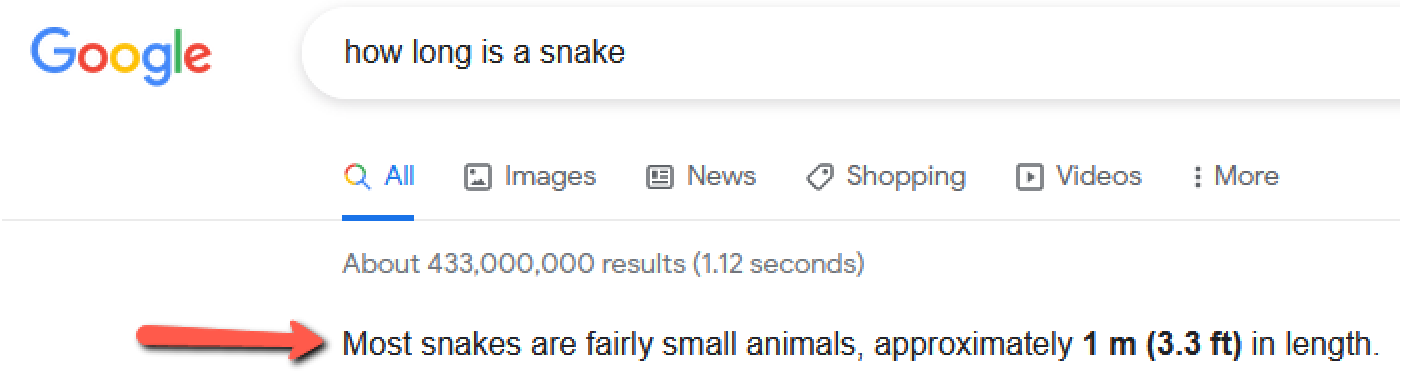
VENTANA

Expressing Priors

- **A prior is a lot like a data point!**

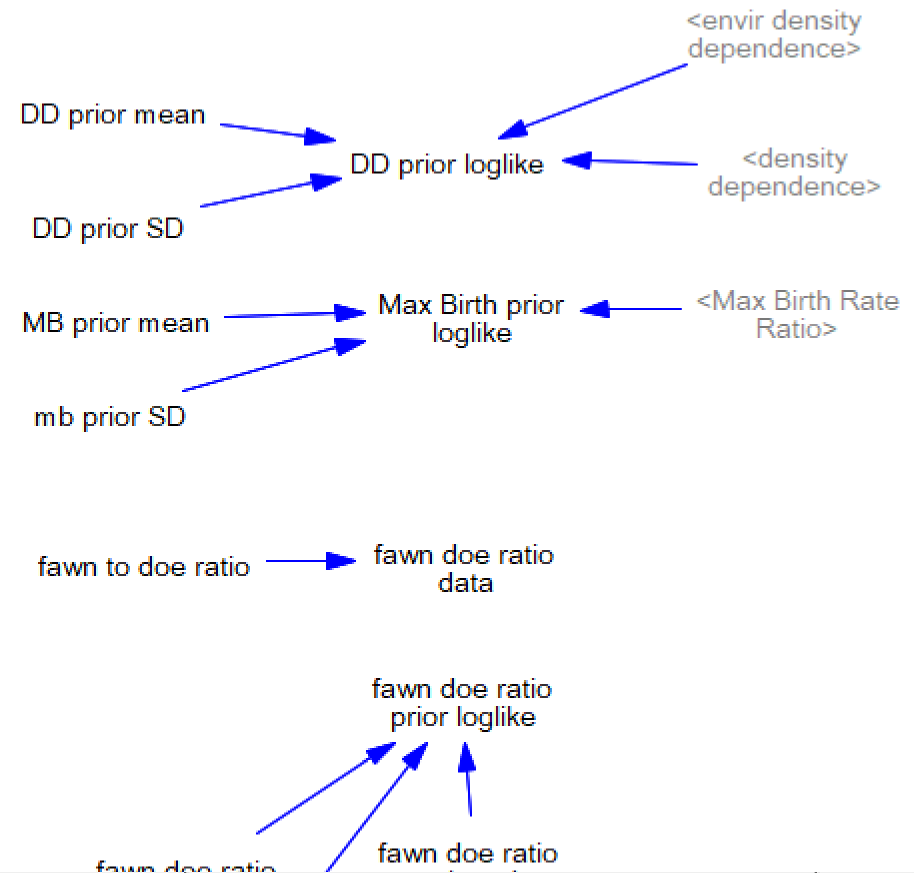
- **If our belief is Normal (Gaussian):**

- **Likelihood =** $\frac{1}{\sigma\sqrt{2\pi}} e^{-\boxed{\left(\frac{param-\mu}{\sigma}\right)^2 / 2}}$
  - For an MCMC log likelihood ratio, we only need the last term
  - μ represents our belief about the mean value of the parameter, i.e. best guess
  - σ represents our belief about the plausible variation; high uncertainty = large σ



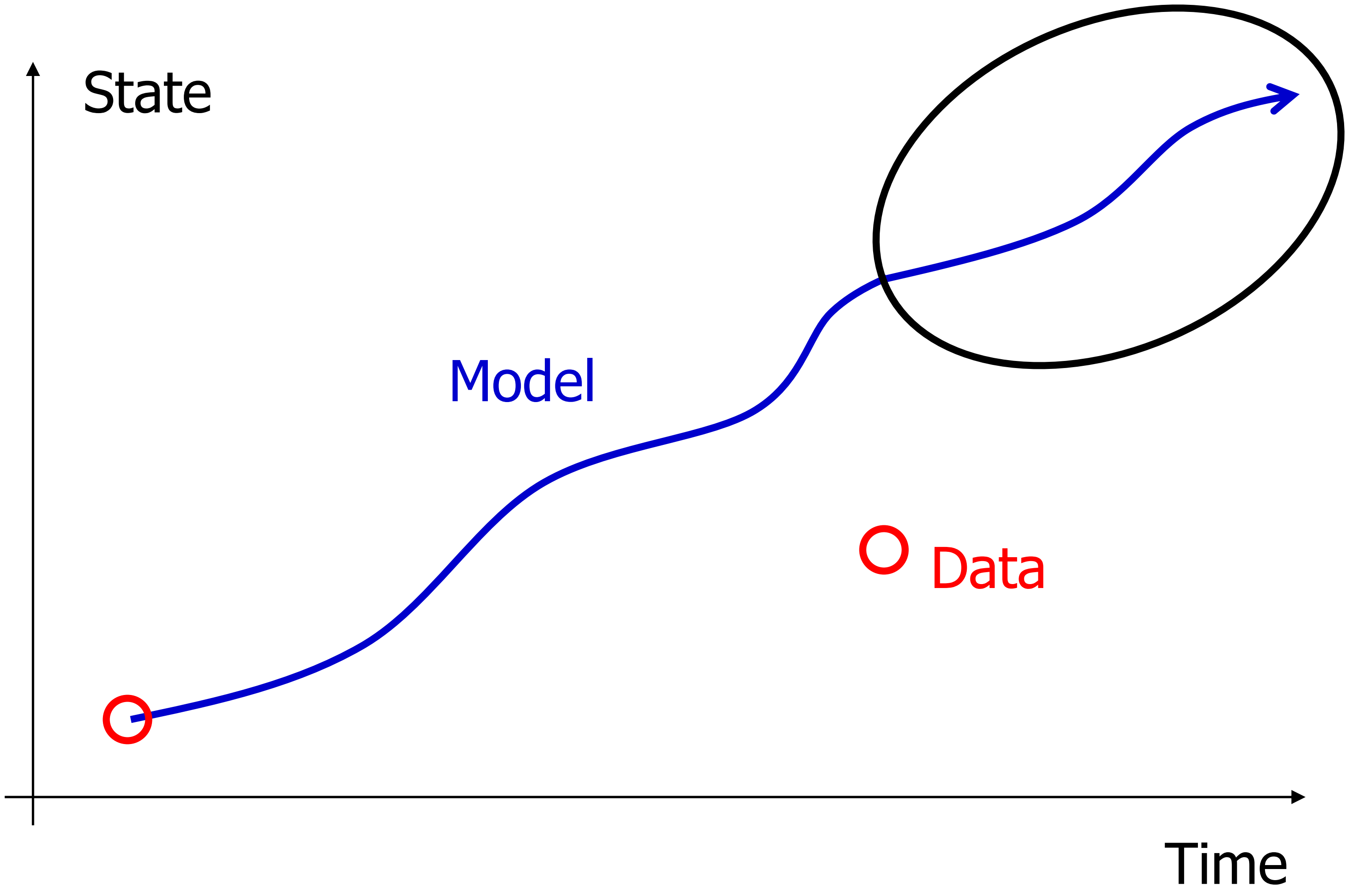
Example

- **Our CWD model has a view dedicated to priors**
- **Most are simple (geometric mean & standard deviation)**
- **A few operate on composites, like generation time (combination of several delays)**
- **A few express hierarchy: how much variety of behavior is plausible across counties?**
- **The prior likelihoods are additional terms in the payoff**



Policy	Distribution	Timing	Transformation	Variable name	Compare To	Weight
Calibration	Gaussian		Log	antlerless harvest[County]	harvest antlerless[County]	0.1
Policy		Always	None	fawn doe ratio prior loglike[County]		1
Calibration	Normal		Log	yearling harvest share[County,buck]	yearling bucks share of all bucks[County]	0.1
Calibration	Normal		Log	yearling harvest share[County,doe]	yearling does share of all does[County]	0.1
Policy		Initial	None	DD prior loglike[County]		1
Policy		Initial	None	Max Birth prior loglike[County]		1

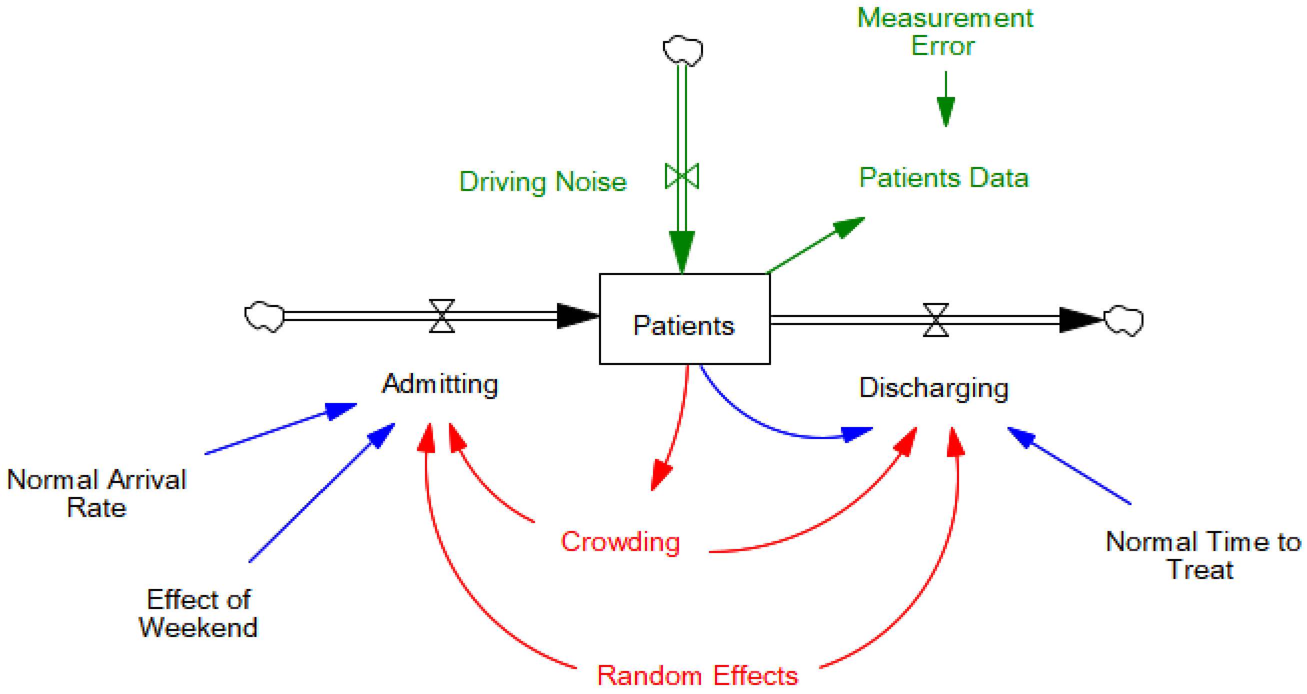
If the model is in the wrong part of the state space, its response may be wrong



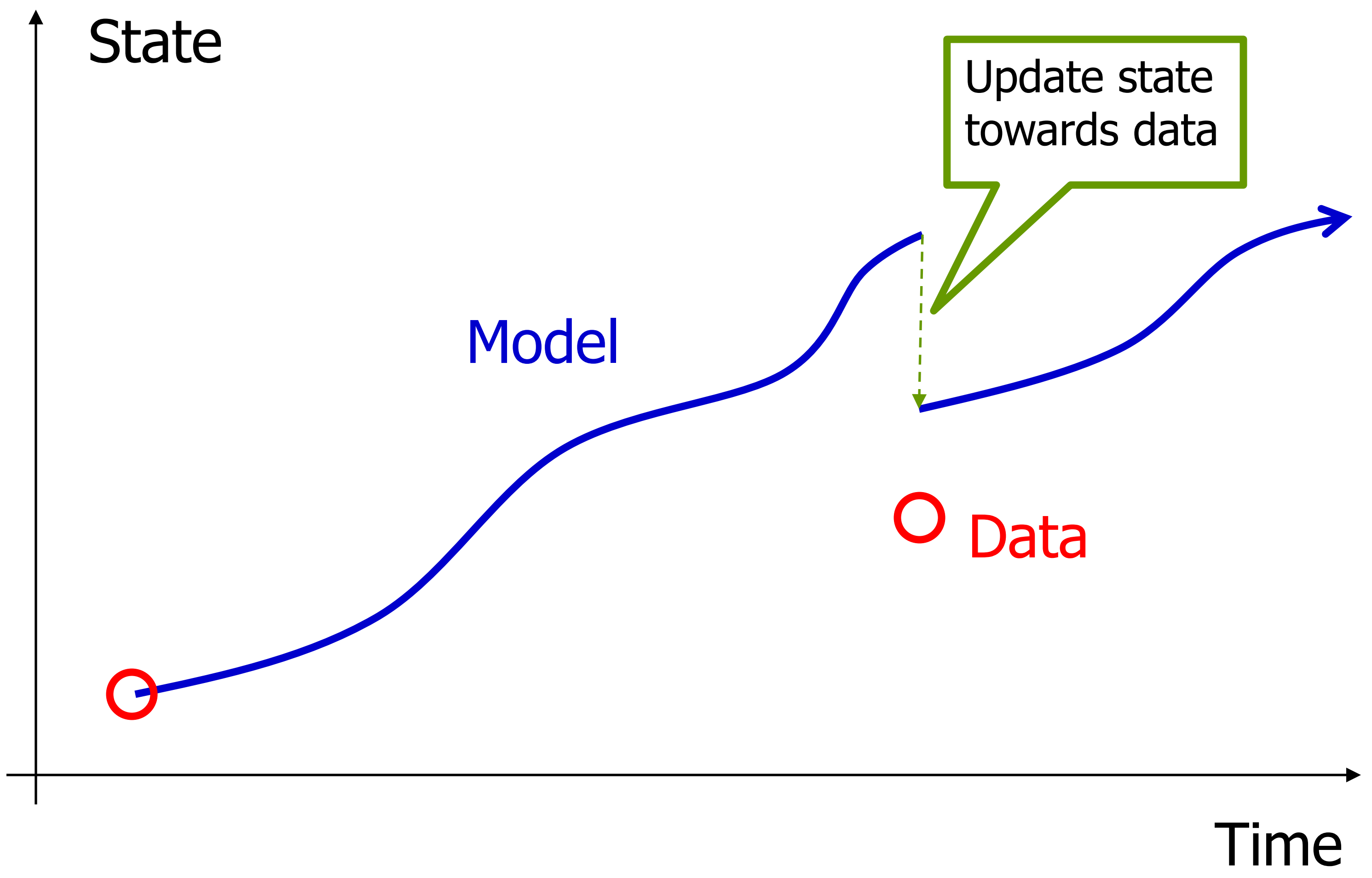
If you're estimating a parameter that affects this part of the trajectory, the response may be particularly wrong

Kalman filtering

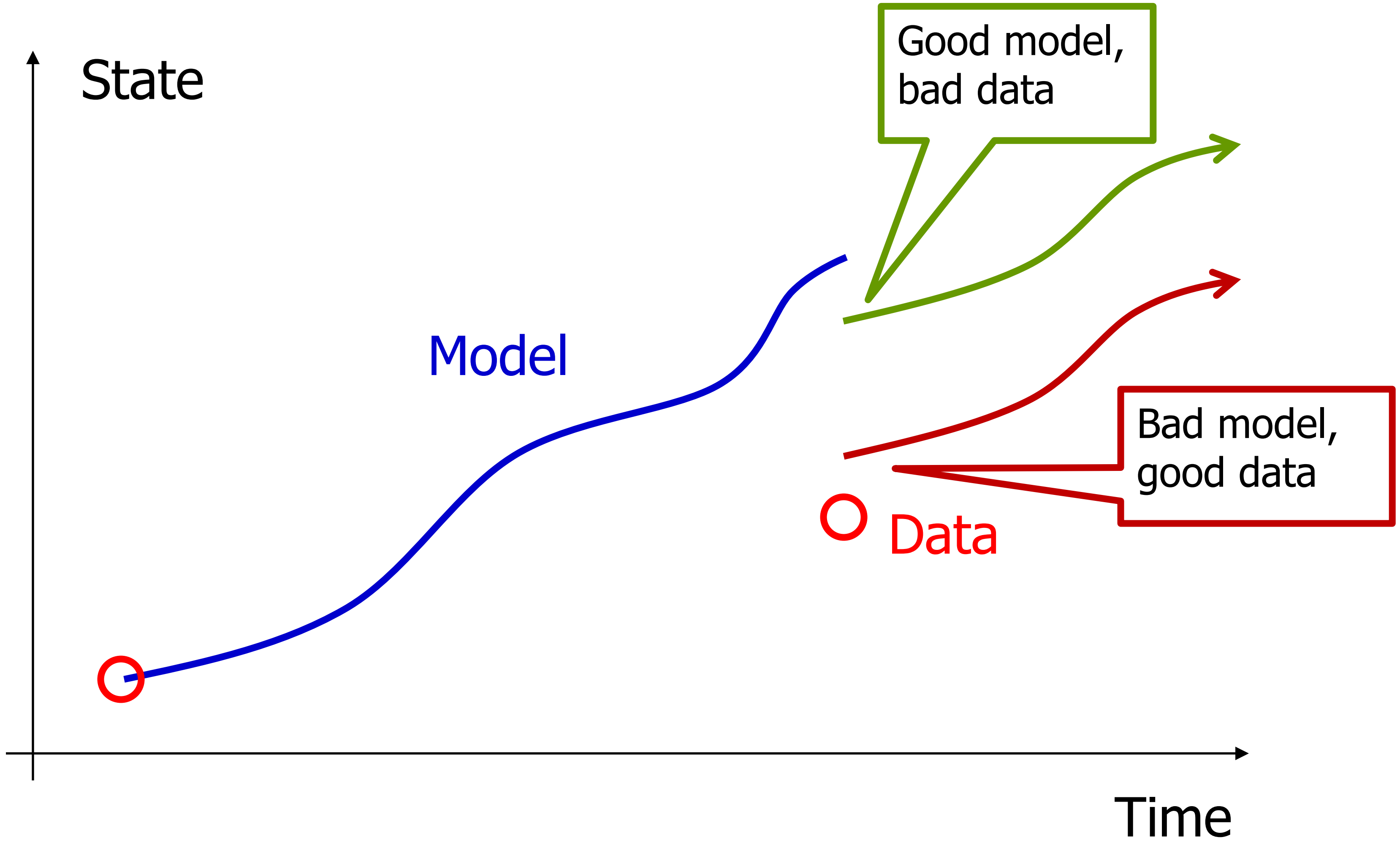
- **Motivation**
  - The model may not reflect everything that affects the system state
    - Random noise from events (e.g., Poisson arrivals)
    - Structure we don't know about
  - Over time, the model state drifts away from reality
- **The Kalman filter is a special case of Particle Filtering with Gaussian noise**



The Kalman filter updates the model trajectory towards the data



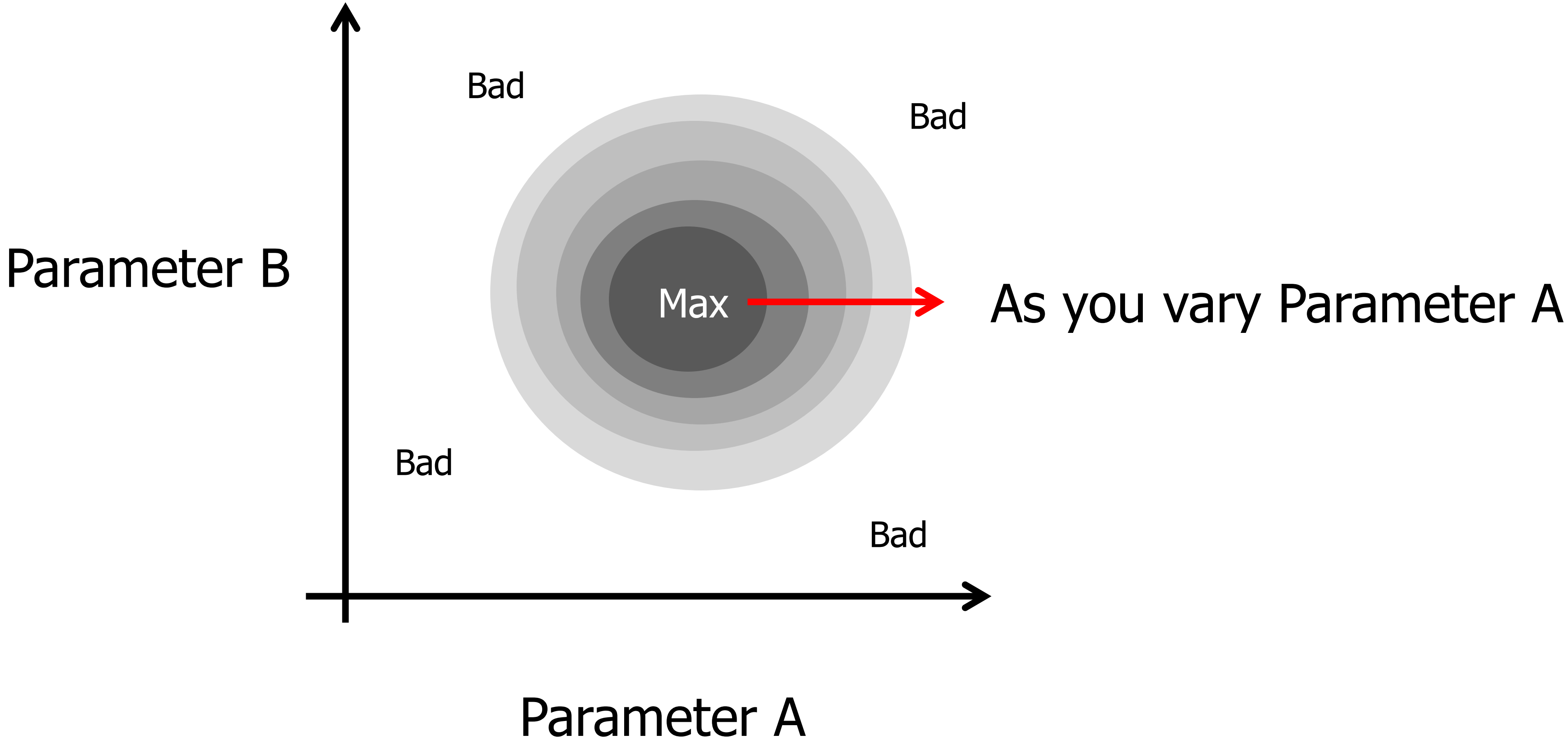
How far? Depends on the estimated quality of the model and data



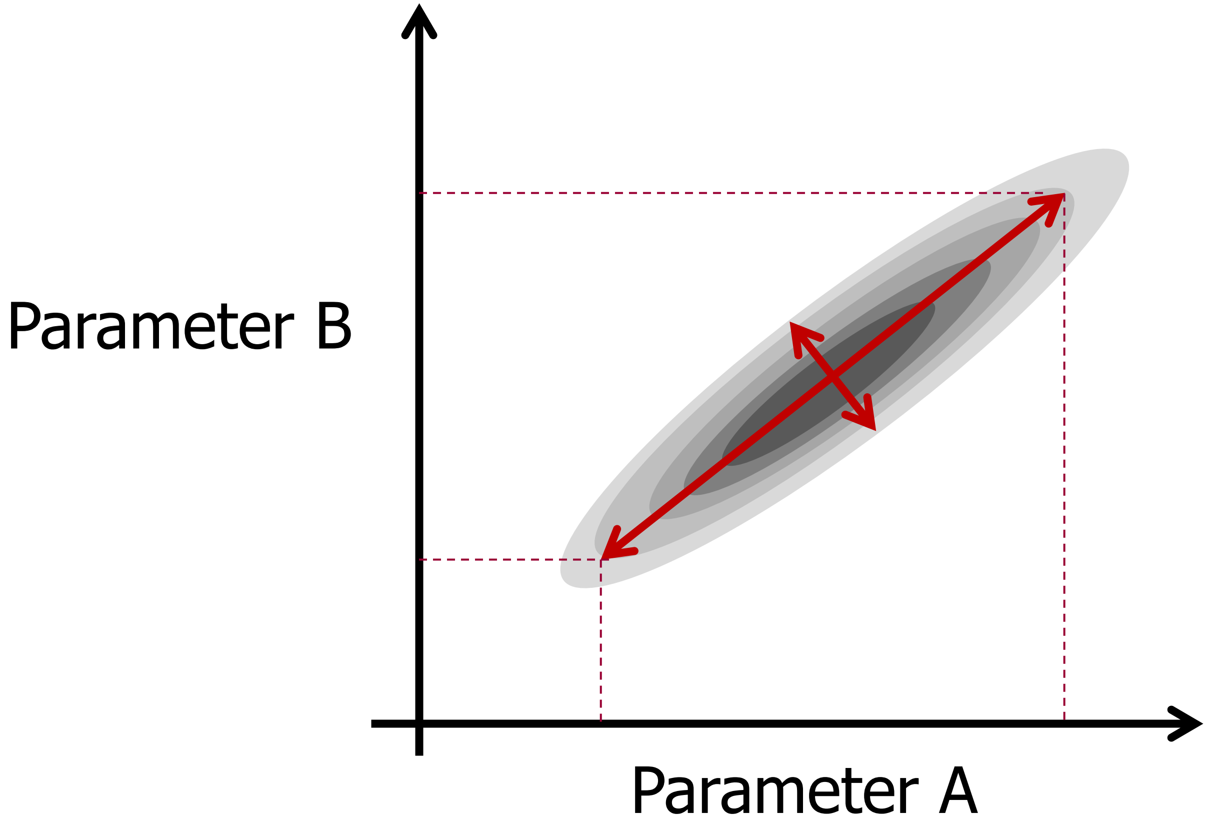
Confidence Bounds & Uncertainty Intervals

- **Motivation**
  - Statistical
    - Is an effect significantly different from zero?
    - After seeing the data, what do we believe about a parameter?
  - Practical
    - What does uncertainty imply for policy?
    - What data might narrow the bounds?
- **Computation**
  - Old way
    - Optimize to find the best fit to data
    - Explore the payoff surface around the maximum
  - New ways
    - Bootstrapping (draw samples from the data)
    - Markov Chain Monte Carlo (MCMC)

Multidimensional Likelihood



Traditional Method: Measure the Ellipse



- **This may be hard if the likelihood surface is shaped like a banana, or a hedgehog, or a bag of 10-dimensional jellybeans…**
- **One-dimensional confidence bounds omit information about the joint distribution of parameters**

VENTANA

Using MCMC to Reveal the Posterior

Posterior

P(Params|Data) = P(Data|Params) * P(Params) / P(Data)

Likelihood **Prior** **Ignore**

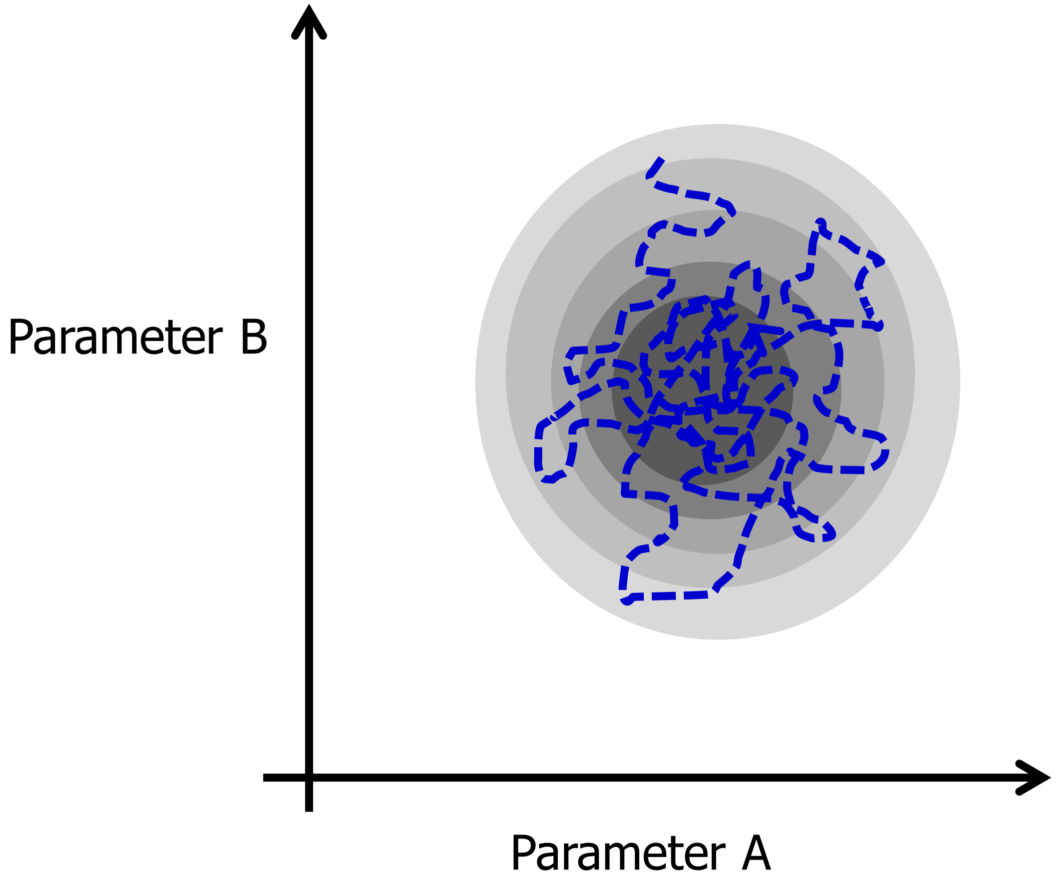
We'd like to know this distribution.

Without this term, the posterior isn't a properly-scaled probability distribution, i.e. it doesn't integrate to 1. So, we need a way to sample the posterior that only cares about the *relative* likelihoods of the data & priors.

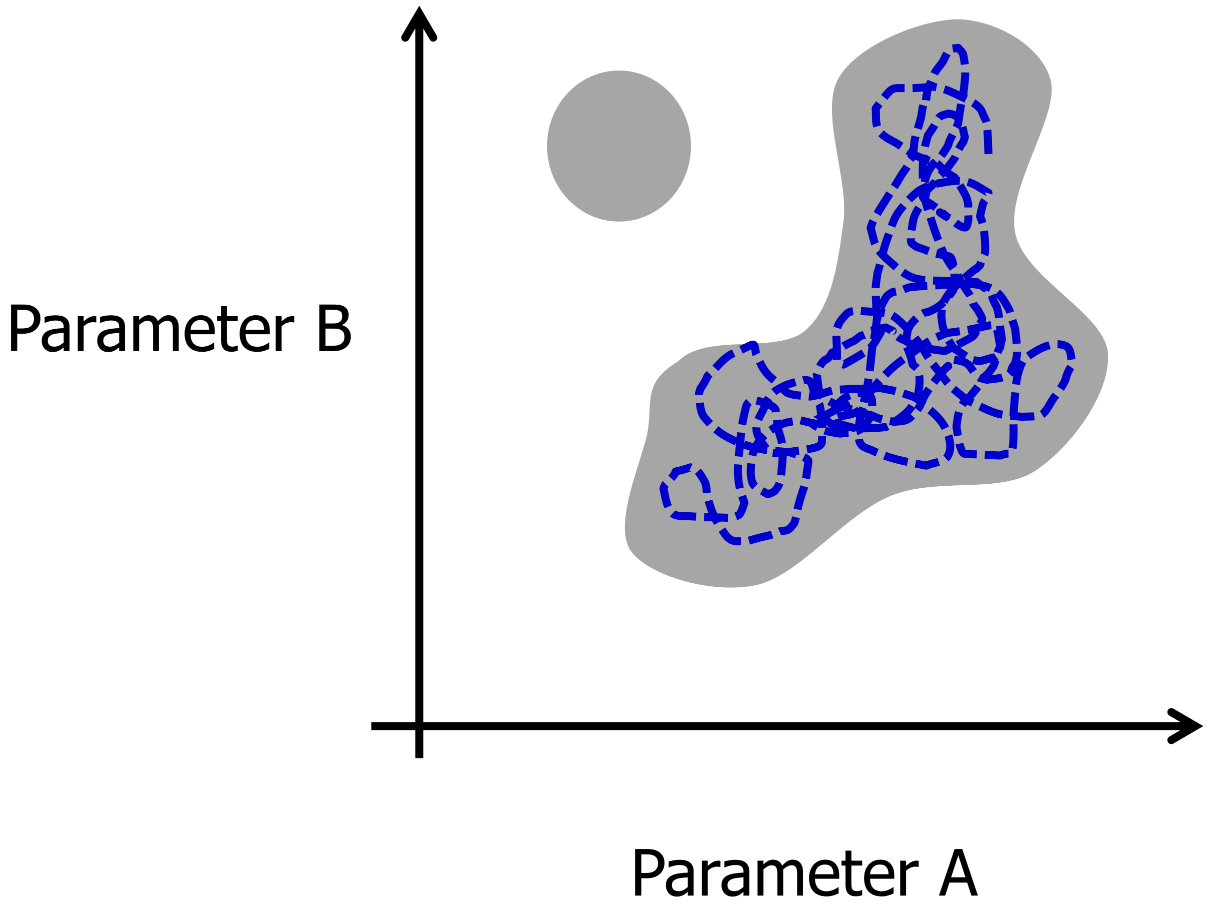
VENTANA

MCMC

- **Basic idea: unleash a random walk on the likelihood surface**
- **Probability of accepting a proposed step is proportional to likelihood**
- **Density of the resulting path converges to the underlying likelihood**

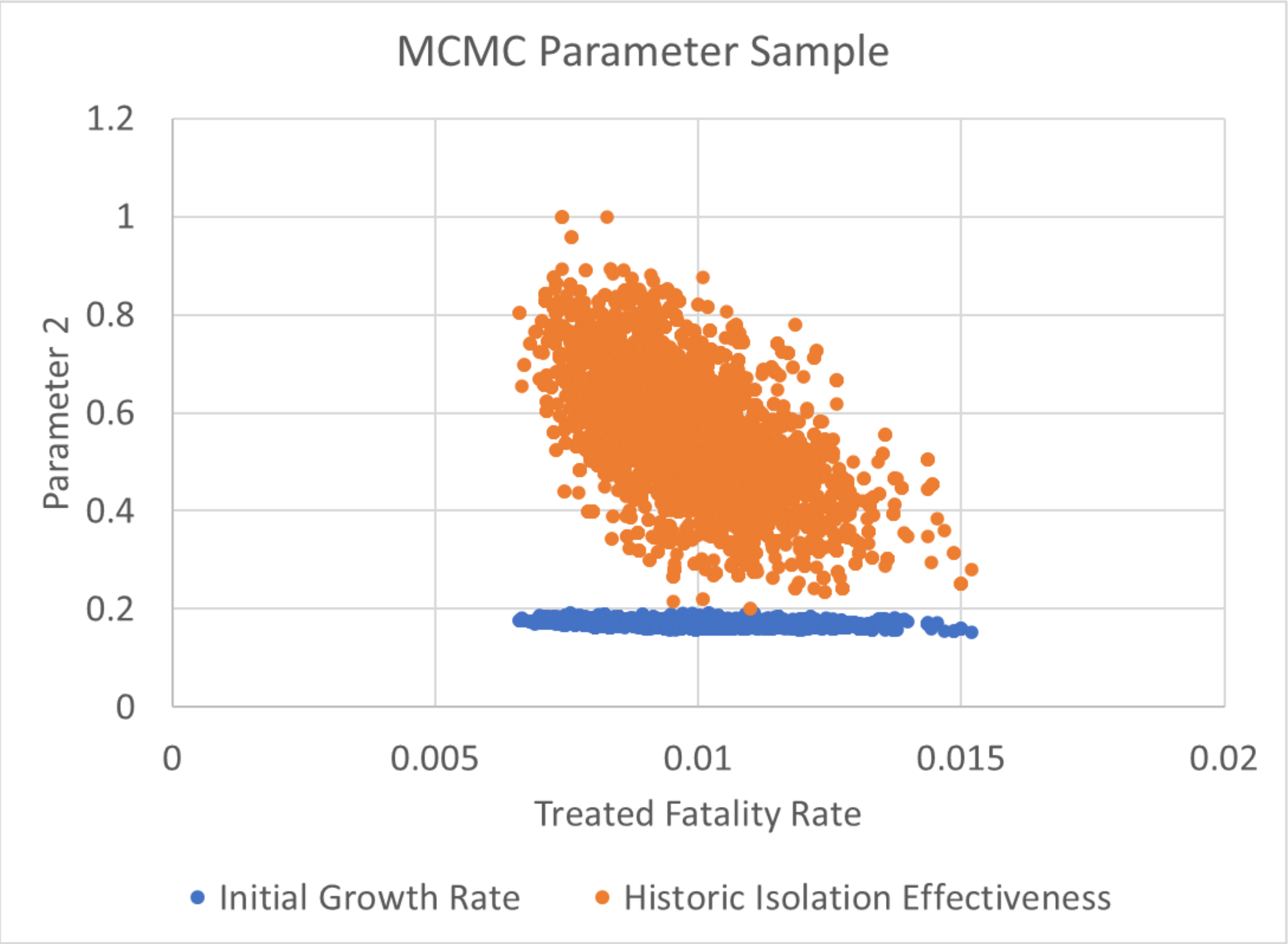


MCMC



MCMC Output

- **A sample of points describing the joint distribution of parameters**
- **Diagnostics**
- **You can then use this to generate sensitivity runs reflecting the sampled parameters**

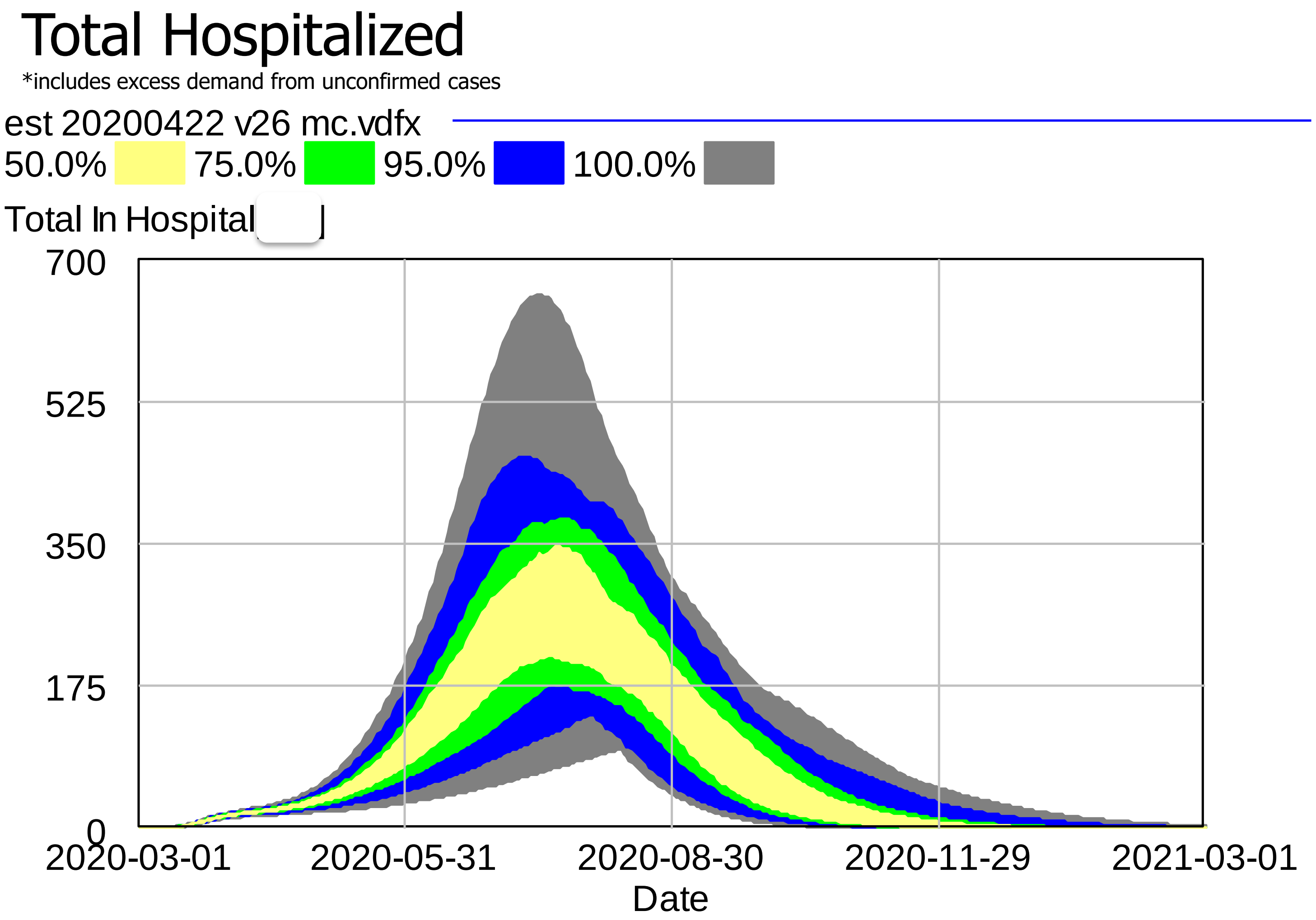


est 2020 05 23 v37e_MCMC_sample+scenario

Using the MCMC Sample for Sensitivity Runs

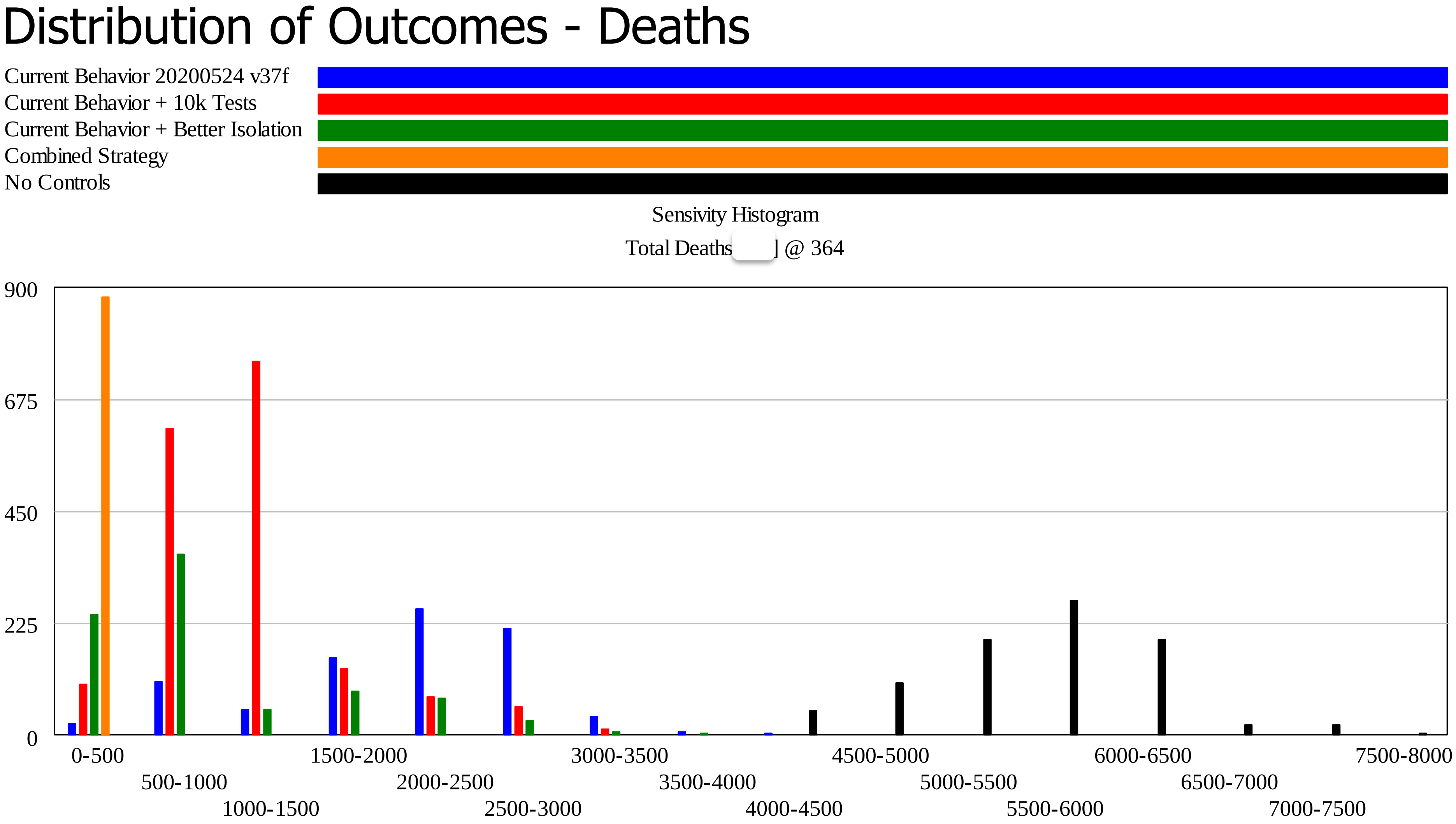
- **What does the parameter distribution imply for the distribution of behavior?**

- **Does the data lie within the uncertainty interval?**



What policy performs best under combined uncertainties?

- **Simple: every run is a sensitivity run**
- **Fancy: stochastic policy optimization**



VENTANA

Synthetic data

- **Purpose:**
  - Test your procedures end to end
  - Can you get useful parameter estimates from limited data?
  - How important are sources of noise or other features of the data?
  - What if the model structure is a simplified version of reality?
- **Procedure**
  - Interpret your model as the truth
  - Change some parameters to make things harder
  - Add noise to some model outputs and/or states
    - Measured cases = RANDOM FUNCTION(true cases)
    - Patients = INTEG(admitting – discharging + NOISE, initial patients)
  - Truncate the frequency and duration of the measurements
  - Use the synthetic data to see if you can recover parameters and distributions
  - More fun if you have an adversary!

My Typical Playbook

- Build/refine structure

- Load data, create a data model
- Create an interface view with model-data comparisons
- Do some hand calibration to see what parameters are interesting

- Do a quick & dirty calibration
  - Weight payoff with log transform and wild guesses at fractional errors
- Evaluate fit, work with model more, ponder what is really problematic or uncertain

- Develop a more carefully weighted payoff, consider Kalman filtering, priors
- Do MCMC to generate a confidence sample
- Do sensitivity runs based on the sample

- Design policies
- Test policies deterministically

- Do policy experiments with sensitivity

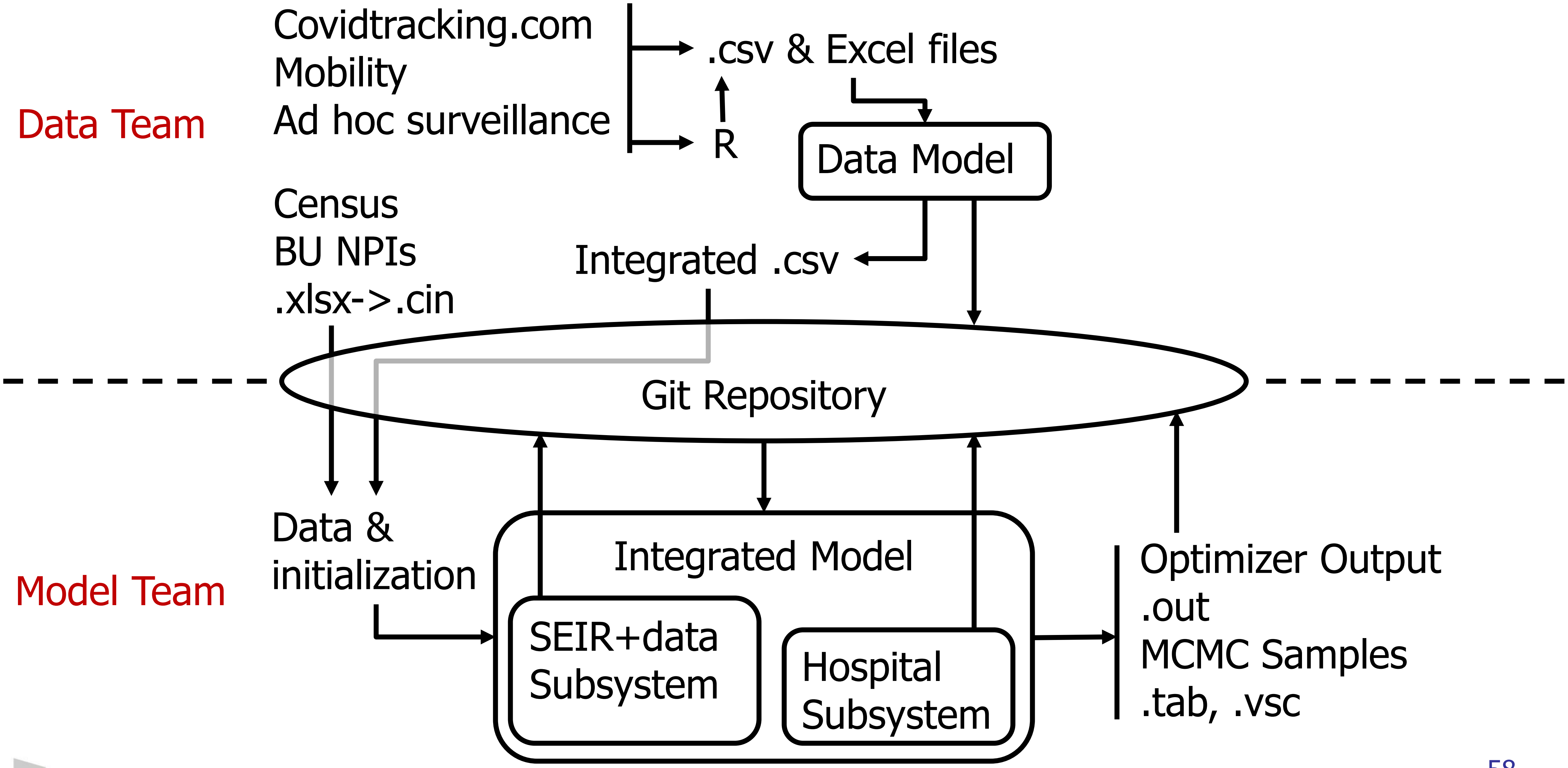
Lots of Iteration!

VENTANA

Some Useful Habits

- Dynamics first – no point devoting statistical horsepower to a dumb model
- Eyeballs first – learn about the model response before you automate fitting
- Consider the data, and data model, an investment
- Use source control and scripting to make things replicable
- Model the data reporting system
- Seek balance between dynamics, statistics and process
- Bring the clients along
- Guide data collection with decision needs
- Remember that this is analysis – it's just a tool, not a religion
- Always be done! – Jim Hines

Contrast: State COVID Project Architecture



Bottom Line

Jay was wrong!

- **Fitting to data doesn't make the model worse**
- **It's hard to make a sensible model fit arbitrary data**
- **If you can't reproduce history, you have some explaining to do**
- **Data is an important information source (not sufficient but necessary)**
- **Intuitive characterizations of system behavior or decision rules may be wrong**

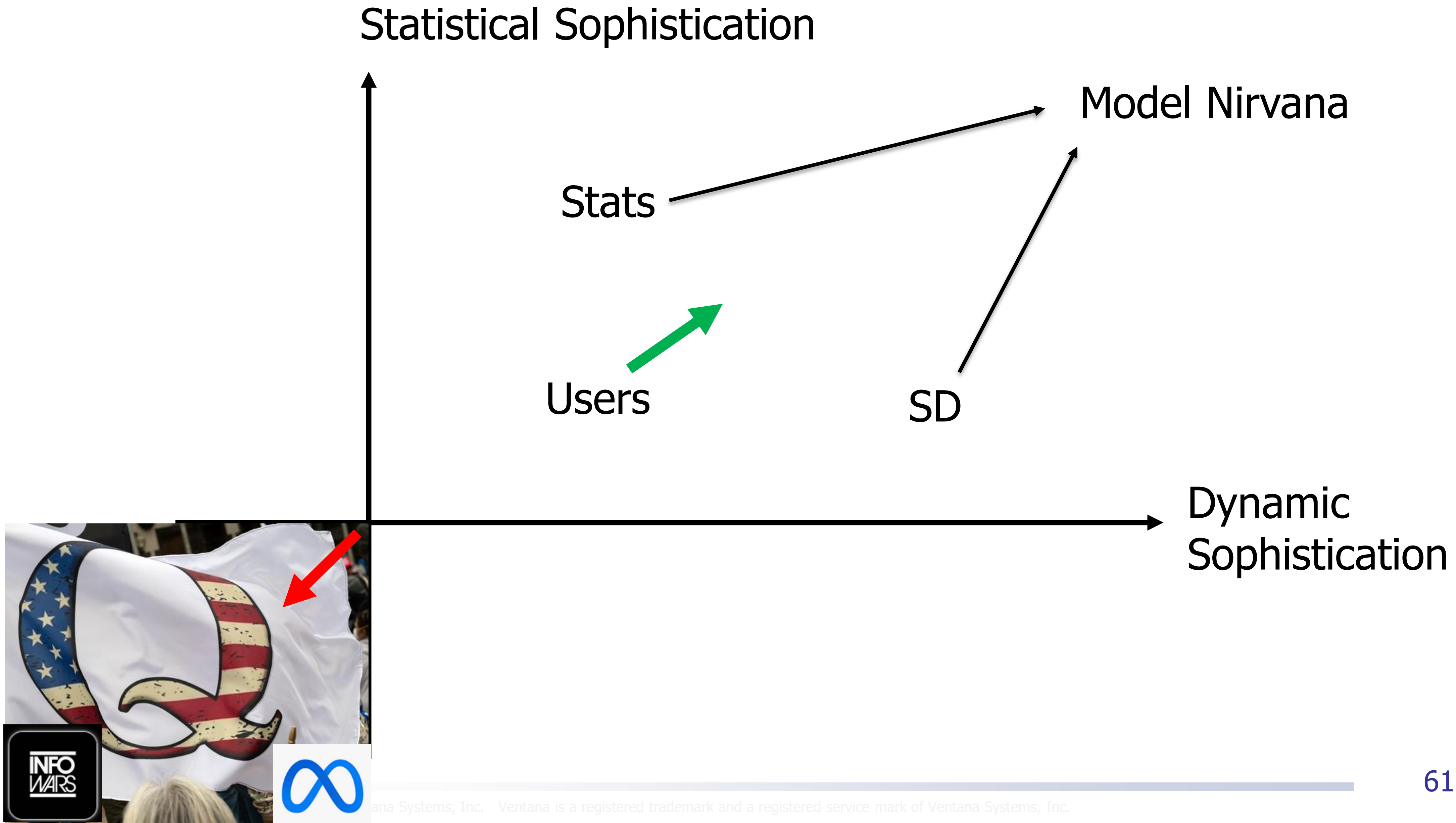
Jay was right!

- **A good model has to come first**
  - Appropriate stocks, flows, feedback, nonlinearity
  - Dimensional consistency and material conservation
  - Decisions use available information
  - Robustness to extreme conditions
- **Models should reproduce all possible realizations of the data and test policies outside the historical range**
- **There are opportunity costs to intensive use of data**

Challenges for the Future

- What's the right way to mix data, calibration and dynamics?
- How do we make calibration and data easier and cheaper?
- What new skills do we need on modeling teams?
- How do we work with other communities, like machine learning?
- How can we bring users along without overwhelming them?

A New Challenge



Selected References & Resources

- **Vensim Data & Calibration workshops (ISDC 2022) https://vensim.com/conference/#using-data-in-vensim (*)**
- **Vensim manuals and sample models (*)**
- **VenPy https://github.com/VensimOfficial/venpy**
- **Gelman, Carlin, Stern & Rubin (1995-2020) Bayesian Data Analysis, http://www.stat.columbia.edu/~gelman/book/ (*)**
- **Nathaniel Osgood & Juxin Liu (2015) Combining Markov Chain Monte Carlo Approaches and Dynamic Modeling, in Rahmandad & Oliva, Analytical Methods for Dynamic Modelers, MIT Press**
- **Nathaniel Osgood (2022) Using Particle Filtering with Dynamic Models in Health: Overview & Intuition, https://youtu.be/dHf-MM9WlIg (*)**
- **Jair Andrade and Jim Duggan (2021) A Bayesian approach to calibrate system dynamics models using Hamiltonian Monte Carlo, SDR https://onlinelibrary.wiley.com/doi/pdf/10.1002/sdr.1693 (*)**
- **"Behavioral dynamics of COVID-19: estimating underreporting, multiple waves, and adherence fatigue across 92 nations." Rahmandad H, Lim T, Sterman J (2021) *System Dynamics Review* 37(1):5-31.**
- **"Simulation-based estimation of the early spread of COVID-19 in Iran: actual versus confirmed cases." Ghaffarzadegan, N., Rahmandad, H. (2020) System Dynamics Review, 36(1):101-129**

- **(*) at least these items are open access last time I checked**

VENTANA

Thanks!