Data and Uncertainty in System Dynamics

Forrester, Kalman, Markov & Bayes

Tom Fiddaman
SD Health SIG
September 1, 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

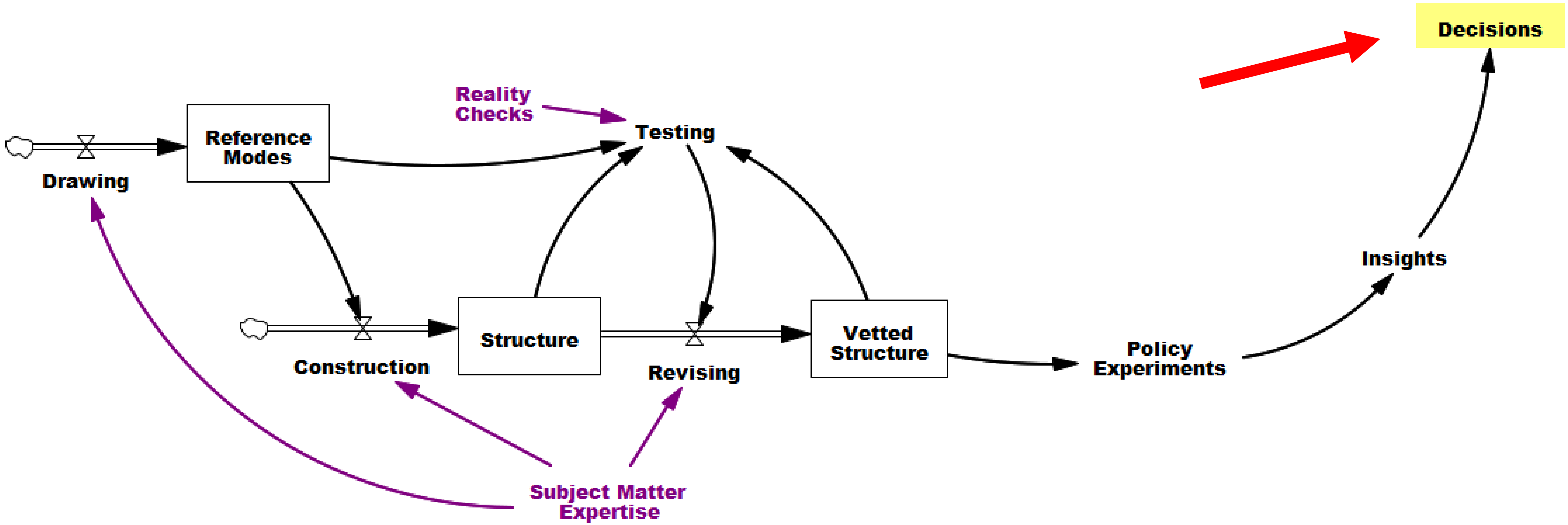
- **Introduction**
- **Example – State COVID19 Policy**
- **Methods**
  - Naïve calibration
  - Maximum likelihood
  - Kalman filtering
  - Bayesian inference
  - Markov Chain Monte Carlo (MCMC)
  - Synthetic data
- **References**

A Classic SD Perspective on Data

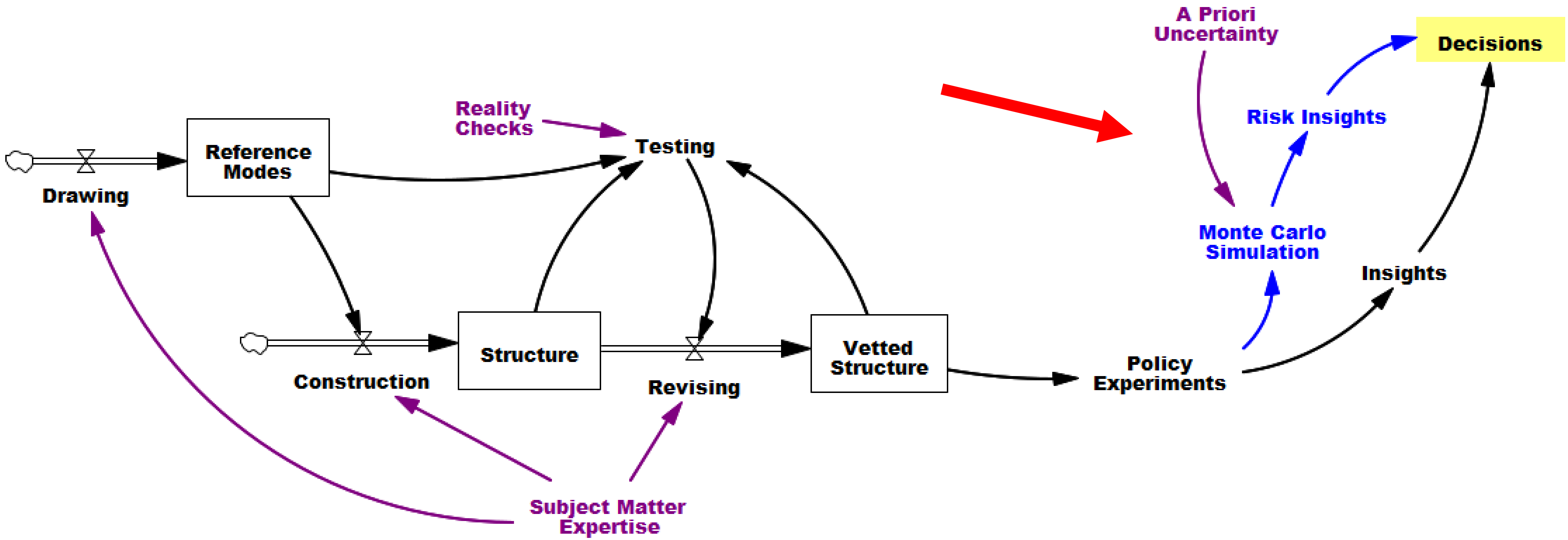
- **fitting curves to past system data can be misleading**
- **given a model with enough parameters to manipulate, one can cause any model to trace a set of past data curves**
- **adjusting model parameters to force a fit to history may push those parameters outside of plausible values as judged by other 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)

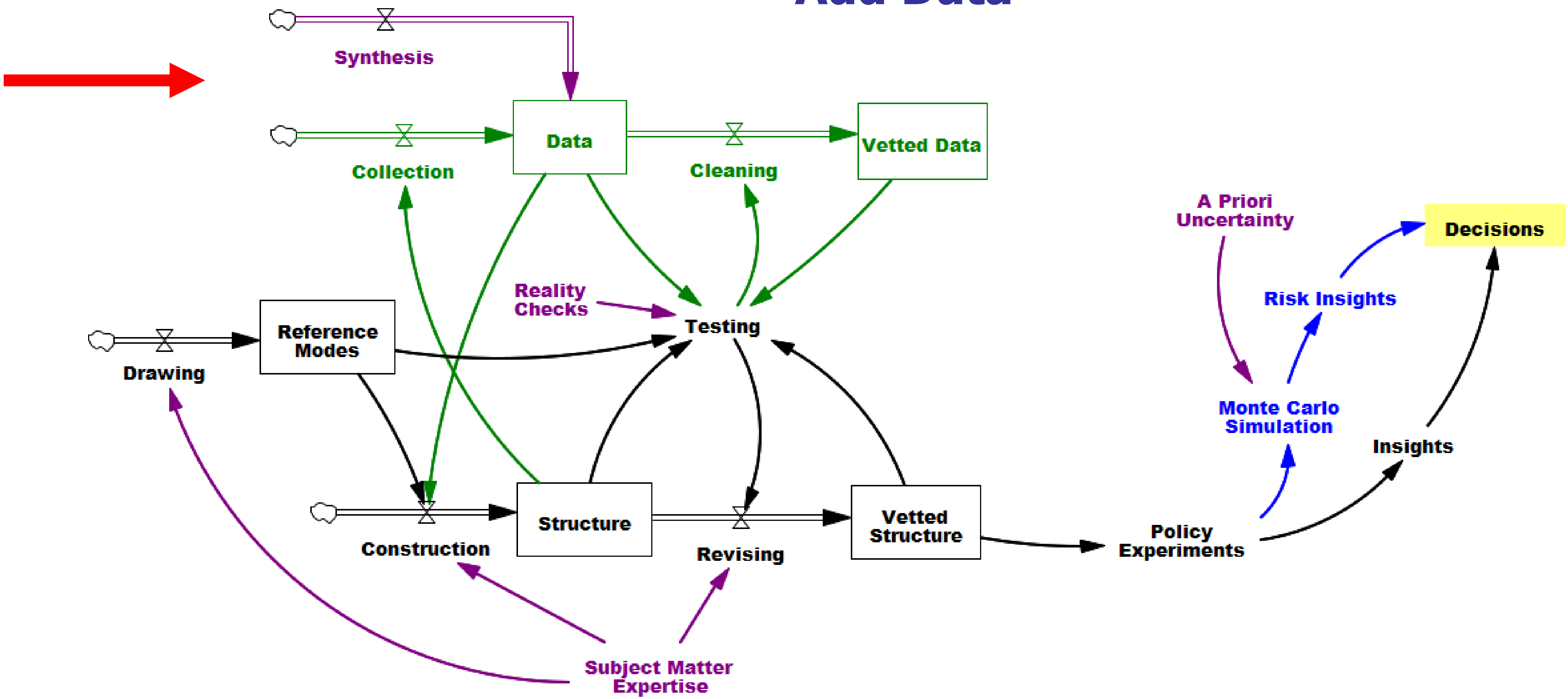
Classic SD


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

Add A Priori Risk & Uncertainty

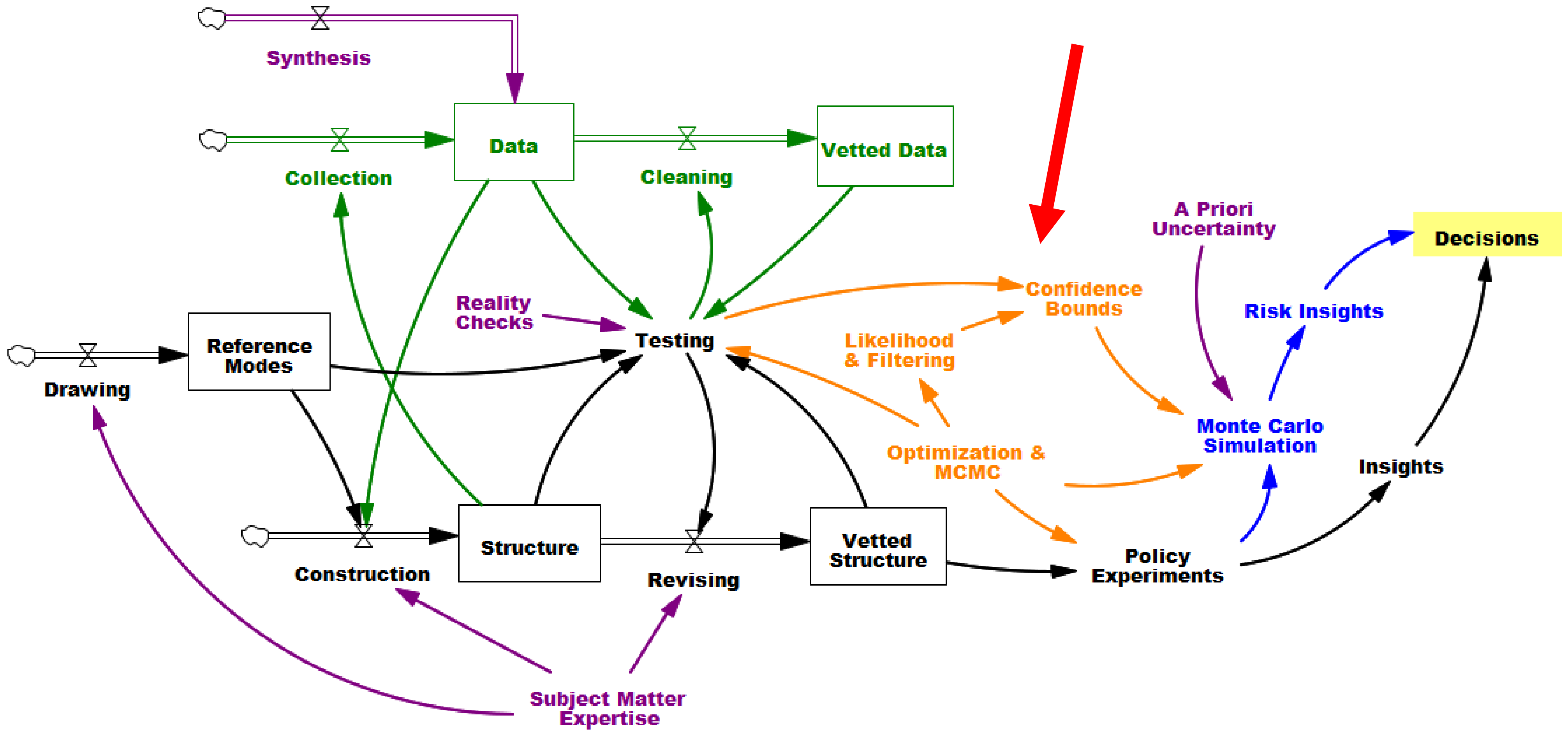


Add Data

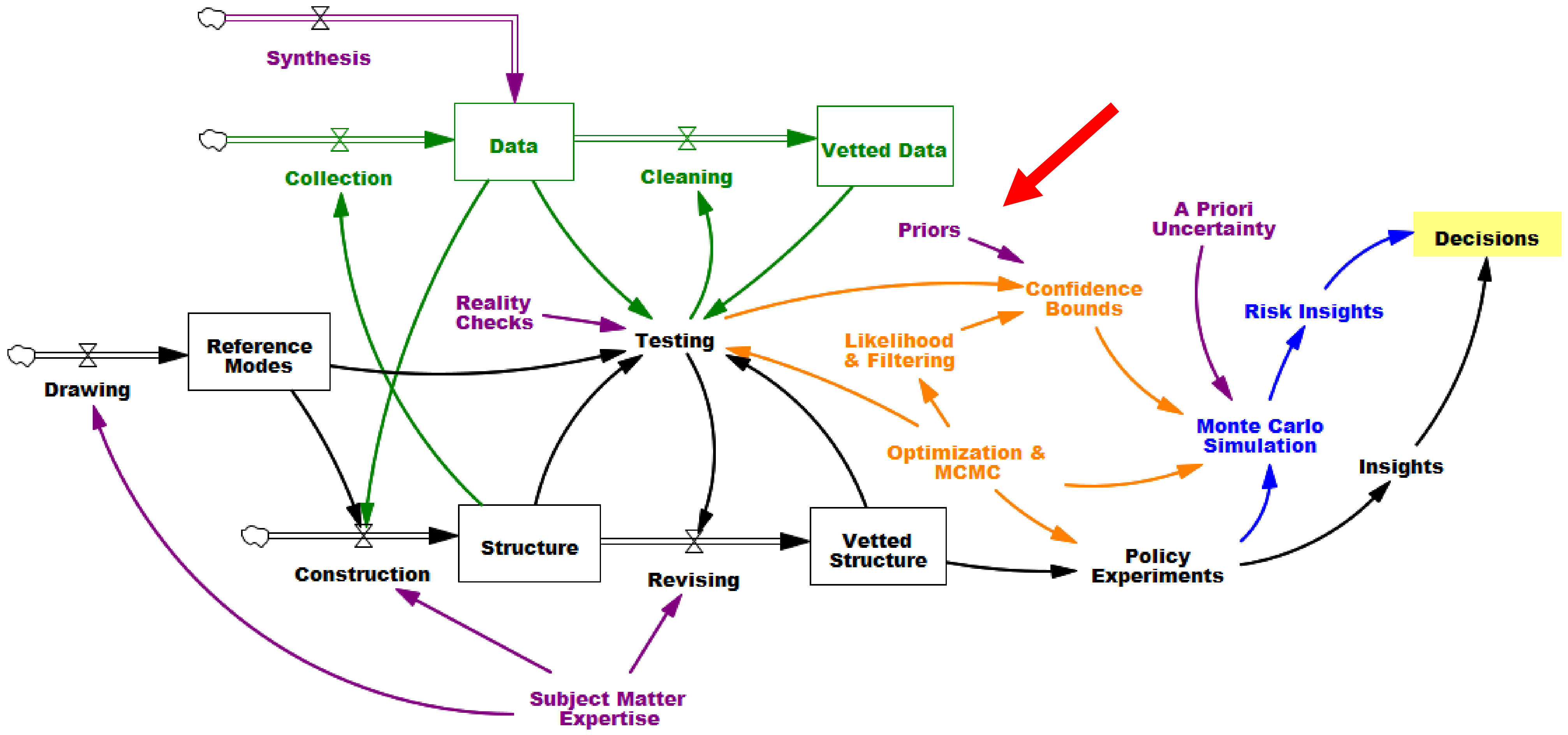


Synthesis
Data
Collection
Cleaning
Vetted Data
A Priori
Uncertainty
Decisions
Reality
Checks
Testing
Risk Insights
Reference
Modes
Drawing
Monte Carlo
Simulation
Insights
Structure
Vetted
Structure
Construction
Revising
Policy
Experiments
Subject Matter
Expertise

Add Likelihood & Algorithmic Support



Add Priors & Mindset = Bayesian SD



VENTANA

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**

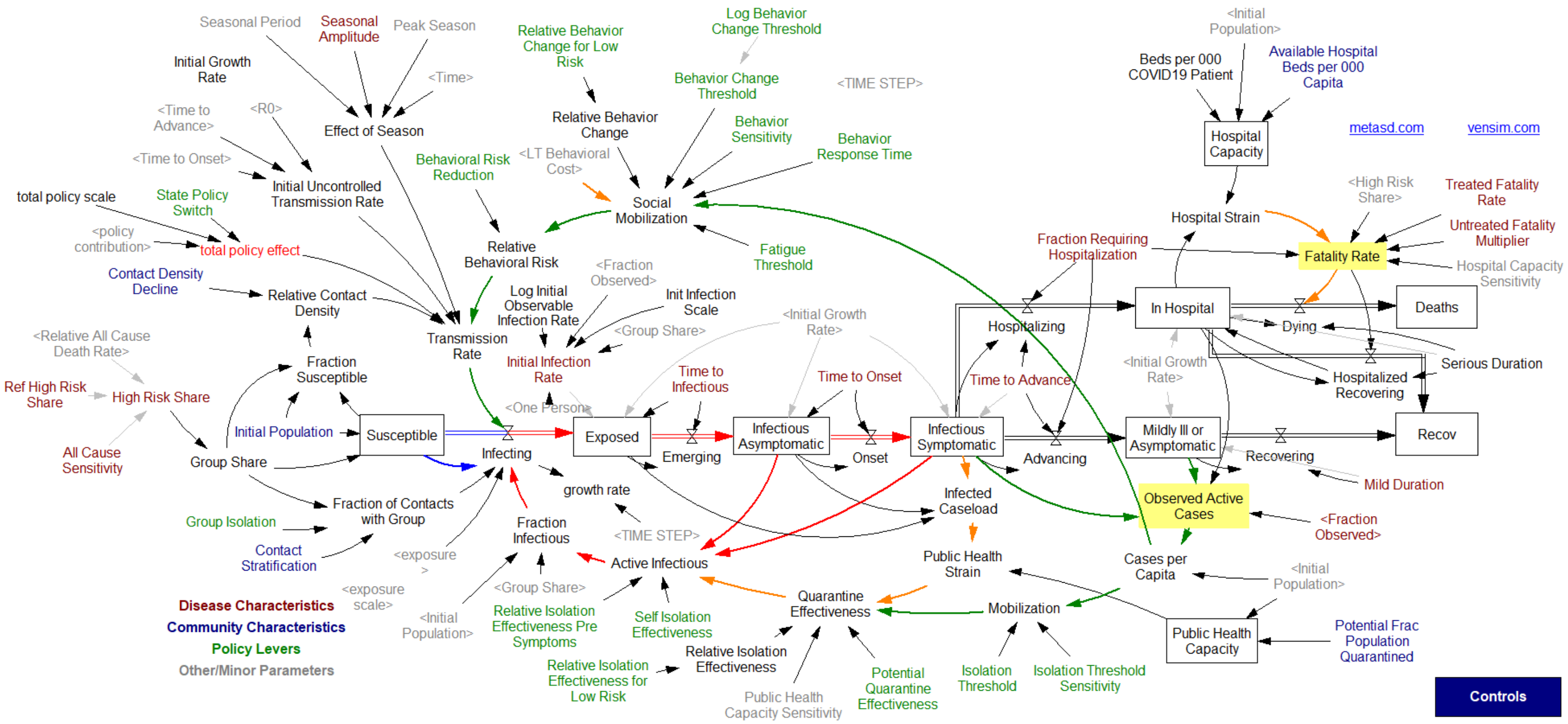
– **No learning about the data, or from the data directly**
– **No contribution to model quality from tests against data**
– **Hard to verify that asserted reference modes or decision structures match reality**
– **Less face validity of historical runs**
– **Difficulty understanding the gaps between 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**

State COVID19 Modeling

- **Context**
  - Early days of the pandemic – April-June 2020
  - Red state with a tech-savvy governor
- **Questions**
  - Tactical interpretation of new data (almost daily)
  - What emergency medical resources will be needed (i.e. when will hospitalization peak, at what level)?
  - How many tests are needed?
  - What are the consequences of reopening (thereby cancelling many NPIs)?
- **No data = no project**

The Model

- **Enhanced SEIR**
  - Higher-order delay structure
  - Test coverage effects
  - NPI policies
  - Behavior, compliance
  - Detailed hospital sector
- **Integrate data streams**
  - Cases, hospitalization, deaths
  - Test composition
  - Mobility (cell phones)
  - Weather, population, etc.

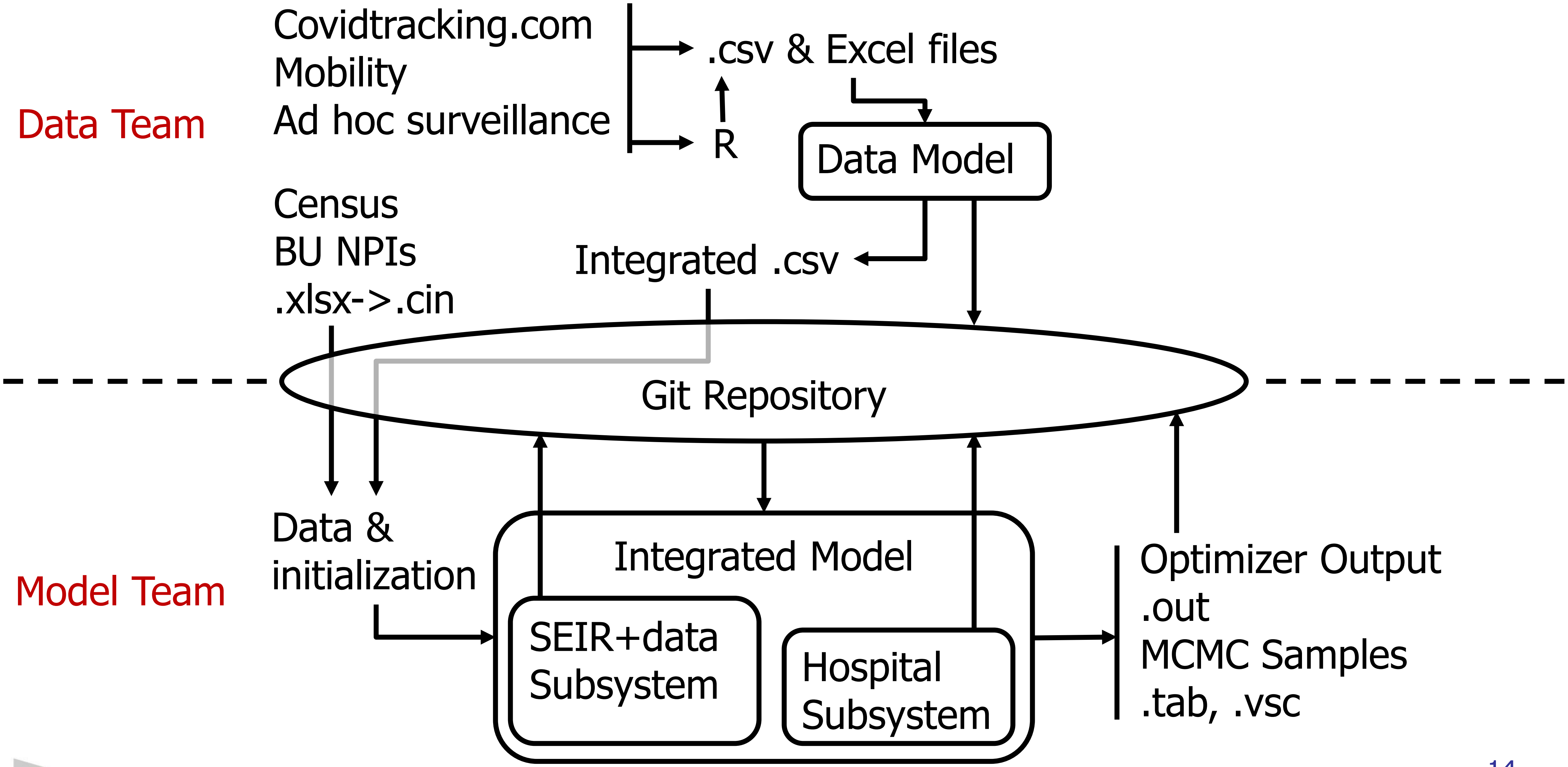


Detail

- 2 risk groups
- 50 states
- 50 x 50 interstate transmission (reduced by adjacency)
- 45 NPIs (essentially exogenous step functions)
- 5 hospital resource classes

- 7353 stocks
- 13003 constants (about 300 estimated)
- 2351 time series data (about 400 used for estimation)

Project Architecture



Daily Routine

- **Data**
  - Update and push to SVN
  - Discuss
- **Model**
  - Pull data and import to vdf
  - Launch optimization c
  - Review calibrati ts
- **Policy discu**
  - Revi rent R and prognosis for growth/decline
  - er how new policies might influence dynamics
  - xplore contingencies and uncertainty

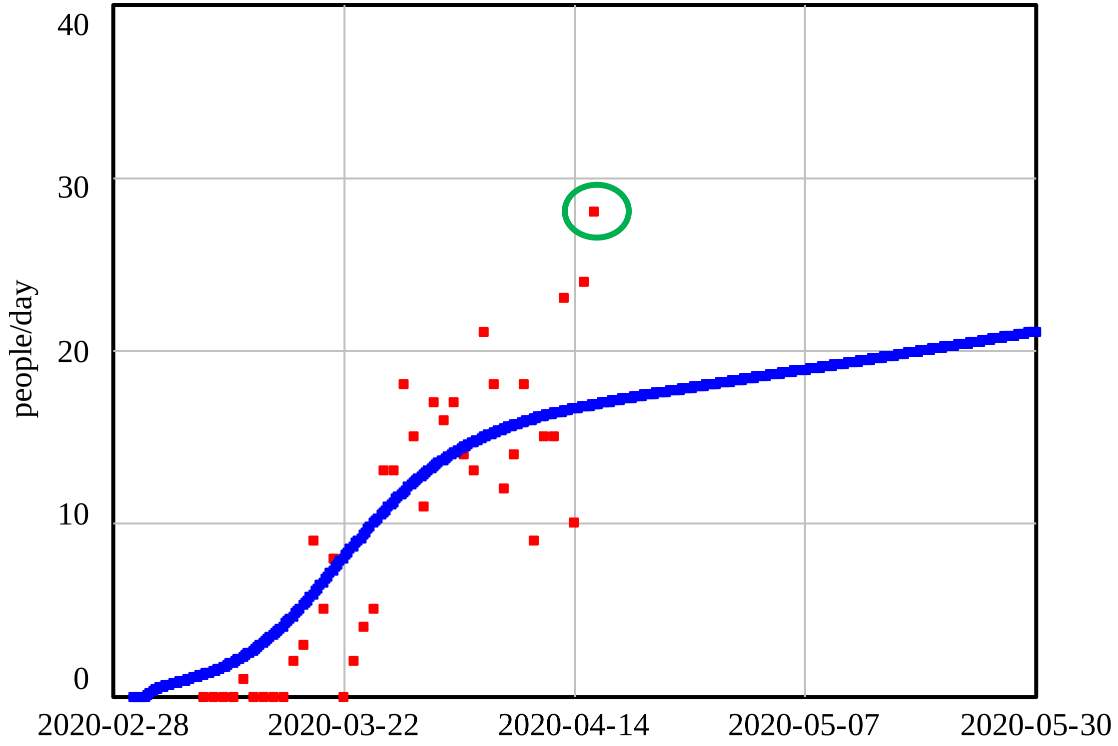
Disrupt cycle for an impending press or legislative event

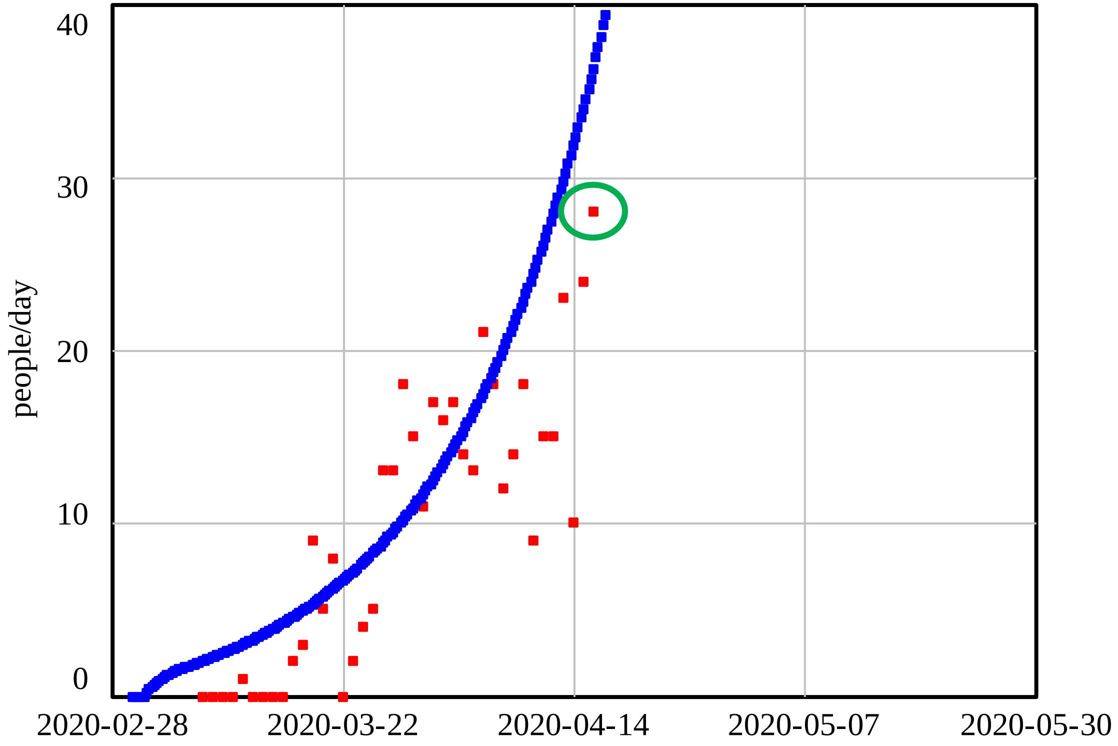
Calibration

- **Mix of conventional hill-climbing (Powell) and MCMC**
- **Case, hospitalization and death counts turned out to have an overdispersed Poisson distribution – least squares (assuming Normal distribution) didn't work well**
- **Timely computation required parallel Vensim and a 60-core HPC server.**

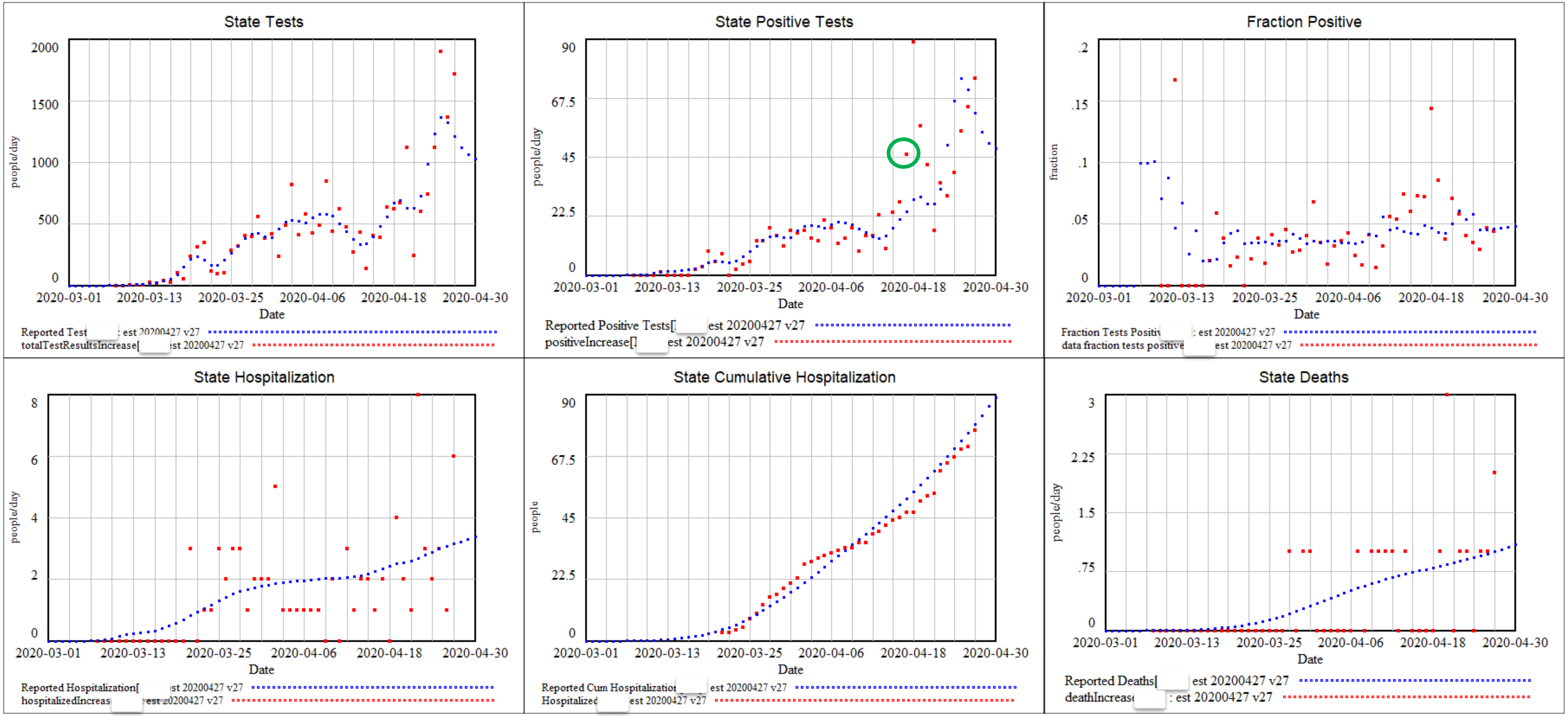
Tax Day, 2020: Reason for concern?

Positive Tests

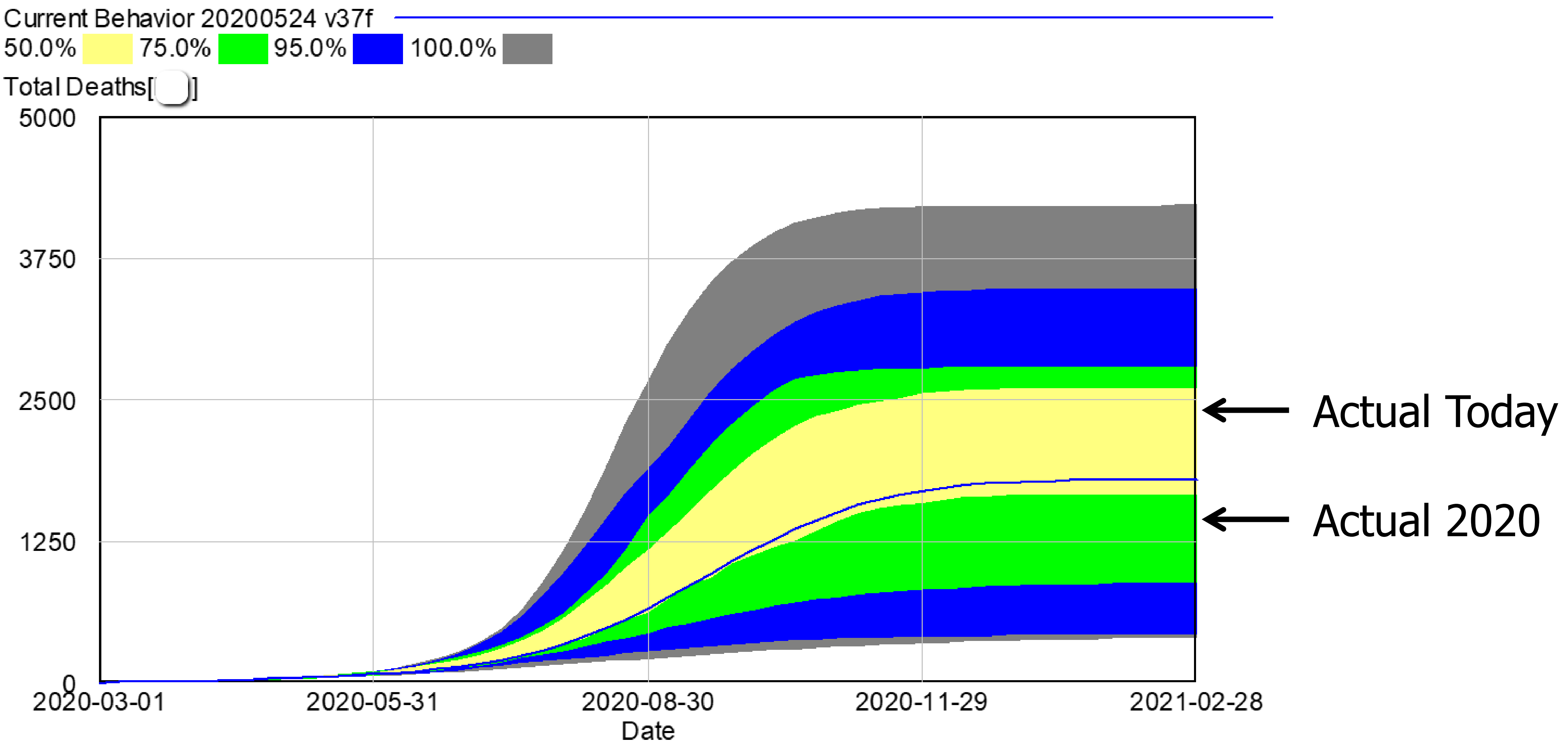




10 Days Later



Outputs



Current Behavior 20200524 v37f
50.0% 75.0% 95.0% 100.0%
Total Deaths[ ]
5000
3750
2500
1250
0
2020-03-01
2020-05-31
2020-08-30
2020-11-29
2021-02-28
Date
Actual Today
Actual 2020

Contingency Tables for Uncertain Outcomes

Average Deaths across Uncertain Scenarios

Average of Total Deaths — Reopening Extent (Column Labels)

Tests/day	Row Labels (Isolation Effectiveness)	0%	20%	40%	60%	80%	100%
625	0%	971	2599	2132	2401	3501	4983
	20%	524	1475	2022	2689	3401	4029
	40%	438	657	1005	1468	1827	3442
	60%	248	560	308	690	1202	3141
	80%	178	197	257	333	281	401
	100%	140	125	135	173	169	163
1250	0%	1059	1569	2150	2827	3240	4176
	20%	788	1088	1936	2409	2498	3854
	40%	377	605	1100	[illegible]	[illegible]	3423
	60%	278	458	454	[illegible]	[illegible]	2024
	80%	156	205	284	[illegible]	[illegible]	776
	100%	150	148	148	[illegible]	[illegible]	213
2500	0%	904	2112	2469	[illegible]	3873	4053
	20%	990	1106	[illegible]	[illegible]	[illegible]	3831
	40%	470	531	947	1243	2406	3393
	60%	223	349	[illegible]	703	1639	[illegible]
	80%	174	151	189	238	581	804
	100%	140	141	144	191	152	216
5000	0%	1278	1719	2234	2638	3750	4187
	20%	469	1411	1686	2070	3406	4027
	40%	346	[illegible]	1056	1438	2373	3681
	60%	272	276	446	508	812	2176
	80%	197	142	308	262	551	550
	100%	121	150	154	161	181	163
10000	0%	1074	1557	1787	2837	3233	4322
	20%	650	951	2281	2128	3031	4266
	40%	333	565	1337	1517	2041	3592
	60%	171	325	528	561	768	999
	80%	131	182	238	259	1279	650
	100%	135	[illegible]	157	131	217	180

Best Guess Current State

VENTANA

Some representative findings

Question	Model-based advice	Outcome
Is a 3x spike in cases in a few days a looming problem, or a blip?	Don't panic.	Didn't panic. It was a blip.
Should a sports arena be converted to a field hospital?	Probably not needed, and the model would provide some early warning if it is.	Did it anyway. Not used.
Is it worth it to buy enough tests for 10,000/day?	Probably don't need that many, but at least 5,000 helps mitigate worst scenarios	Testing about 5,000/day; leading nation per capita.
Is contact tracing worth it?	Yes, if followed up by other measures.	Leading nation per capita. Later overwhelmed and challenged by noncompliance, largely abandoned.
Do masks work?	No data. Any cheap measure that reduces transmission without distancing is a win.	No mask order; low usage.
What happens when you reopen the economy?	Resurgence of disease, unless you have a strong NPI strategy.	Reopened without a strategy. Big fall/winter wave.

Outcomes

- **Stopped chasing noise in the data**
- **Demonstrated that hospital loads requiring top-tier response were unlikely**
- **Warned that early reopening with optional NPIs incurred some very nonlinear risks**

- **Within a few months,**
  - Predicted death tolls were modest
  - Response choices were fully politicized

- **As a result,**
  - Our contract was not renewed
  - Models were basically squeezed out of the debate
  - Health-economy tradeoffs were made with little or no explicit reasoning
  - It was probably possible to have lower health and economic consequences together

Some challenges

Early

- **Extremely sparse data overall**
- **Low test coverage and resulting biases**
- **Uncertainty about everything, especially**
  - True prevalence
  - Fatality rates
  - Hospital treatments
  - Policy effectiveness
- **Time pressure**

Ongoing

- **Long delays**
- **Limited surveillance of the general population**
- **Difficulty observing actual behavior**
- **Data restatements and definitional changes**
- **Difficulty connecting to economic effects**

Methods

- **Prerequisite: Data**
- **Synthetic data**
- **Naïve Calibration**
- **Maximum likelihood**
- **Kalman filtering**
- **Bayesian inference**
- **Markov Chain Monte Carlo (MCMC)**

My Typical Playbook

- Build/refine structure

- Load data
- 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

- Design policies
- Test policies deterministically

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

- Do policy experiments with sensitivity

Fundamentally, what are we doing?

- **Create a model of the process that generated the data**
  - Dynamic structure
  - Distribution of errors in 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: I'm 95% certain that the parameter lies within my credible interval

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 – discharching + NOISE, initial patients)
  - Truncate the frequency and duration of the measurements
  - Use the synthetic data to see if you can recover the parameters
  - More fun if you have an adversary!

Naïve Calibration

- **Build a control panel that compares model and data series**
- **Do some hand calibration to discover what is important**

- **Automate using the optimizer**
- **Don't worry 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

Maximum Likelihood

- **Choose the value of parameters that maximizes the likelihood of observing the data given the model**

- **This yields a Maximum Likelihood Estimator (MLE)**

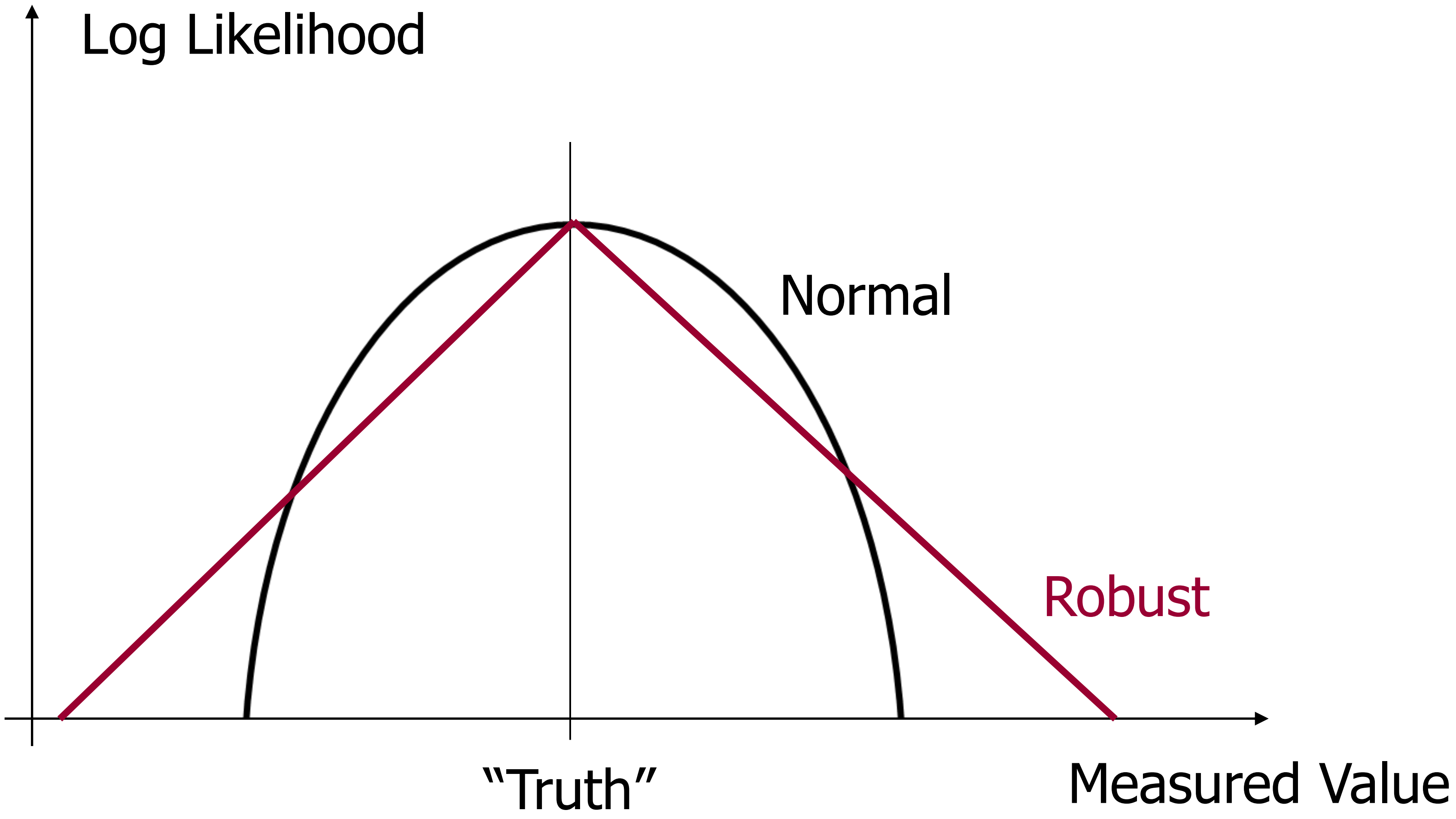
- **Suppose there is more than one observation**
  - Then the likelihood is the product of the individual likelihoods for each data point
  - Working with log likelihood is easier, because ln() converts the product to a sum

- **Likelihood expresses the probability of getting the data observed from your model, not the chance that the model is right**

Log-Likelihood
Gaussian errors

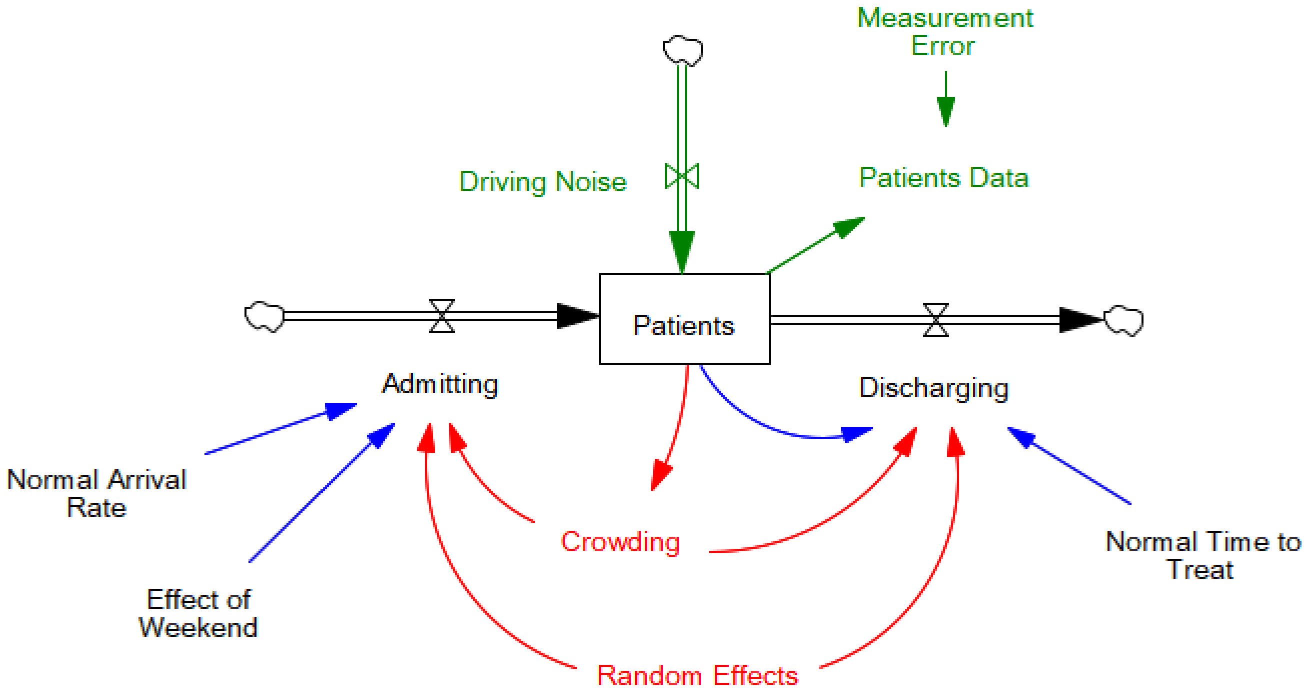
- **Likelihood =** $\frac{\mathbf{1}}{\boldsymbol{\sigma}\sqrt{\mathbf{2\pi}}}\boldsymbol{e}^{-\left(\frac{\boldsymbol{model-data}}{\boldsymbol{\sigma}}\right)^{\mathbf{2}}/\mathbf{2}}$
  - This is the PDF of the Gaussian (Normal) distribution
  - σ represents the scale of the error associated with a 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 multiplication becomes addition**
- **Log(Likelihood) =**
  – LN($\boldsymbol{\sigma}$) - the bigger the $\boldsymbol{\sigma}$, the lower the likelihood, as it's spread thinner
  – LN($\sqrt{\mathbf{2\pi}}$) - this is a constant we can ignore
  – $\left(\frac{\boldsymbol{model-data}}{\boldsymbol{\sigma}}\right)^{\mathbf{2}}/\mathbf{2}$ - the weighted sum of squares, adjusted by the divisor /2

What does the likelihood look like?

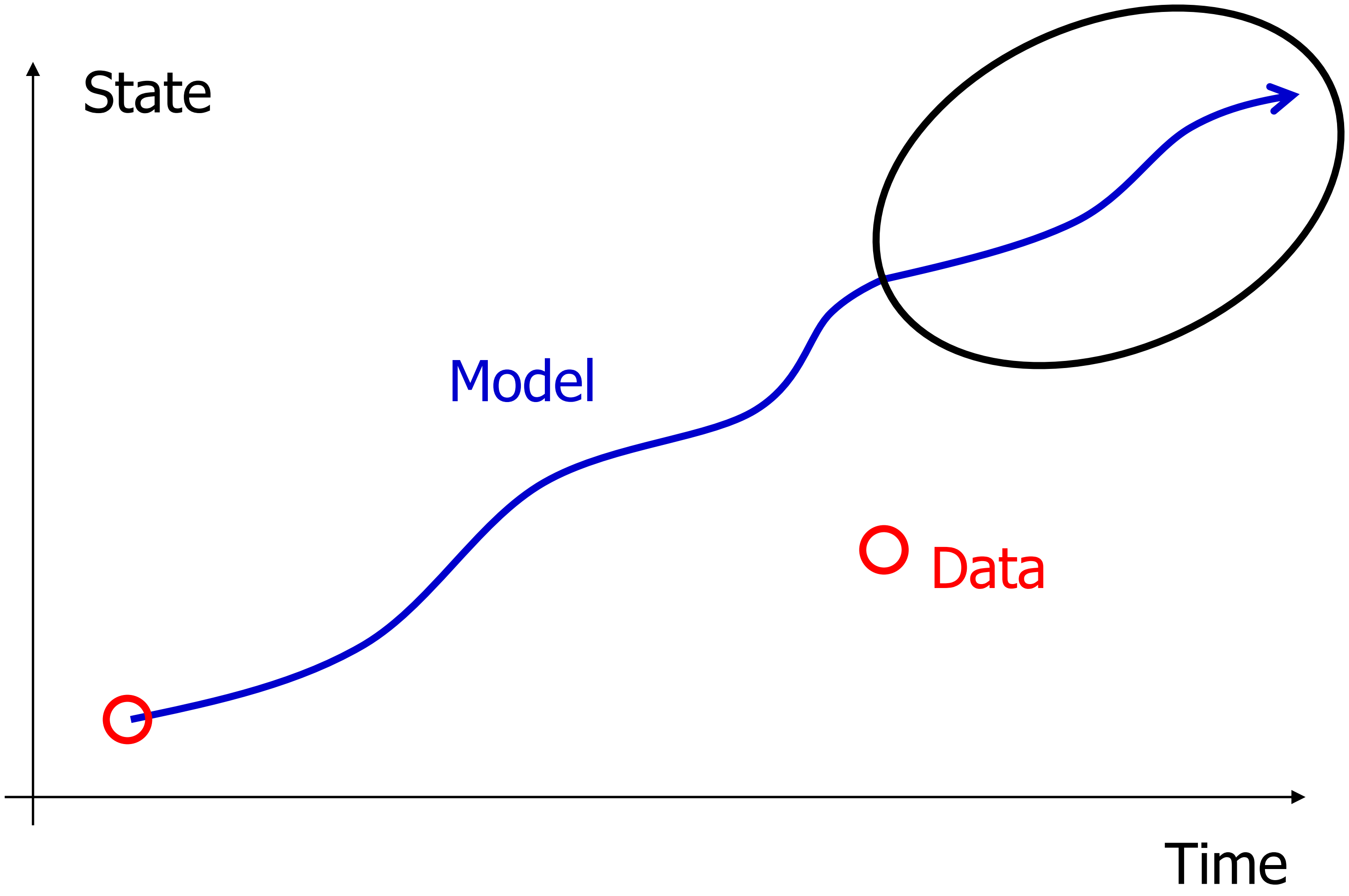


Kalman filtering

- **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**

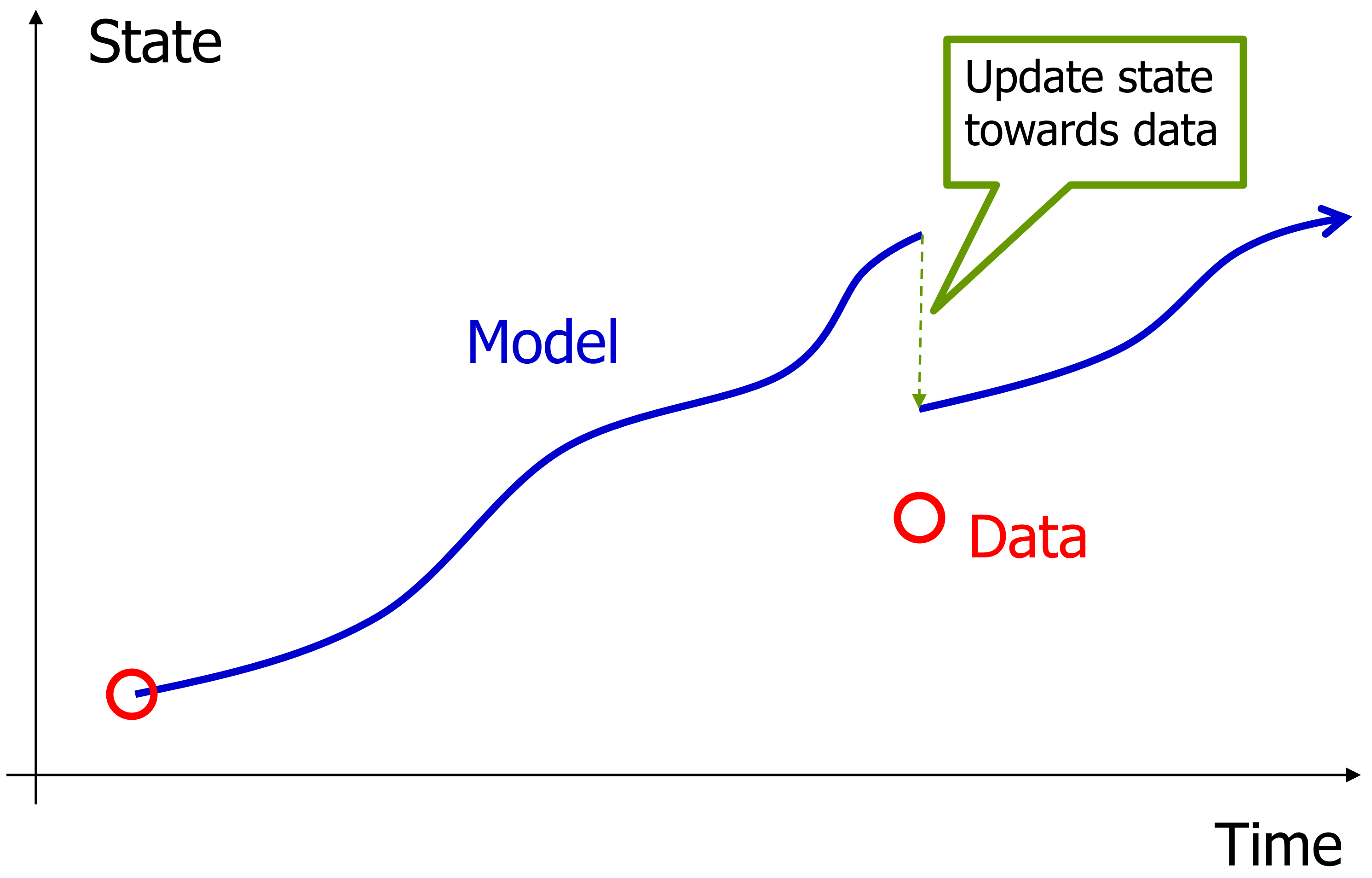


If the model is in the wrong part of the state space, its response may be wrong

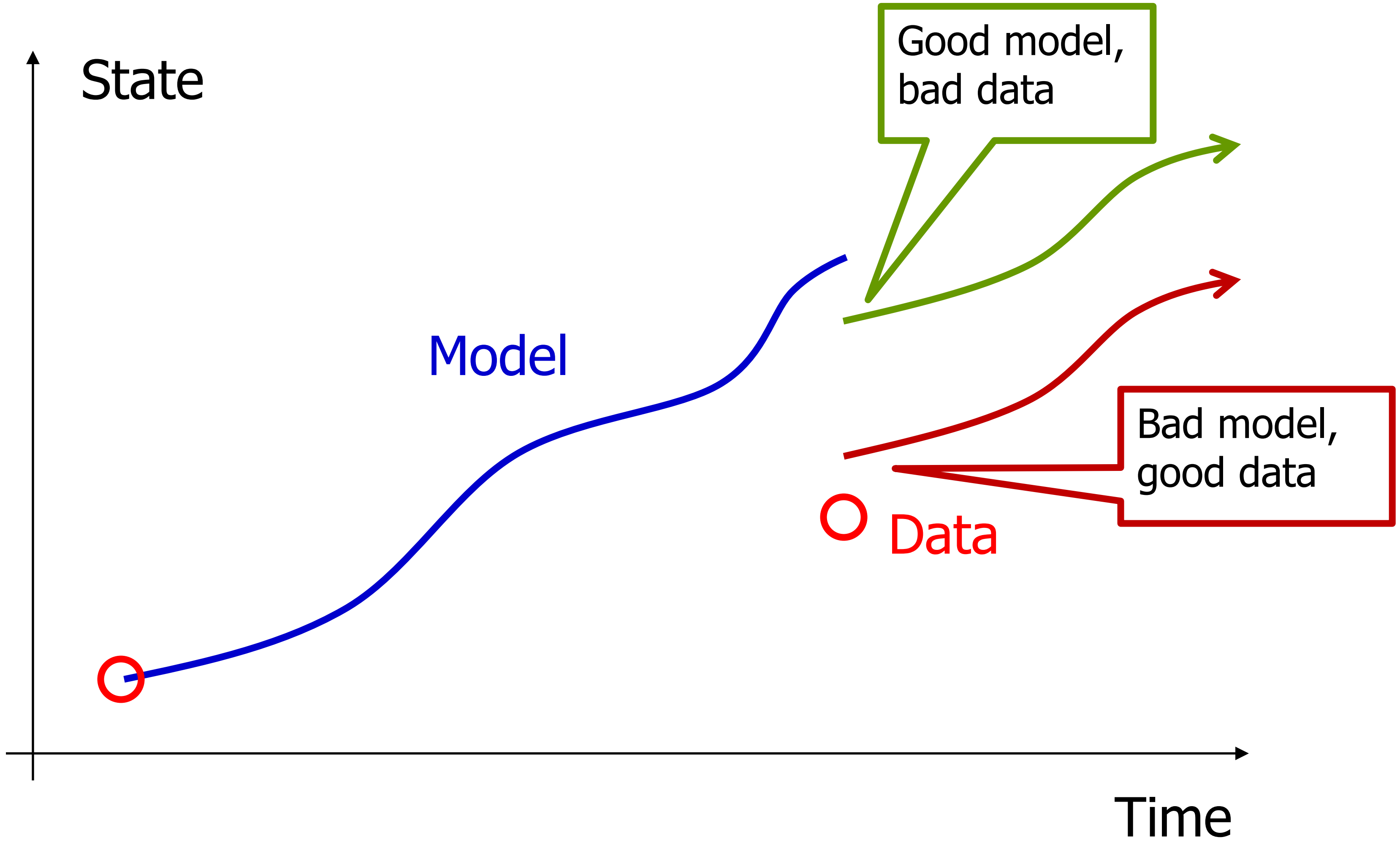


If you're estimating a parameter that affects this part of the trajectory, the response may be particularly wrong

The Kalman filter updates the model trajectory towards the data



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



Bayesian Inference

https://www.nature.com/articles/s41409-021-01473-w

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)**

Posterior

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

Likelihood **Prior** **Ignore**

The Answer

"Traditional" Calibration

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

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

Priors

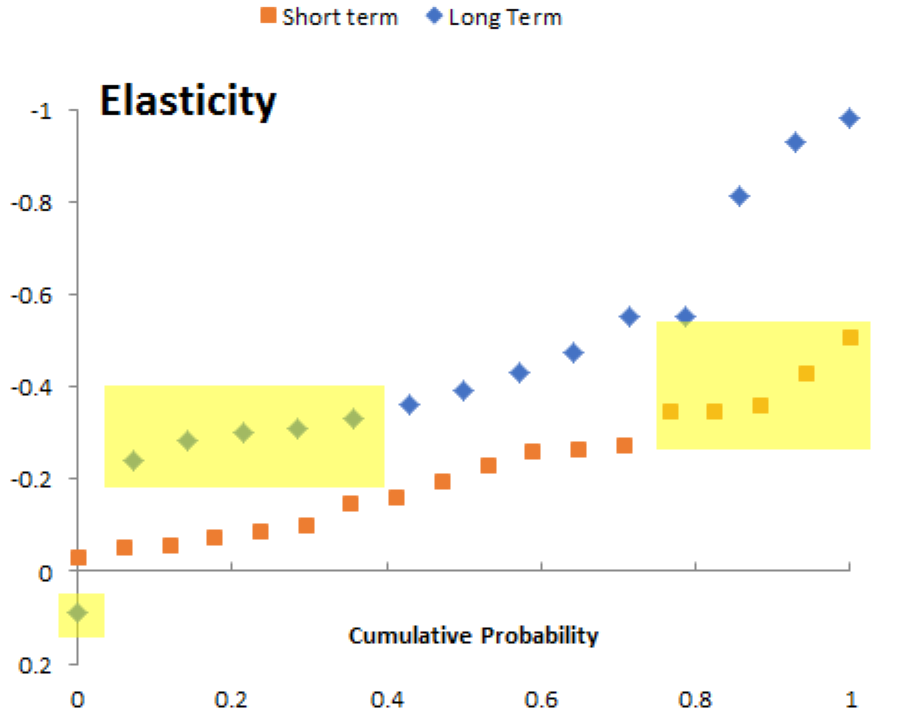
- **No priors = uniform priors**
  - This is essentially what we've been doing so far
  - It's not always a good choice, *but* if you have lots of data, it might not matter.
- **Non-informative or Maximum Entropy priors**
  - Contribute as little information as possible, i.e. assume maximum ignorance a priori
  - For a scale parameter like a time constant, this is -LN(param) for positive parameters
  - This can be tricky to construct
- **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

Problems with Intuitive Sourcing of Parameters

- **Example: price elasticity**
- **Simple concept, but …**

- **Estimates of short and long term overlap**
- **Most ignore behavioral phenomena**
- **Many short term estimates are non-robust (implying explosive economic response)**

Table 1. Price elasticities of fuel demand reported in the literature, by average year of observation.



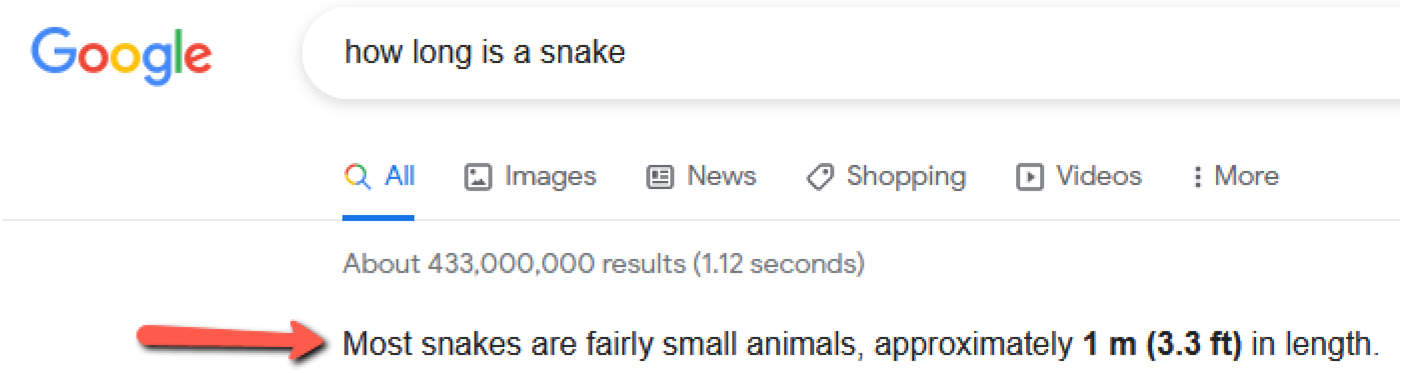
https://metasd.com/2019/03/challenges-sourcing-parameters-dynamic-models/
https://doi.org/10.1016/j.tra.2017.06.001

Expressing Priors

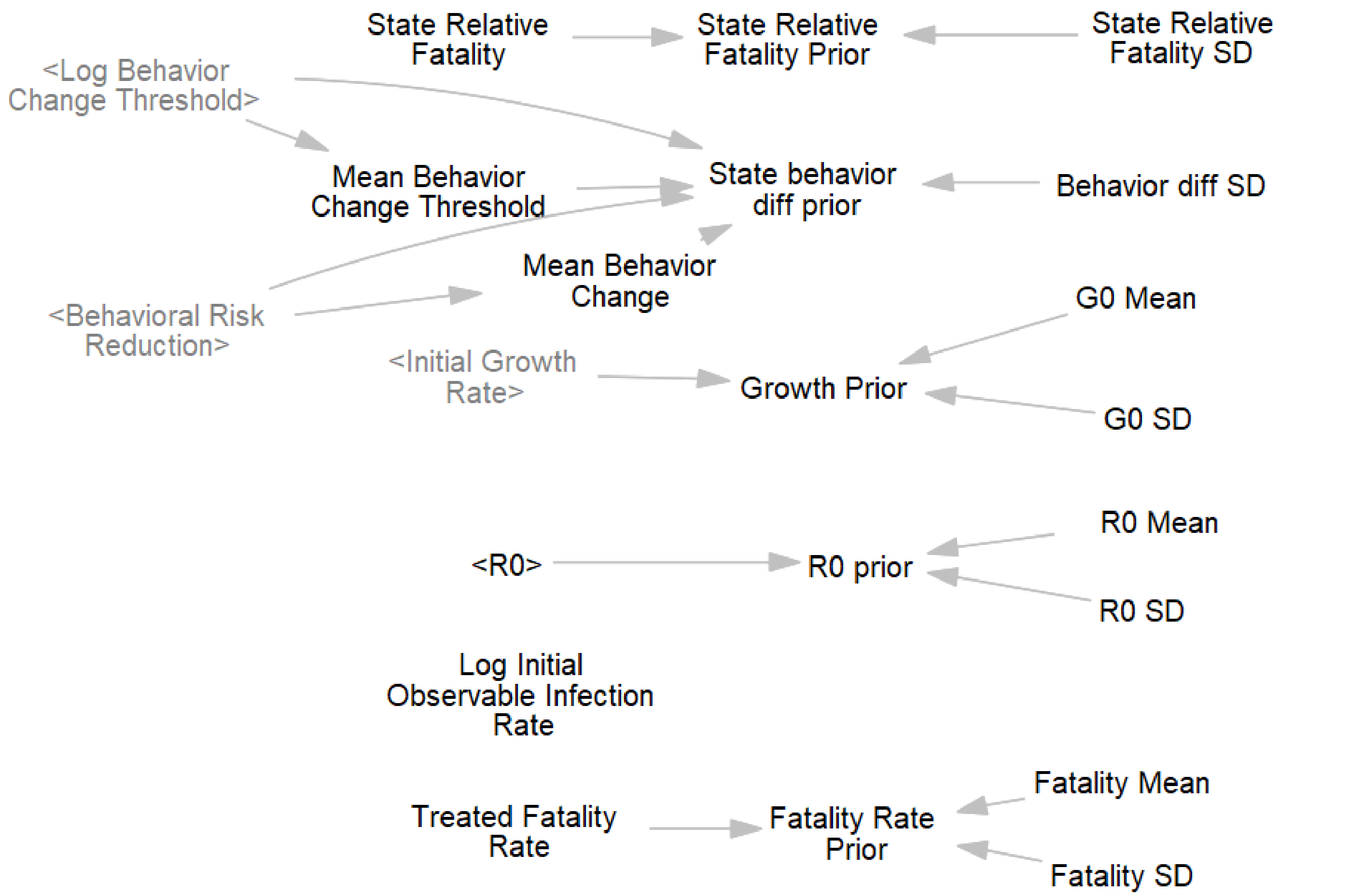
- **A prior is a lot like a data point!**

- **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 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 states?**
- **The prior likelihoods are additional terms in the payoff**

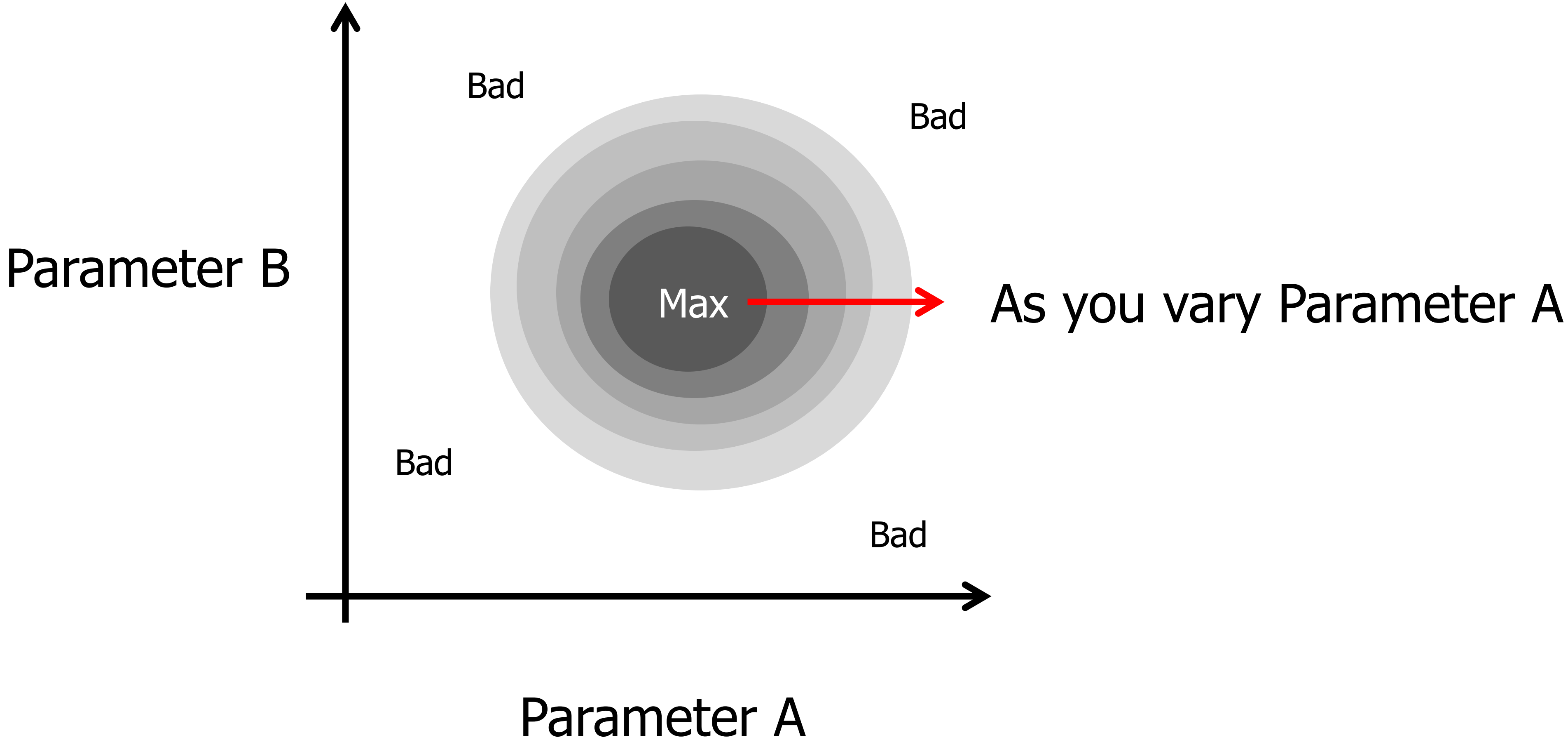
Enabled	Policy	Distribution	Timing	Transformation	Variable name	Compare To	Weight
☑	Calibration	OD Poisson		None	Reported Deaths[zone]	data deaths incr[zone]	death data w
☑	Calibration	OD Poisson		None	Reported Hospitalization[z	data Hospitalized incr[zone]	hospitalizatic
☑	Calibration	OD Poisson		None	Reported Positive Tests[zor	data test positive Incr[zone]	case data we
☑	Calibration	Gaussian		None	observed mobility effect[zc	data Google mobility[zone]	Mobility Effe
☑	Policy		Initial	None	R0 prior		1
☑	Policy		Initial	None	Fatality Rate Prior		1
☑	Policy		Initial	None	Frac Hosp Prior		1
☑	Policy		Initial	None	High Symptom Prior		1
☑	Policy		Initial	None	Generation Time Prior		1

VENTANA

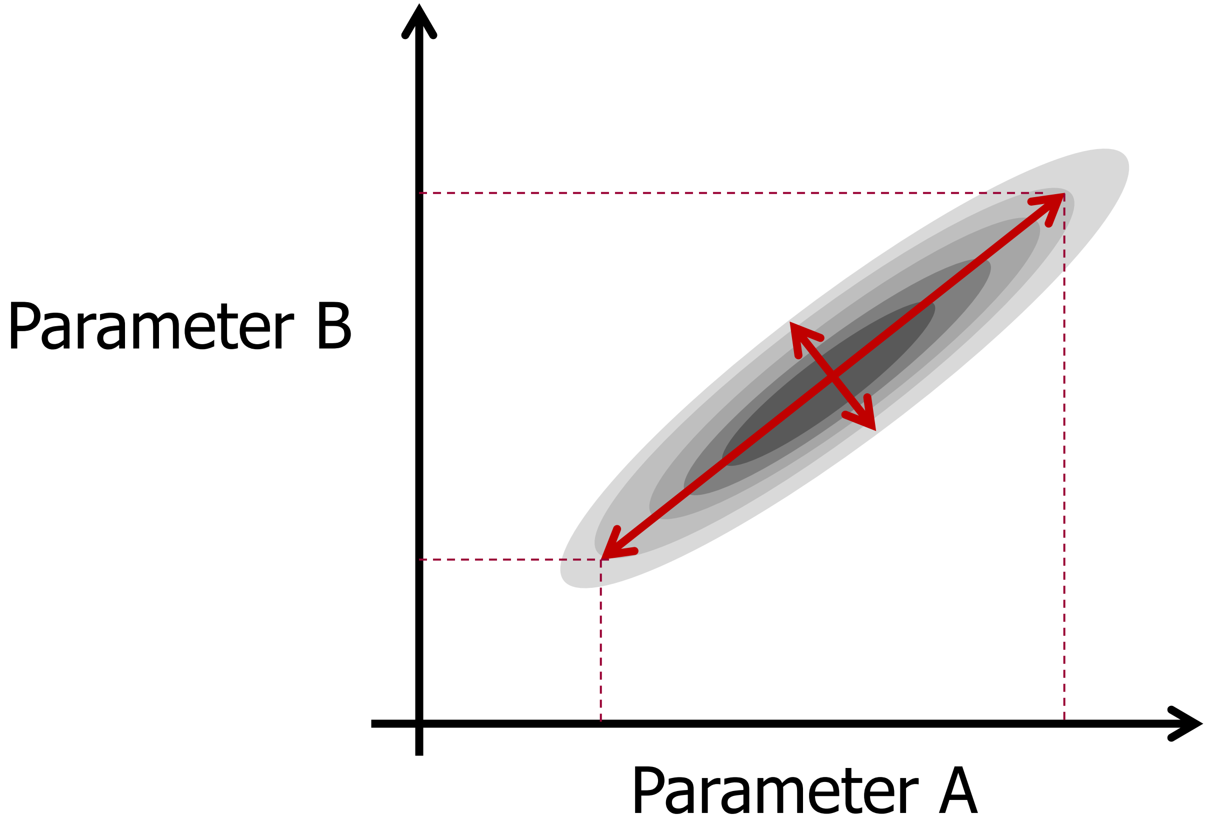
Confidence Bounds & Credible Intervals

- **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**

Using MCMC to Reveal the Posterior

Posterior

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

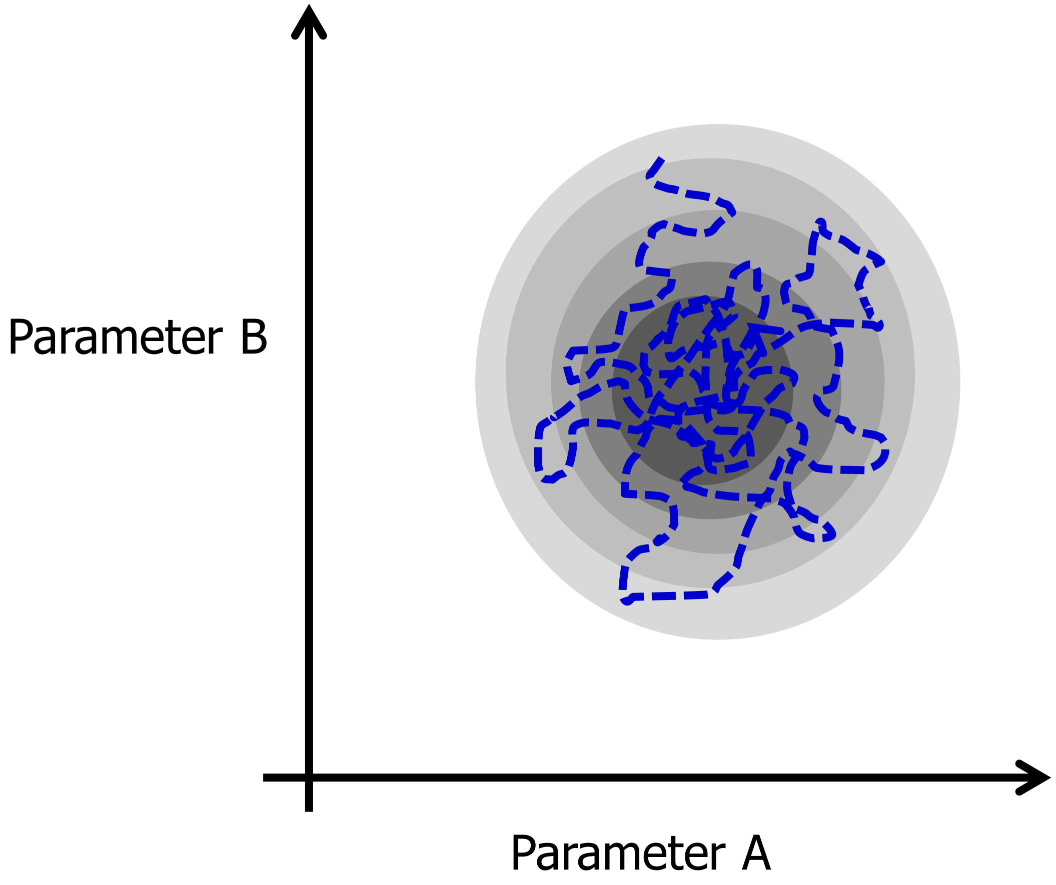
Likelihood **Prior** **Ignore**

We'd like to know this distribution in full.

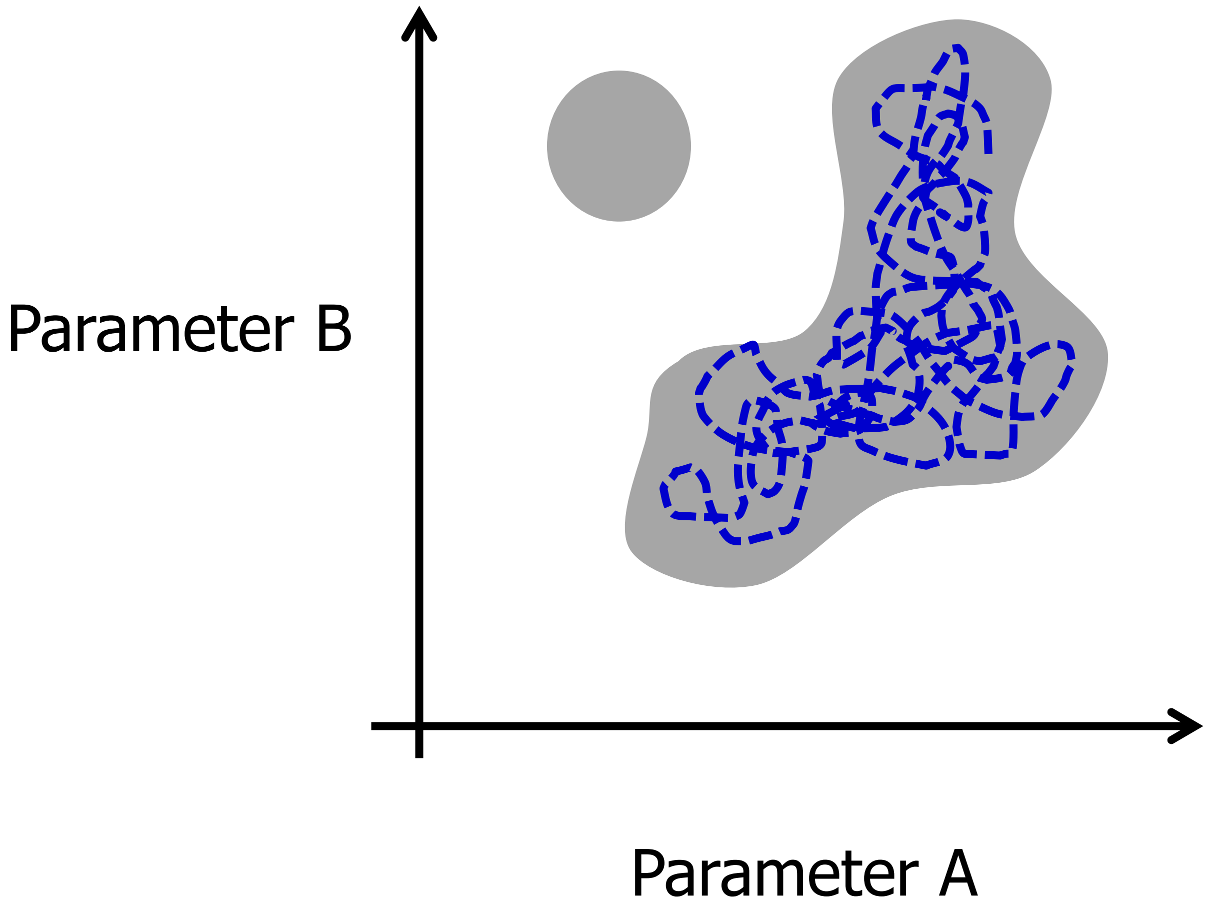
Without this term, the posterior isn't a properly-scaled probability distribution, i.e. it doesn't integrate to 1. So, we need a way to sample the posterior that only cares about the *relative* likelihoods of the data & priors.

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**

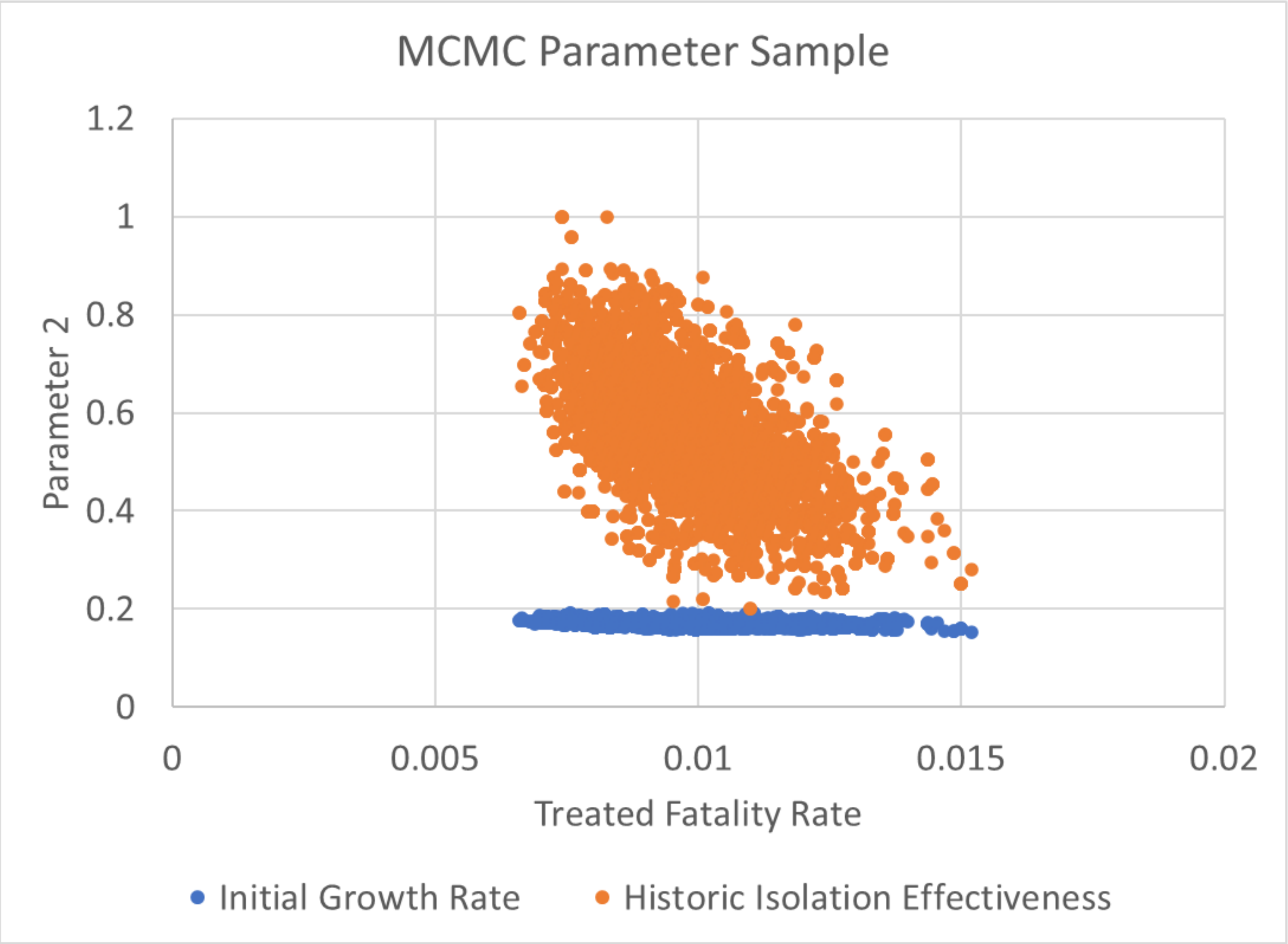


MCMC



MCMC Output

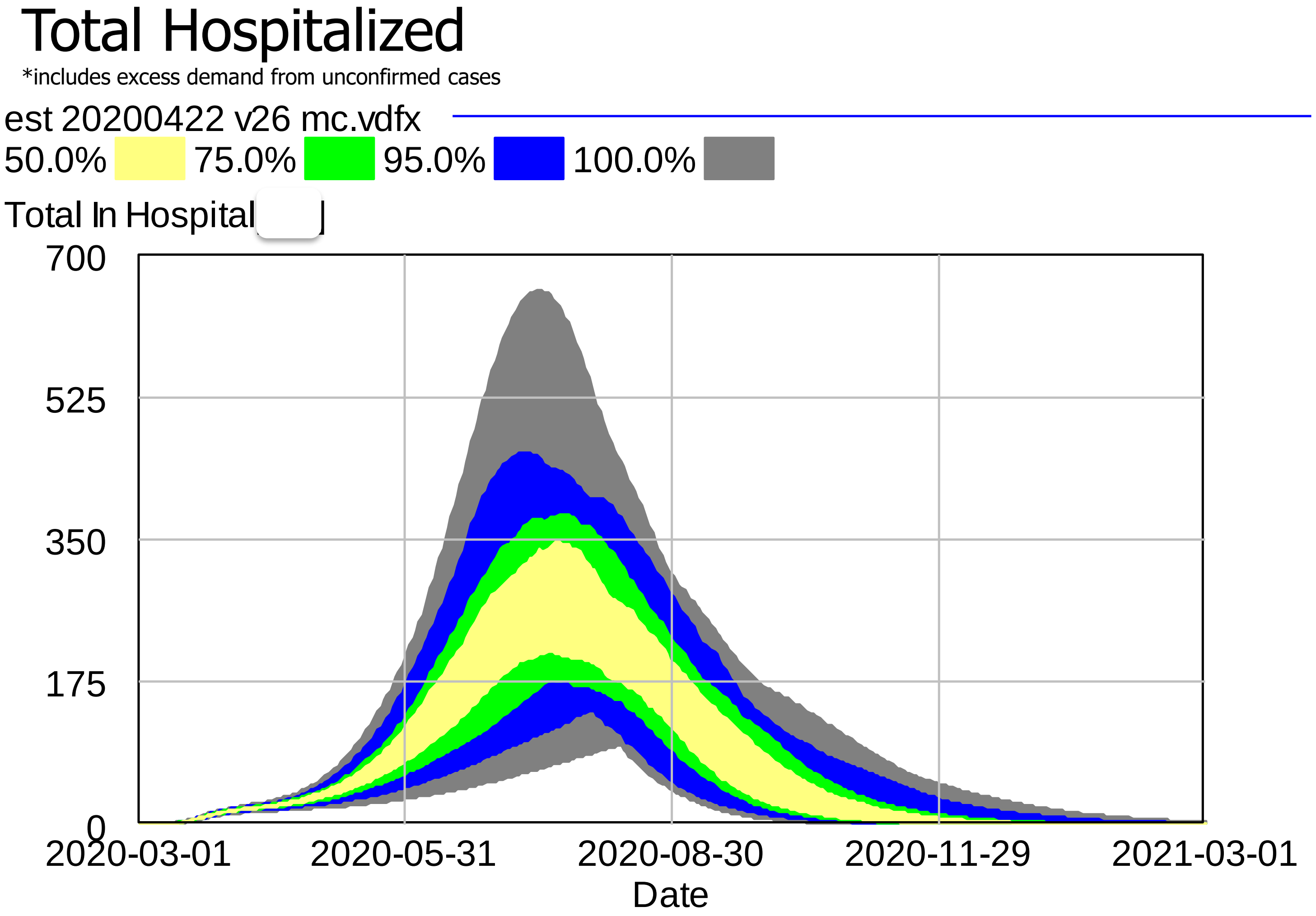
- **A sample of points describing the joint distribution of parameters**
- **Diagnostics**
- **You can then use this to generate sensitivity runs reflecting the sampled parameters**



est 2020 05 23 v37e_MCMC_sample+scenario

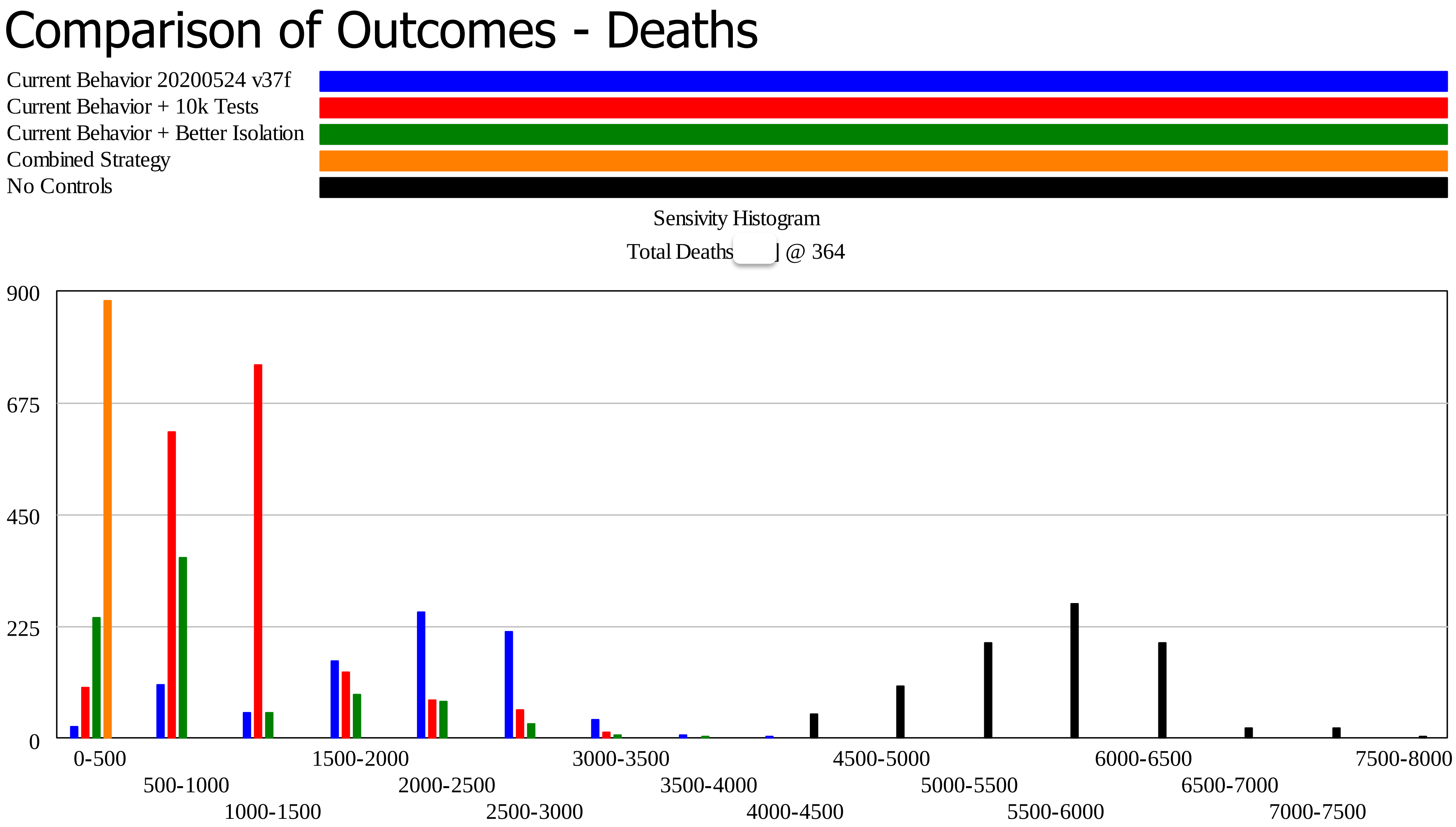
Use the MCMC Sample for Sensitivity Runs

- **What does the parameter distribution imply for the distribution of behavior?**
- **Does the data lie within the confidence/credible interval? (Posterior predictive check in Bayes-speak)**



What policy performs best under combined uncertainties?

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



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**

Selected References & Resources

- **Vensim manuals and sample models**
- **Vensim Data & Calibration workshops (ISDC 2022) https://vensim.com/conference/#using-data-in-vensim**
- **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**