Model Assessment

BEE 6940 Lecture 11

April 10, 2023

Table of Contents

1. Motivation: Nonstationarity
2. Model Assessment through Simulation
3. Graphical Model Checks
4. Human Perception and Graphics
5. Key Takeaways
6. Upcoming Schedule

Motivation: Nonstationarity

Is Our Model Appropriate?

As we've discussed, there are many choices involved in modeling data, *e.g.*:

- Statistical/process model;
- Nonstationarity;
- Residuals;
- Prior distributions

Is Our Model Appropriate?

This means that are a large number of models under consideration.

In general, we are in an $\mathcal{M}$**-open setting**: no model is the "true" data-generating model, so we want to pick a model which performs well enough for the intended purpose.

The contrast to this is $\mathcal{M}$**-closed**, in which one of the models under consideration is the "true" data-generating model, and we would like to recover it.

Motivation: Nonstationarity

Let's think about this from the perspective of whether a dataset is nonstationarity.

Option 1: Treat this as a formal hypothesis test:

- H_0 (null): data is stationary (no trend);
- H_1 (alternative): data is stationary (trend).

Under classical assumptions, can derive Mann-Kendall test (see last class for a reminder) and see if the data is "likely" given the assumption of H_0.

Motivation: Nonstationarity

But:

- This only is tractable in closed form because of classical assumptions (*e.g.* normality) which are unlikely to hold in practice.
- Not very satisfying: we have this "zoo" of statistical tests which apply in highly specific contexts.

Motivation: Nonstationarity

But:

- This only is tractable in closed form because of classical assumptions (*e.g.* normality) which are unlikely to hold in practice.
- Not very satisfying: we have this "zoo" of statistical tests which apply in highly specific contexts.

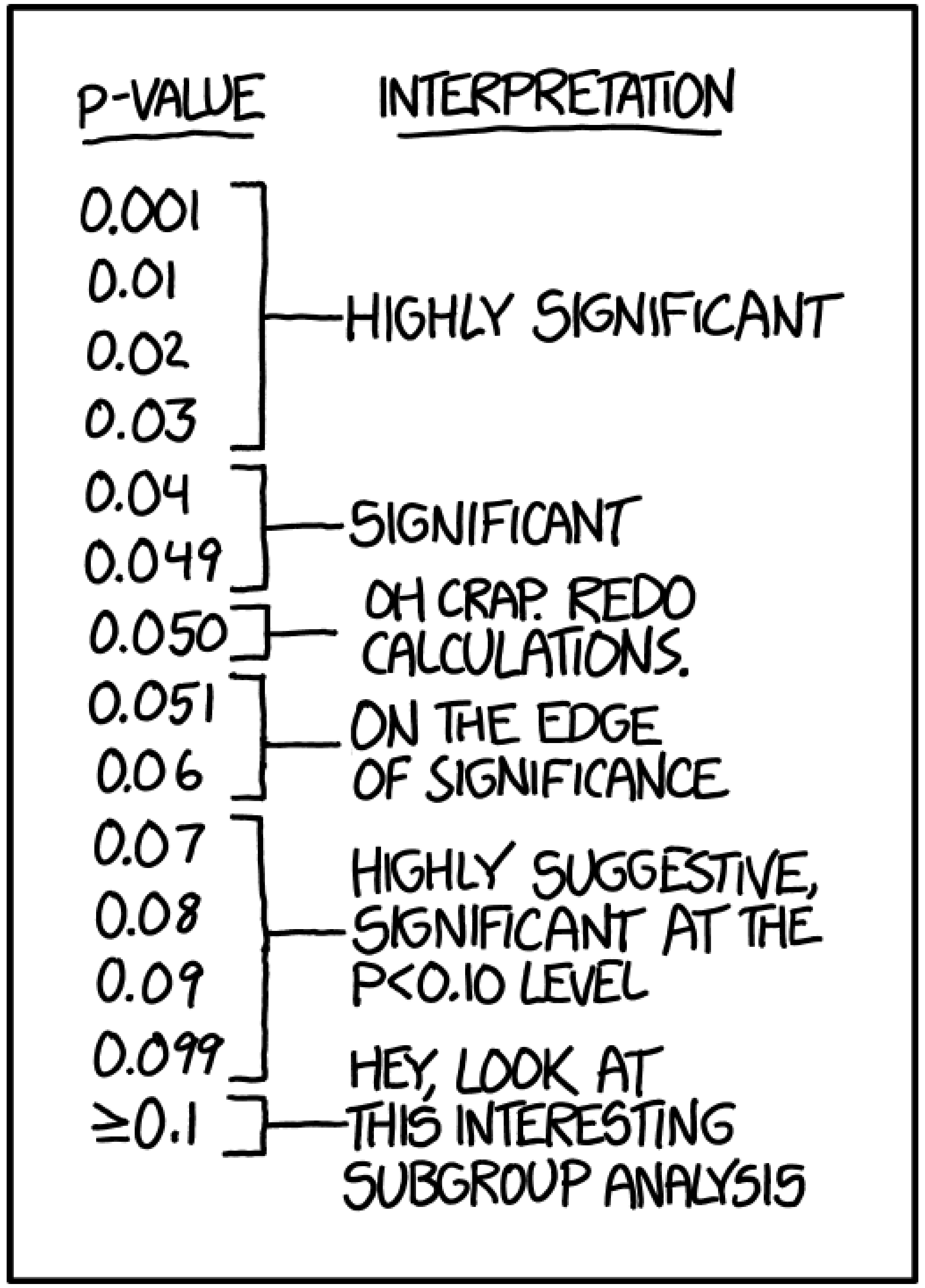
What other approaches are there?

What Is Any Statistical Test Doing?

If we think about what a test like Mann-Kendall is doing:

1. Assume the null hypothesis H_0;
2. *Obtain the sampling distribution of a test statistic* S *which captures the property of interest under* H_0;
3. Calculate the probability of S more extreme than $\hat{S}$ (the p-value).

Aside: P-values



P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Source: XKCD

What Is Any Statistical Test Doing?

If we think about what a test like Mann-Kendall is doing:

1. Assume the null hypothesis H_0;
2. *Obtain the sampling distribution of a test statistic S which captures the property of interest under H_0;*
3. Calculate the probability of S more extreme than $\hat{S}$ (the p-value).

None of this actually requires classical assumptions!

Model Assessment Through Simulation

Simulation for Statistical Tests

Instead, if we have a model which permits simulation (through Monte Carlo or the bootstrap):

1. Calibrate models under different assumptions (*e.g.* stationarity vs. nonstationary based on different covariates);
2. Simulate realizations from those models;
3. Compute the distribution of the relevant statistic S from these realizations;
4. Assess which distribution is most consistent with the observed quantity.

Advantages of Simulation for "Testing"

- More structural freedom (don't need to write down the sampling distribution of S in closed form);
- Don't need to set up a dichotomous "null vs alternative" test;
- Models can reflect more nuanced hypotheses about data generating processes.

Model Assessment Criteria

This raises the question: how do we assess models?

Model Assessment Criteria

This raises the question: how do we assess models?

Generally, through **predictive performance**: how probable is some data (out-of-sample or the calibration dataset)?

Variations on Predictive Distributions

Posterior Predictive Distribution: Consider a new realization y^{rep} simulated from

$$p(y^{\text{rep}}|y) = \int_\theta p(y^{\text{rep}}|\theta)p(\theta|y)d\theta.$$

Variations on Predictive Distributions

Posterior Predictive Distribution: Consider a new realization y^{rep} simulated from

$$p(y^{\text{rep}}|y) = \int_\theta p(y^{\text{rep}}|\theta)p(\theta|y)d\theta.$$

Samples from this distribution can be simulated by:

$$p(\theta|y) \xrightarrow{\hat{\theta}} \mathcal{M}(\hat{\theta}) \rightarrow y^{\text{rep}}$$

Variations on Predictive Distributions

Prior Predictive Distribution: Sample $\theta \sim p(\theta)$ instead:

$$p(y^{\text{rep}}) = \int_\theta p(y^{\text{rep}}|\theta)p(\theta)d\theta.$$

Variations on Predictive Distributions

Prior Predictive Distribution: Sample $\theta \sim p(\theta)$ instead:

$$p(y^{\text{rep}}) = \int_\theta p(y^{\text{rep}} | \theta) p(\theta) d\theta.$$

When $y^{\text{rep}} = y$, this is the same as the *marginal likelihood* $p(y)$, which is the normalizing constant in the denominator of Bayes' Theorem.

Why Consider These Distributions?

Model evaluations are often considered using a point estimate $\hat{\theta}$.

Why is this potentially bad?

Graphical Vs. Quantitative Predictive Checks

"Predictive checks" can come in two flavors:

1. Graphical checks: Visualizations of replicated data or test statistics vs. the original.
2. Quantitative checks: p-values, information criteria/cross-validation.

Graphical Vs. Quantitative Predictive Checks

"Predictive checks" can come in two flavors:

1. Graphical checks: Visualizations of replicated data or test statistics vs. the original.
2. Quantitative checks: p-values, information criteria/cross-validation.

We will focus on graphical checks today, and discuss IC/CV next week.

Graphical Model Checks

What Is A Graphical Model Check?

Goal: Look for problems with replications which might reveal model inadequacy.

Common examples:

- A hindcasts or distributions of test statistics might show over- or under-confidence, or that the simulations don't capture key trends;
- Residual plots (distributions, autocorrelations) might show that the discrepancy or error model was mis-specified.

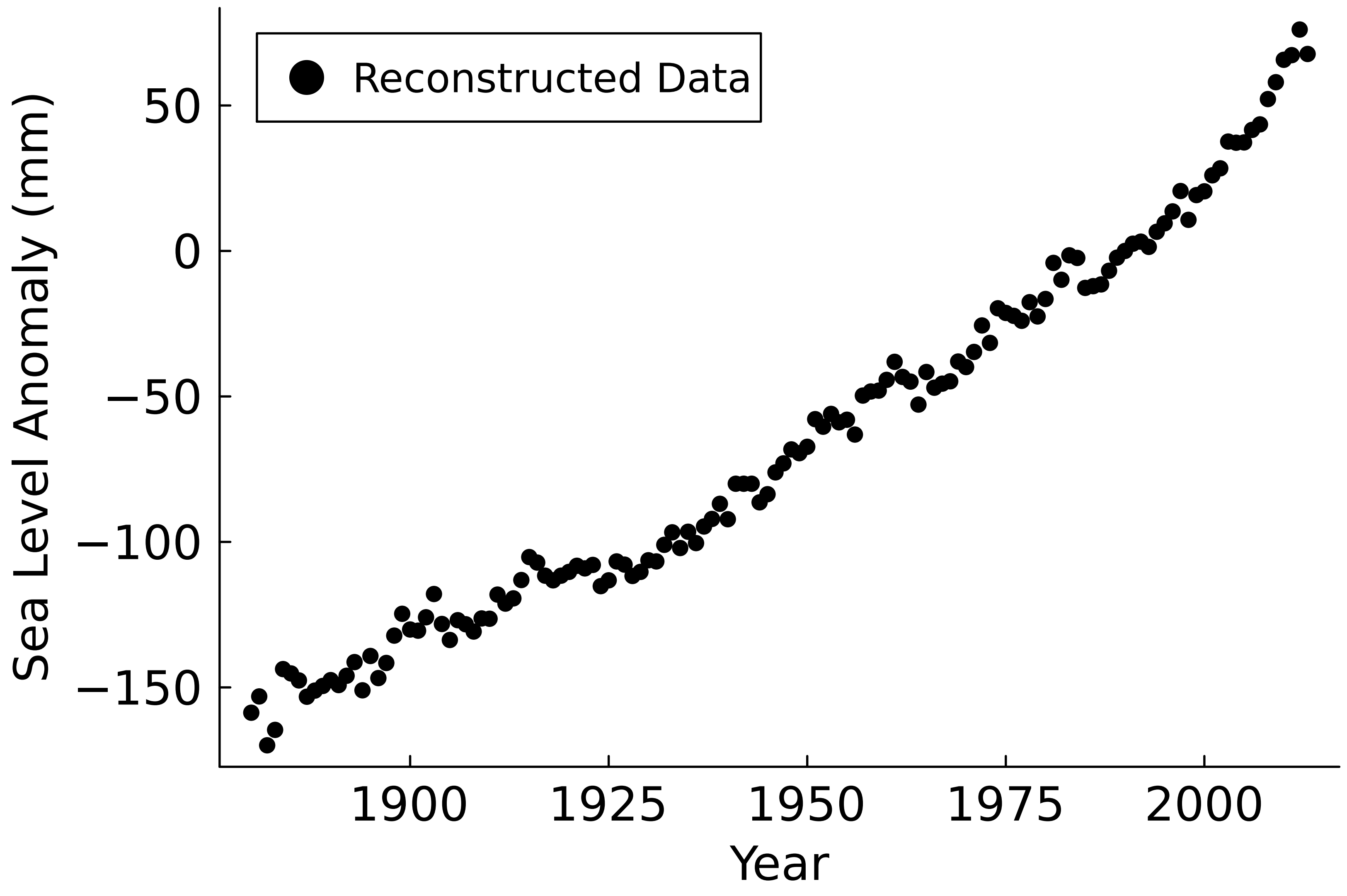
What Is A Graphical Model Check?

Graphical checks are not a substitute for hypothesis-driven model development; they go hand-in-hand.

For example, you might find that your data is less representative of the model simulations. Ask yourself if this makes sense due to internal variability, or if it might be the result of "mis-specification" and could be improved through modeling.

Examples of Graphical Model Checks

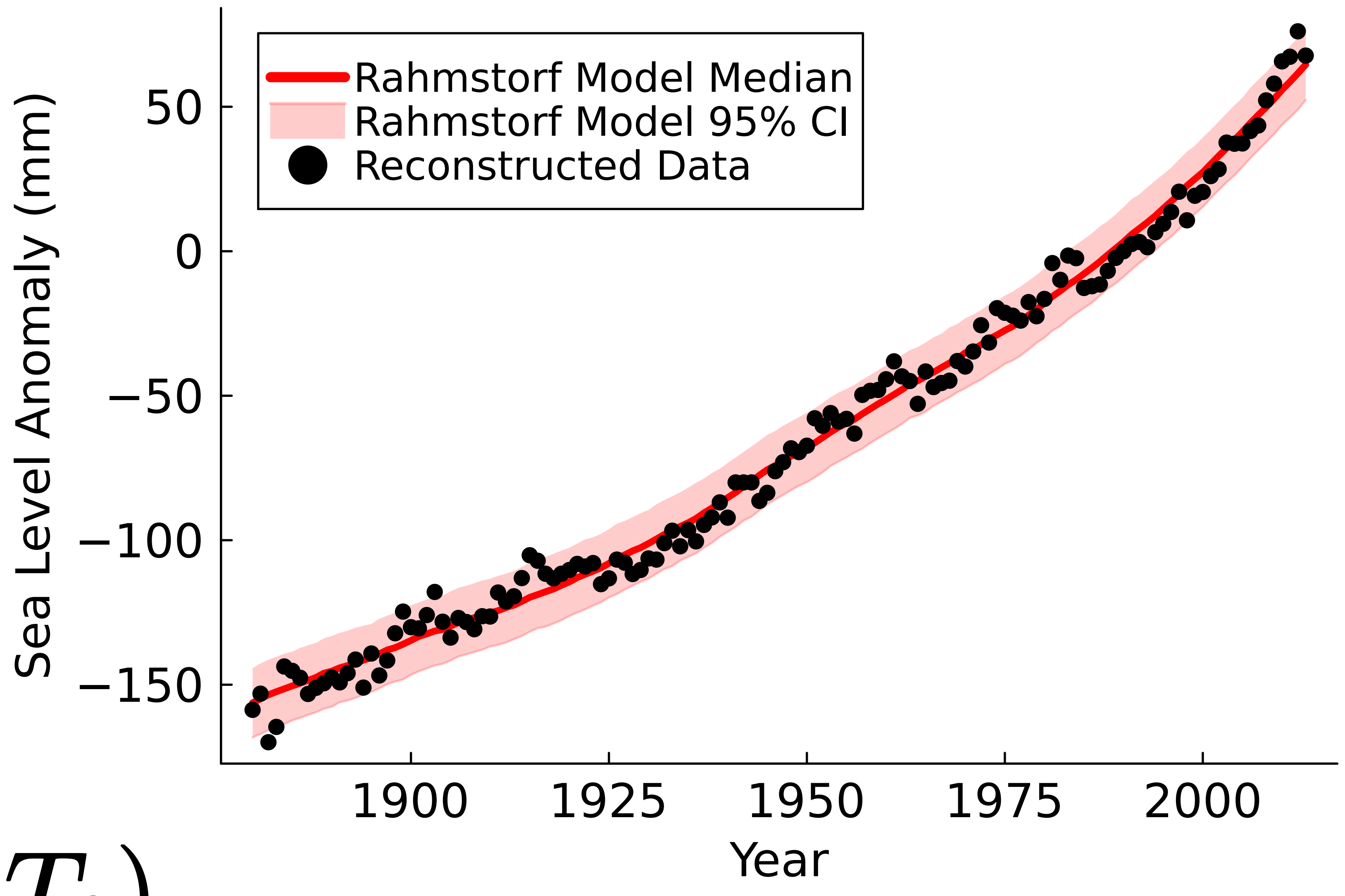
Let's look at the sea-level rise data you've been working with.



Examples of Graphical Model Checks

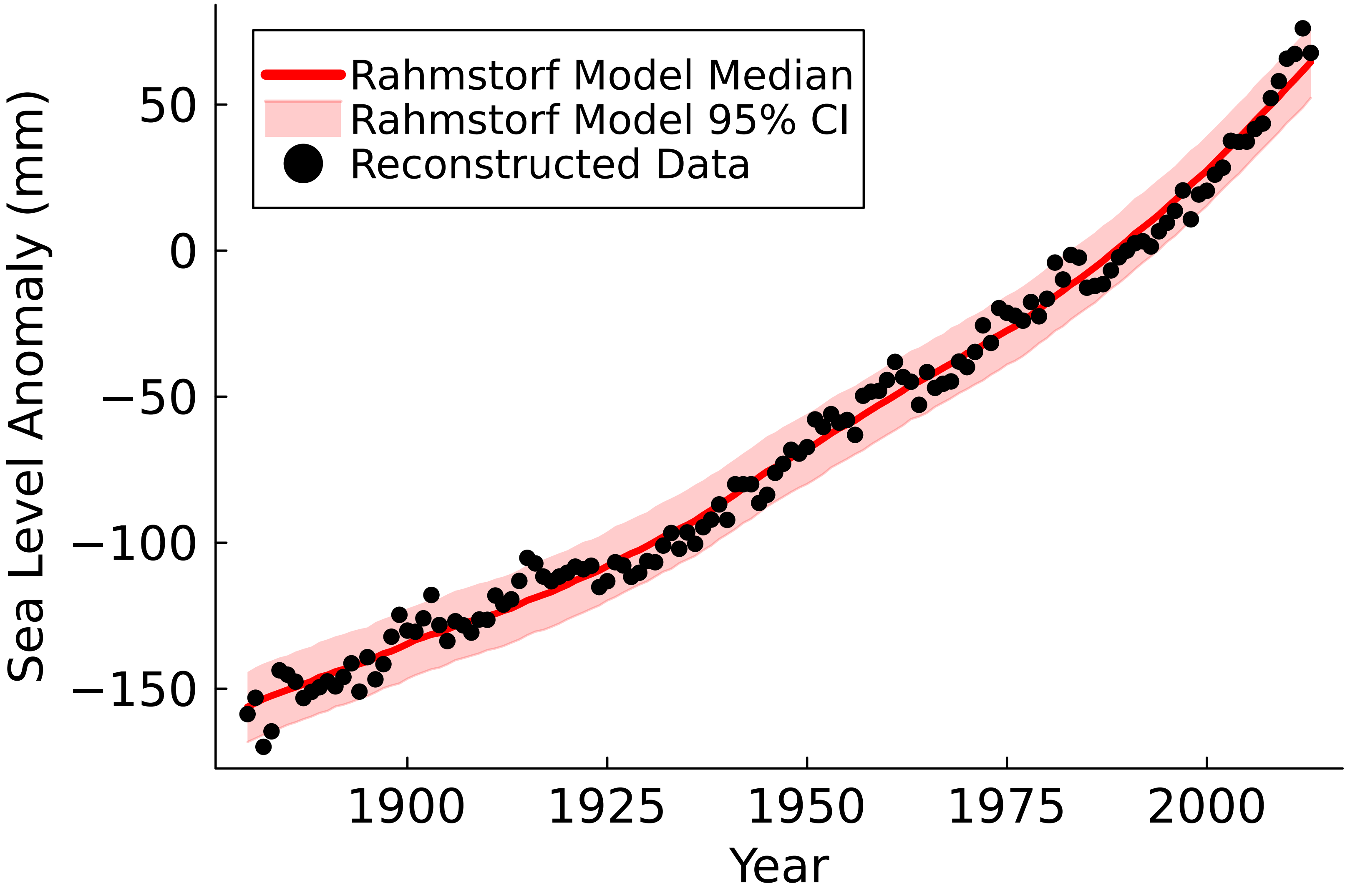
Let's use the Rahmstorf (2007) model with normally-distributed residuals to start:

$$\begin{aligned} y_t &= H(t) + \varepsilon_t, \\ H(t+1) &= H(t) + \alpha(T(t) - T_0) \\ \varepsilon &\sim \text{Normal}(0, \sigma) \end{aligned}$$



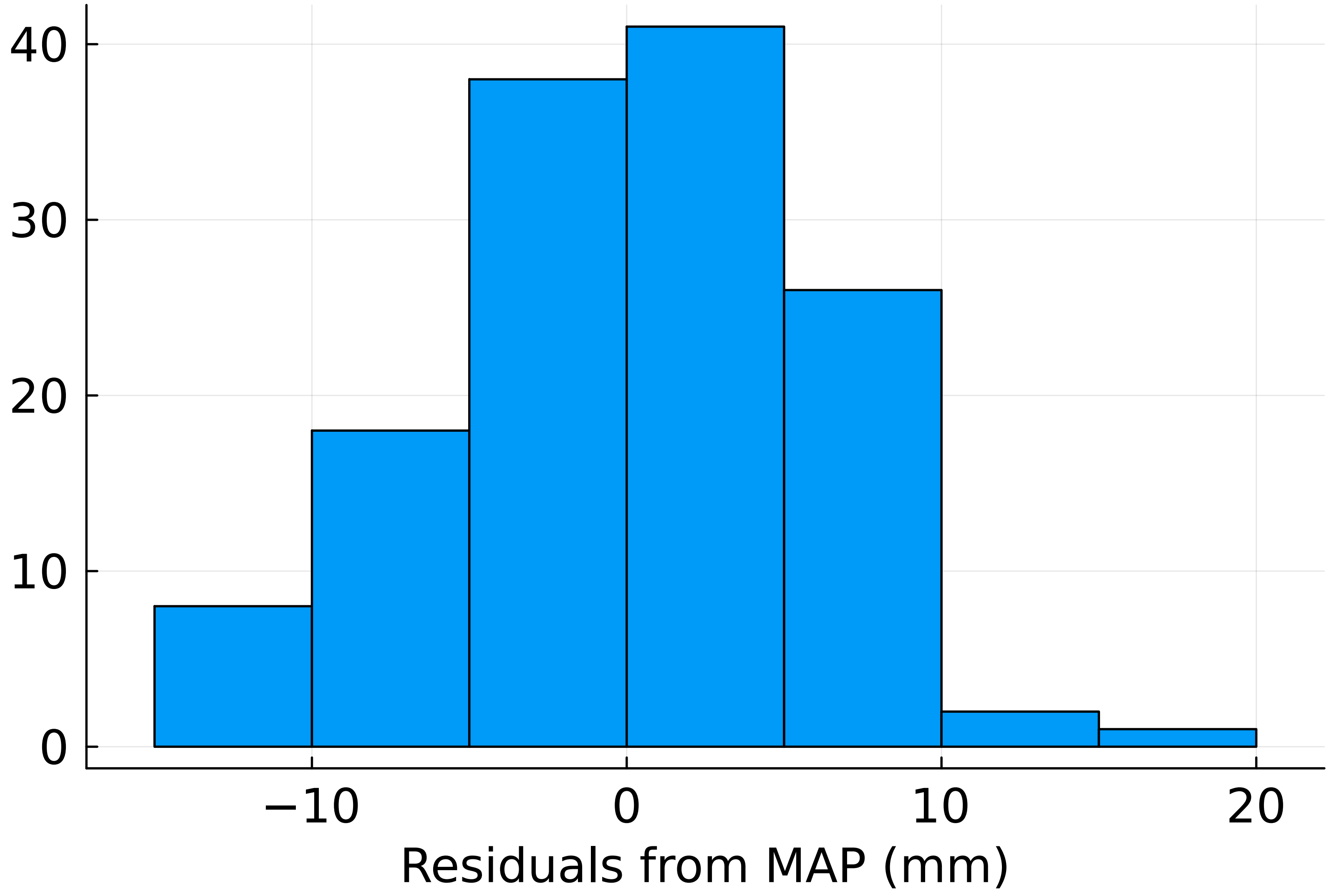
This Is Good, Right?

The surprise index is ~3%, and ideally would be 5%, but that's generally pretty good. But...

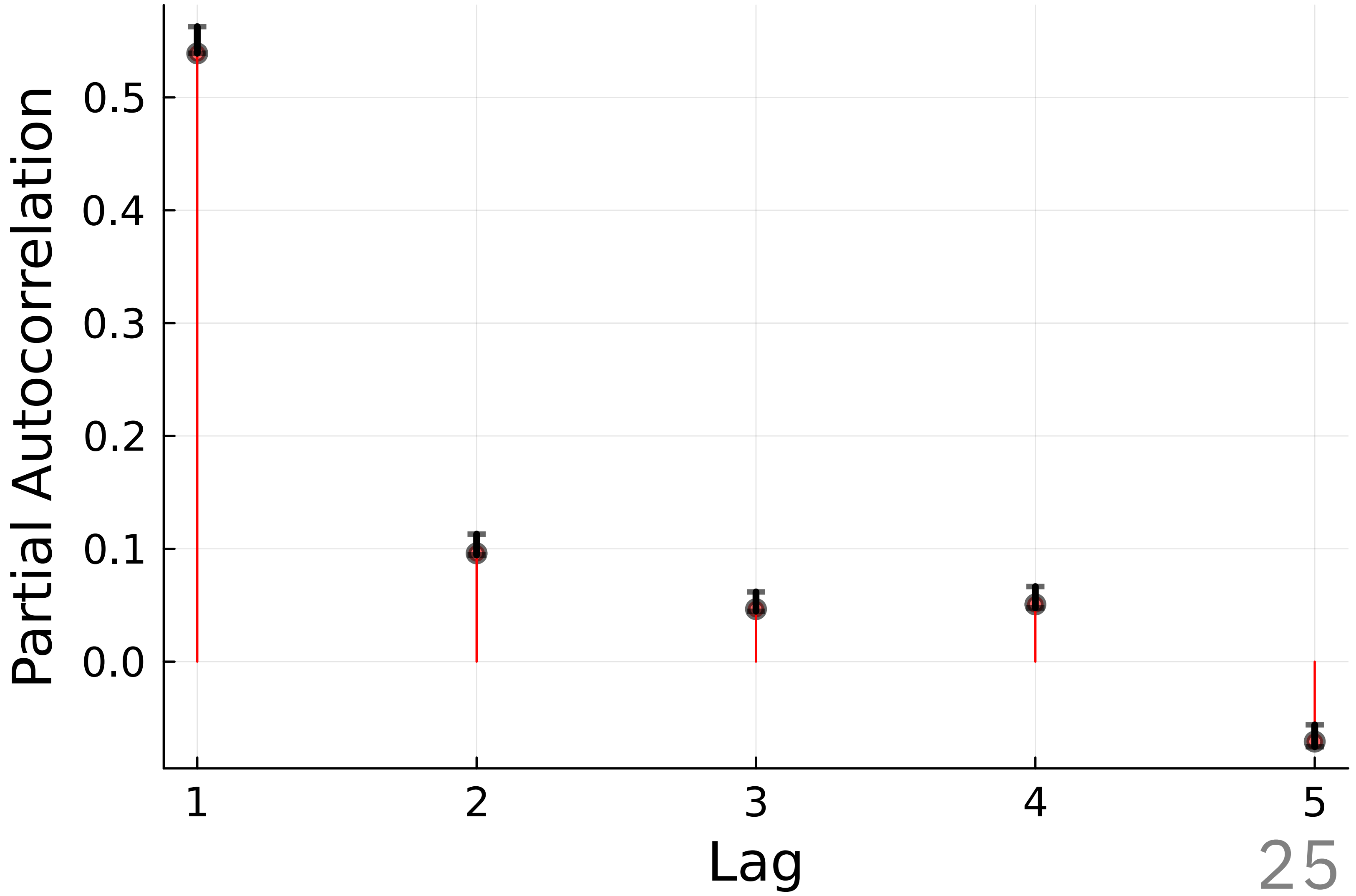


Look at the Residuals...

Residuals relative to the MAP:



Partial Autocorrelation Function:



Result of These Graphical Checks

Based on these few checks, what might we conclude?

Result of These Graphical Checks

Based on these few checks, what might we conclude?

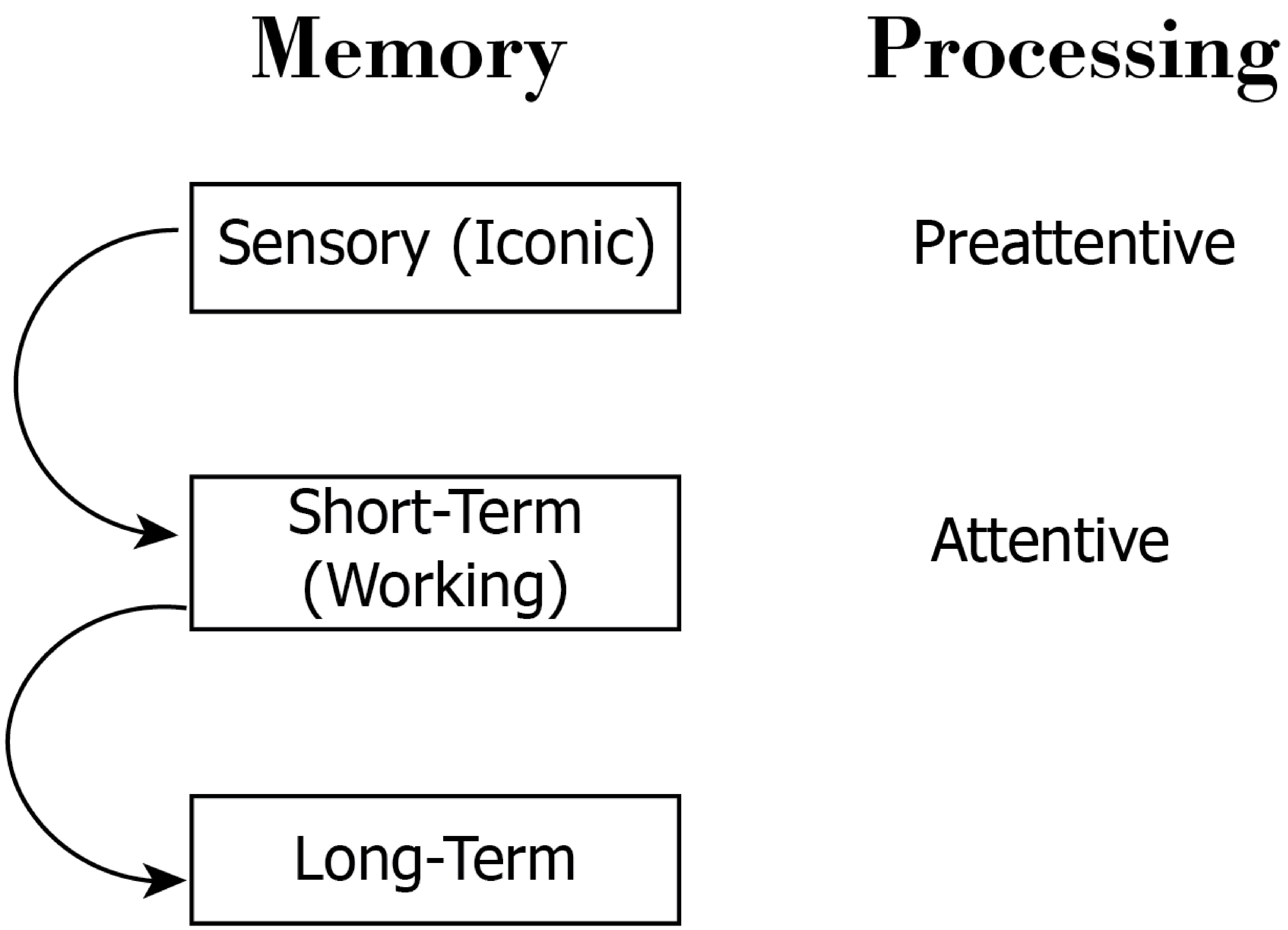
- Priors are possibly slightly restrictive, but this isn't crazy.
- Might want to use a discrepancy structure which accounts for lag-1 autocorrelation.
- Would need to see if the remaining residuals with that discrepancy were normally-distributed, or if some other distribution (*e.g.* with fat tails) might fit better.

Other Checks

What are some other test statistics we could check for this problem?

Human Perception and Graphics

Human Memory Systems



Memory

Processing

Sensory (Iconic)

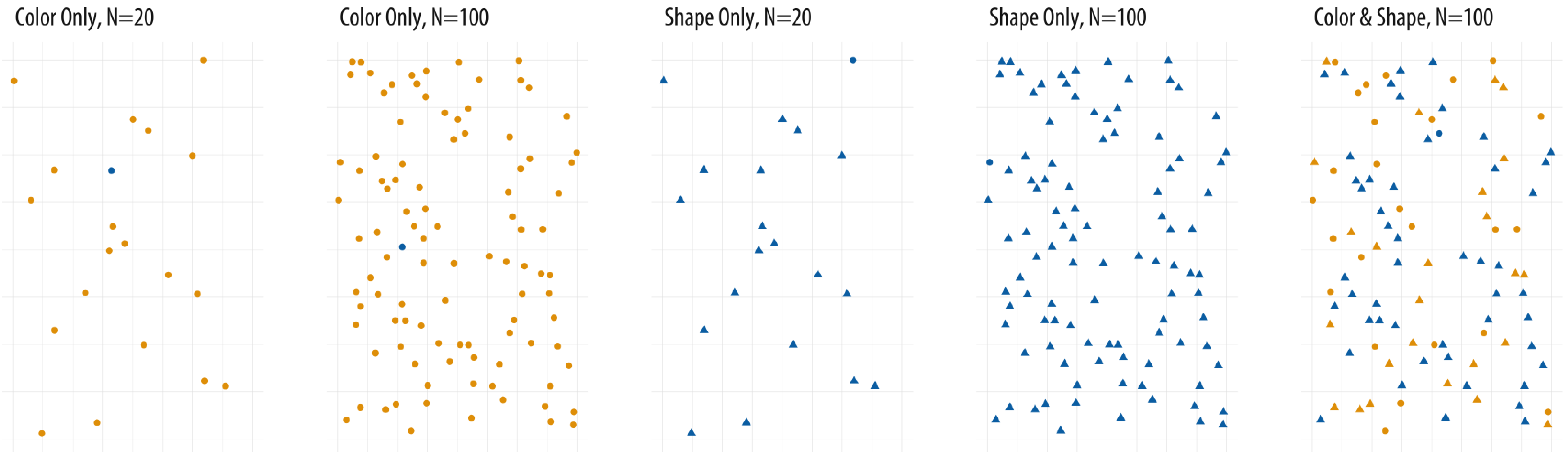
Preattentive

Short-Term (Working)

Attentive

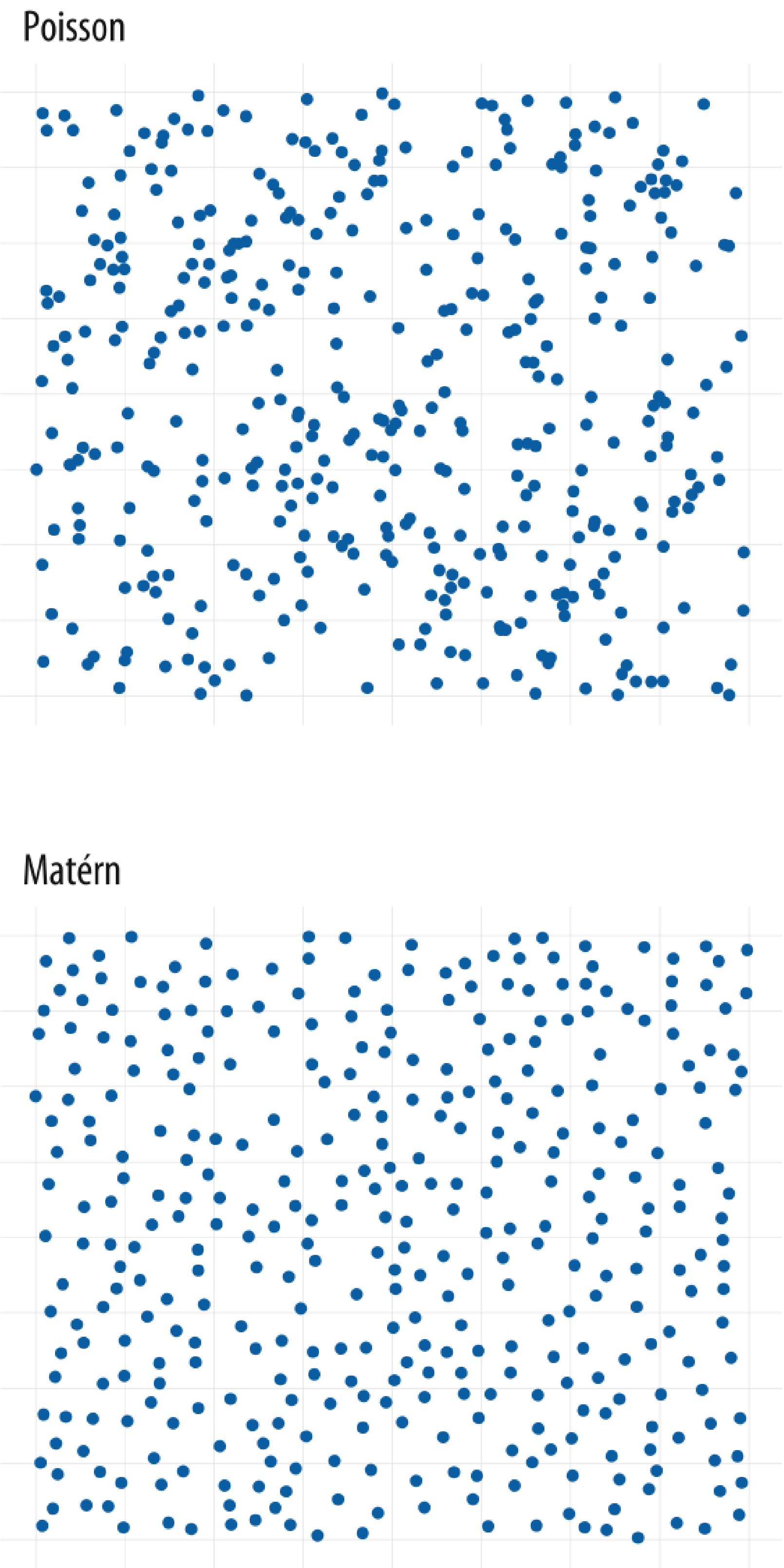
Long-Term

Preattentive "Popout"



Source: Healy (2018)

Gestalt Rules



Which of these two panels is more random?

Source: Healy (2018)

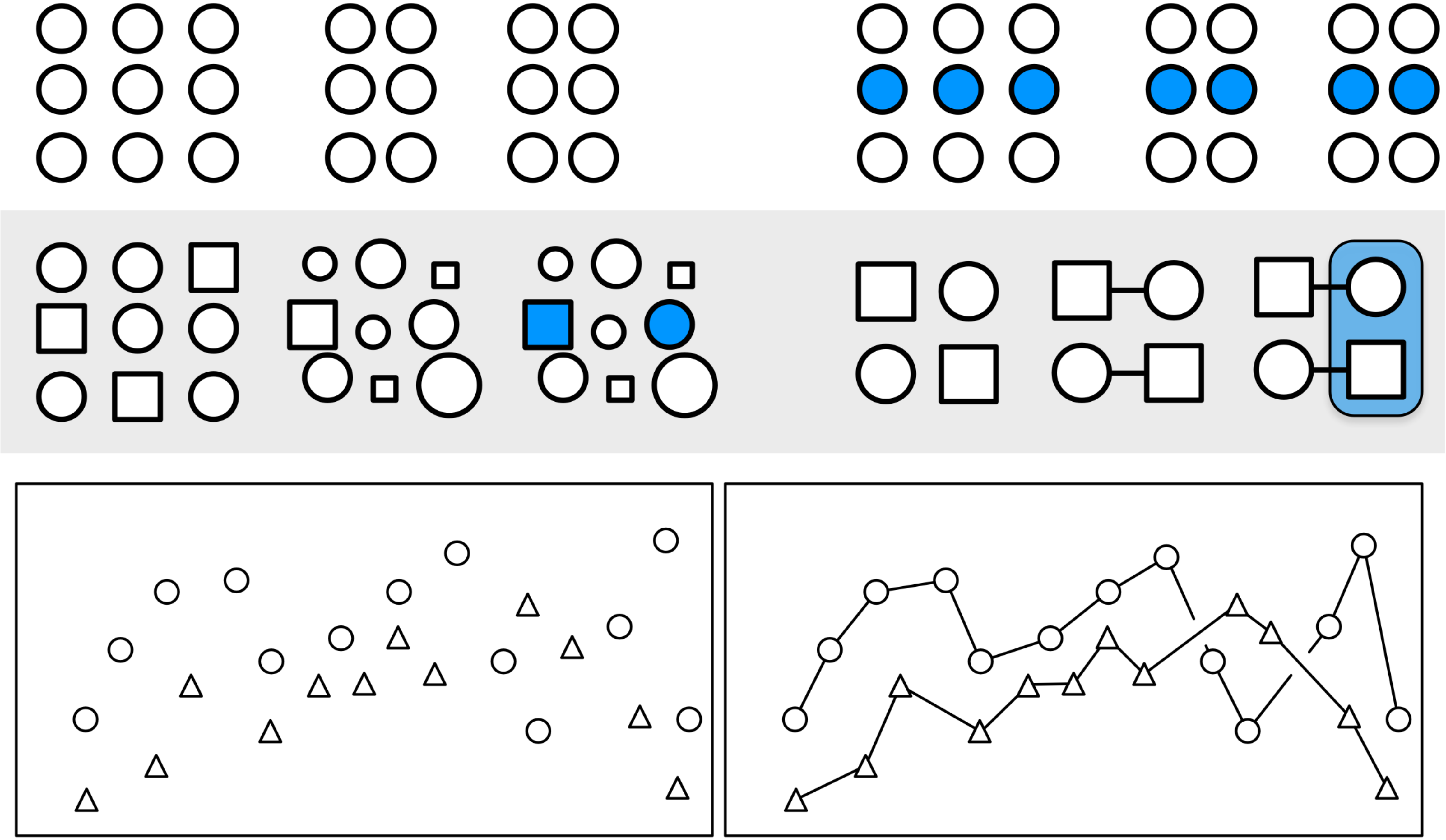
Gestalt Rules

Humans are *really* good at finding structure, even if it doesn't exist (this is why you shouldn't just stare at data and draw conclusions!)

We follow certain rules which allow us to draw inferences from incomplete or sparse visual information. These are called "gestalt rules".

The details aren't critical (but are interesting!), but in general, we try to *group*, *classify*, and *connect*.

Gestalt Rules



Source: Healy (2018)

Implications for Graphical Checks

It's easy to draw misleading conclusions from graphics due to these effects.

To help reduce this risk:

- Look at and provide a variety of visualizations of uncertainty.
- Connect points only when in-between values have meaning or the scatterplot is hard to follow due to the number of points.
- Make key features "pop" with pre-attentive cues to avoid "searching" for meaning.

Key Takeaways

Key Takeaways

- Predictive performance is a common criterion for model evaluation.
- Common statistical tests can be obtained as special cases of simulations from predictive distributions.
- Can consider *prior* or *posterior* predictive distributions.
- Graphical checks can be used to assess fit of replications to assumptions and/or data.
- Be careful about biases and heuristics in perception and visualization to not mislead yourself or others.

Upcoming Schedule

Upcoming Schedule

Wednesday: Discussion of Oreskes et al (1994).

Next Monday: Model Selection, Information Criteria, and Cross-Validation