SAND2017-6611C

Exceptional service in the national interest

Sandia National Laboratories

Using Data-Driven Uncertainty Quantification to Support Decision Making

Matt Peterson

Joint work with:
Charlie Vollmer, David Stracuzzi, and Max Chen

POC: Matt Peterson
mgpeter@sandia.gov

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

UQ for Decision Making

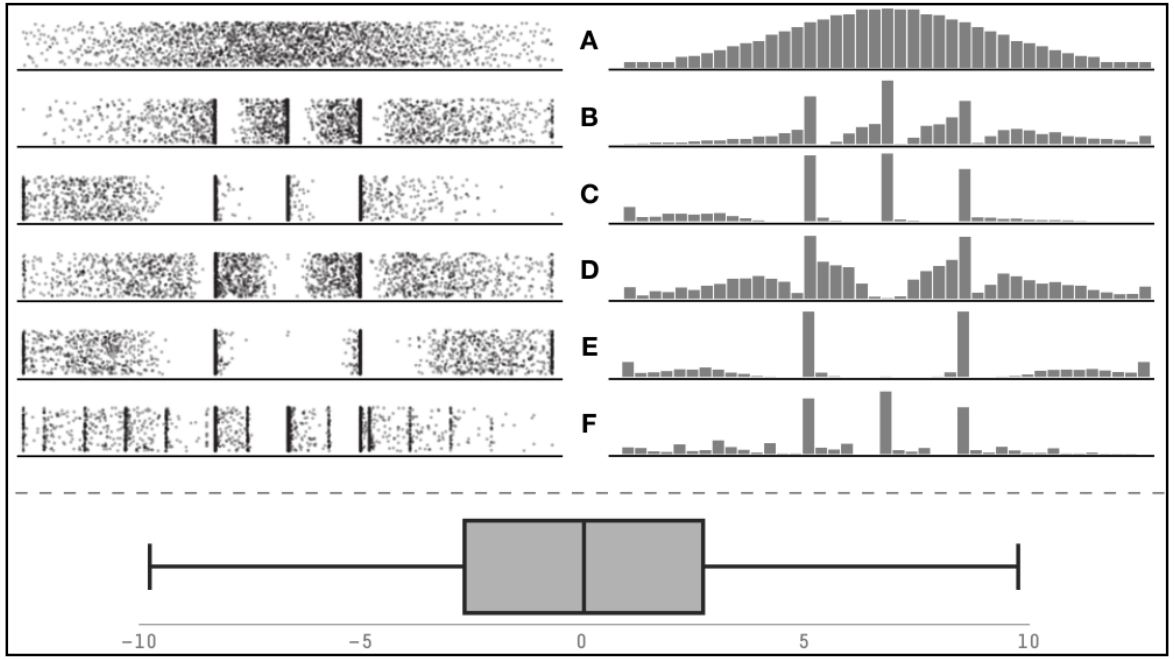
Sandia National Laboratories

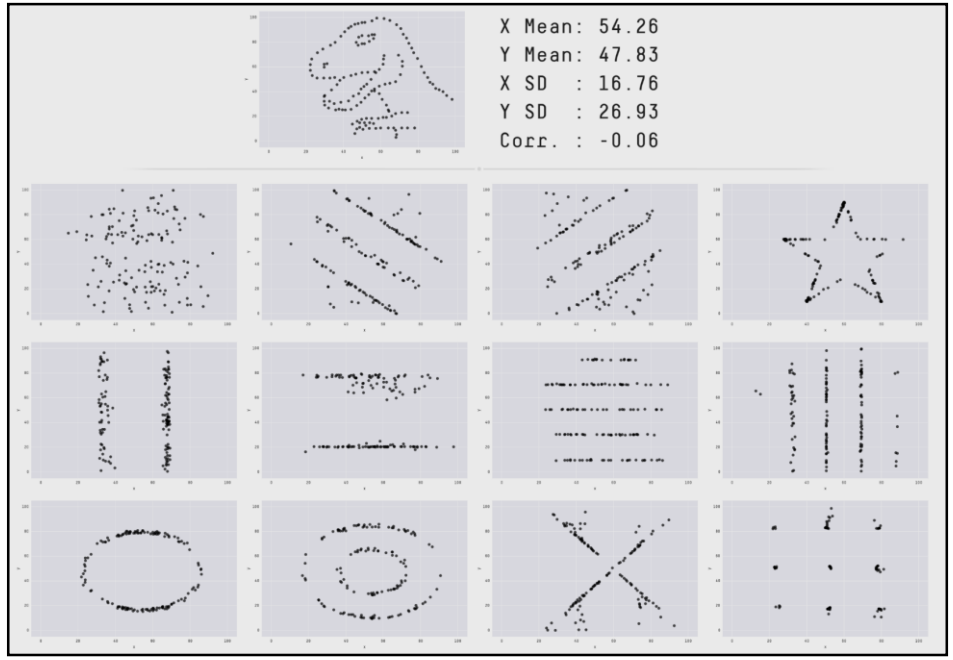
Problem

- Analysts have to make crucial decisions from large datasets
  - Sometimes in real time

Solution

- Simplify the data into a quick and easy way for human understanding
  - Usually done through means, stand deviations, quartiles, etc.





Images curtesy of https://www.autodeskresearch.com/publications/samestats

Example: Seismic Onset Detection

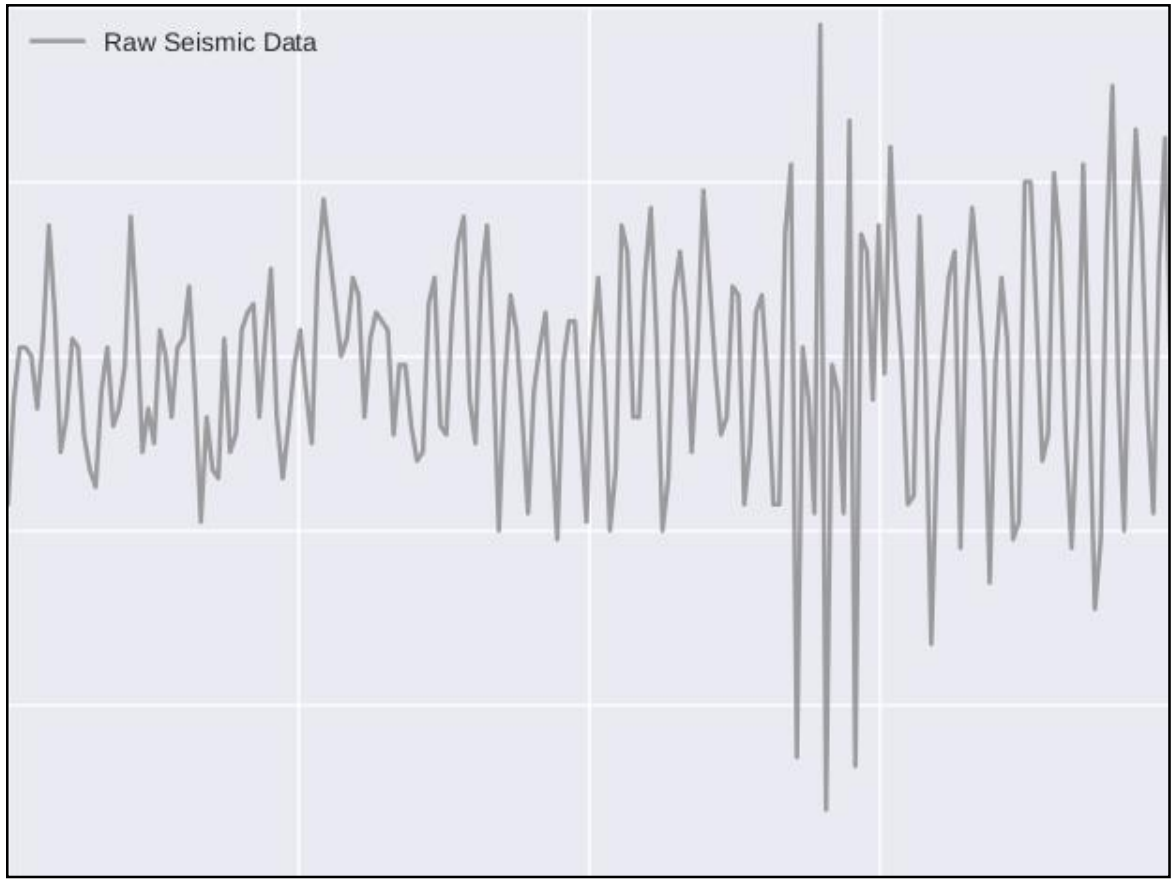
Sandia National Laboratories

Given

- Waveform data containing both noise and seismic signals

Produce

- Signal onset/arrival time
- Precision is critical to downstream processing

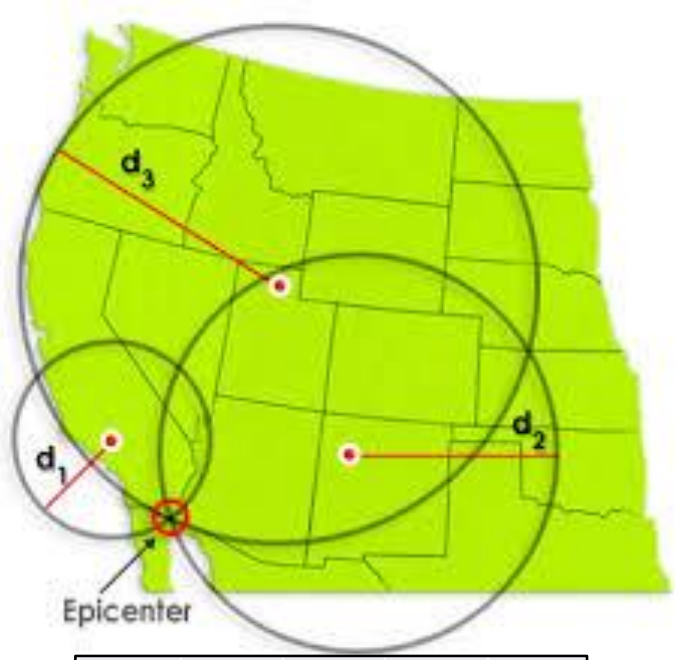


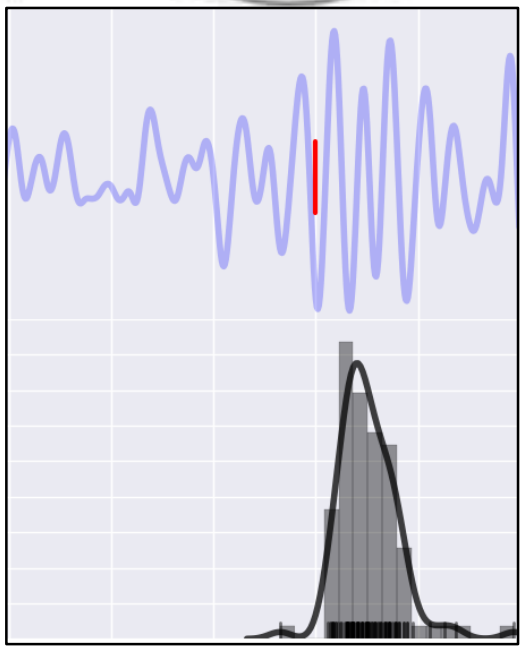
Note:

Ground truth does not exist or is hard/expensive to get

Impact on Downstream Analyses

- Several analyses depend on onset time:
  - Location (hypocenter)
  - Event type (natural or man-made) & size
  - Subsurface tomography
  - Earth model
- Rely more on current data, less on historical data and modeling assumptions
  - Relative reliability of data points and sensors
  - Possibly improve ability to sense smaller, less obvious signals





Analysis Approach

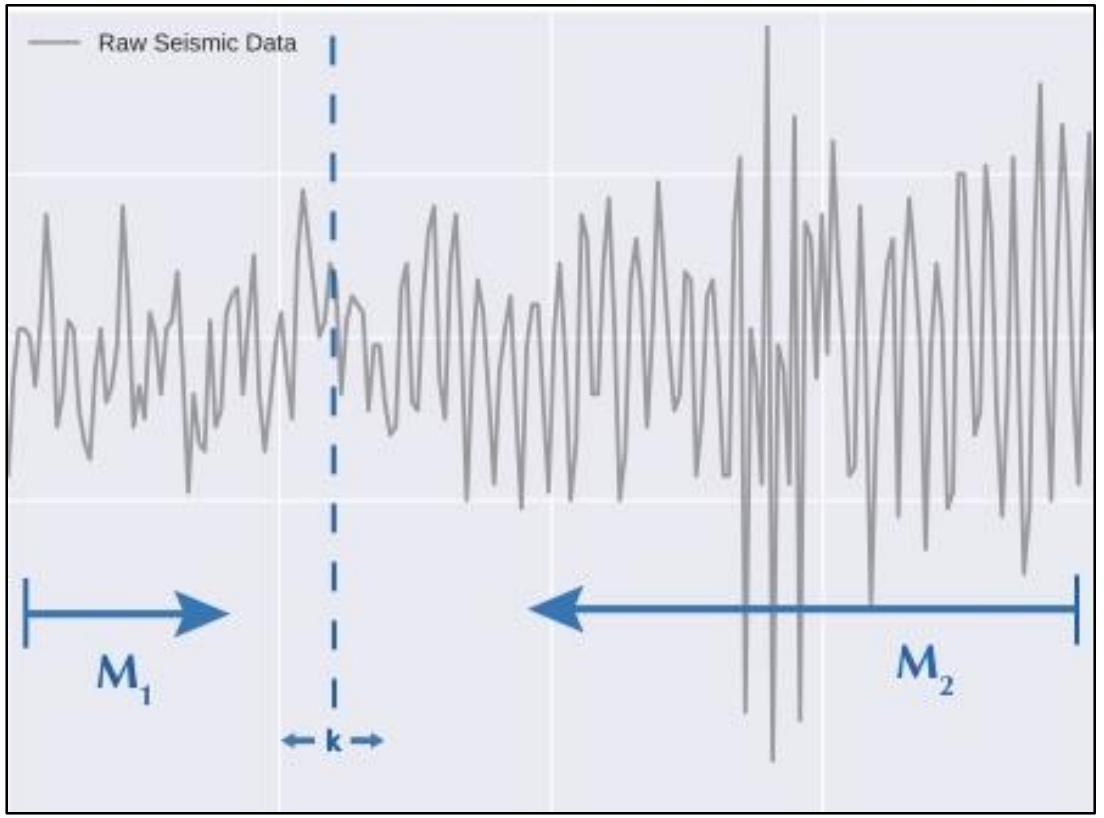
Sandia National Laboratories

Model

- $\mathcal{M}_1$, for the *noise* left of k, is Gaussian: $Y_t \sim N(0, \sigma_n^2)$
- $\mathcal{M}_2$, for the *signal* right of *k*, is ARMA(*p*,*q*): $x_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i x_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$

Basic Approach

- Model the noise (M1) and the signal plus noise (M2) separately
- Optimize model parameters θ, ϕ via maximum likelihood
- Akaike information criterion (AIC) to select transition point *k*
- Point at which two models meet is the "best guess" signal onset



Log-Likelihood

- $\mathcal{M}_1$, for the *noise* left of k, is Gaussian: $Y_t \sim N(0, \sigma_n^2)$
- $\mathcal{M}_2$, for the *signal* right of *k*, is ARMA(*p*,*q*): $x_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i x_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$

$$l(\mathcal{M}) = l(\mathcal{M}_1 | Y_1, \dots, Y_k) + l(\mathcal{M}_2 | Y_{k+1}, \dots, Y_T)$$

$$= -\frac{k}{2}\ln(2\pi) - \frac{k}{2}\ln(\sigma^2) - \frac{1}{2\sigma_n^2}\sum_{t=1}^{k} y_t^2 - \frac{T-k-p}{2}\ln(2\pi) - \frac{T-k-p}{2}\ln(\sigma_s^2) - \frac{1}{2\sigma_s^2}\sum_{t=k+p+1}^{T} \varepsilon_t^2$$

where,

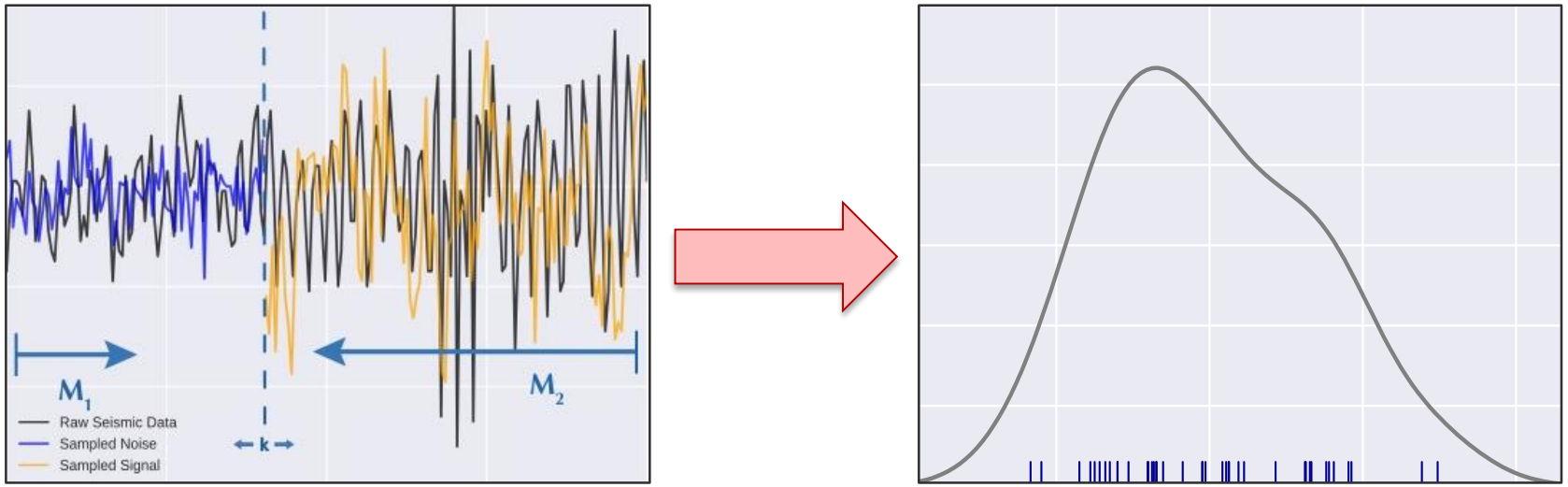
$$\varepsilon_t = Y_t - c - \sum_{i=1}^{p} \phi_i Y_{t-i} - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

for $t = p+1, p+2, \dots, T$

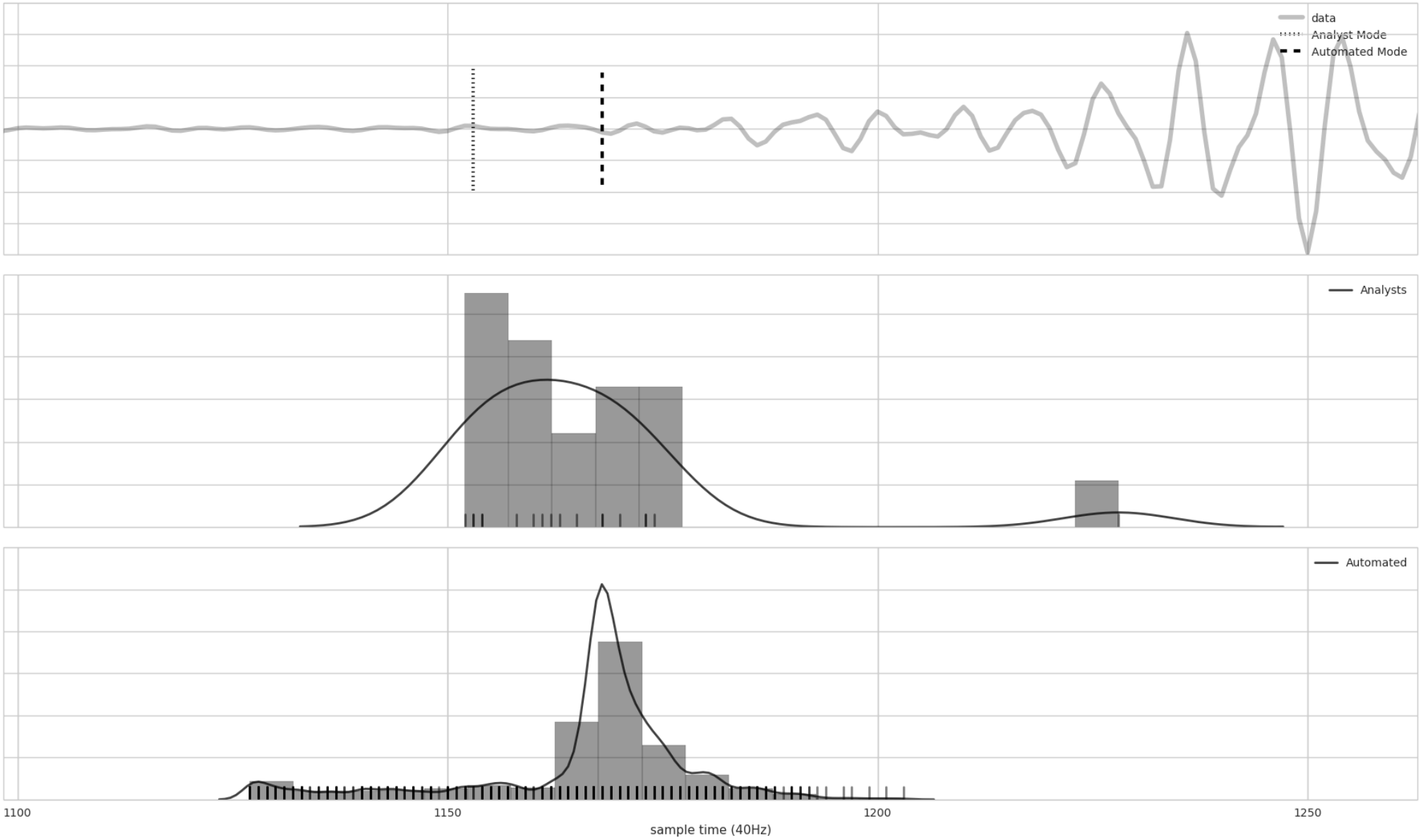
Uncertainty Analysis

Parametric Bootstrap

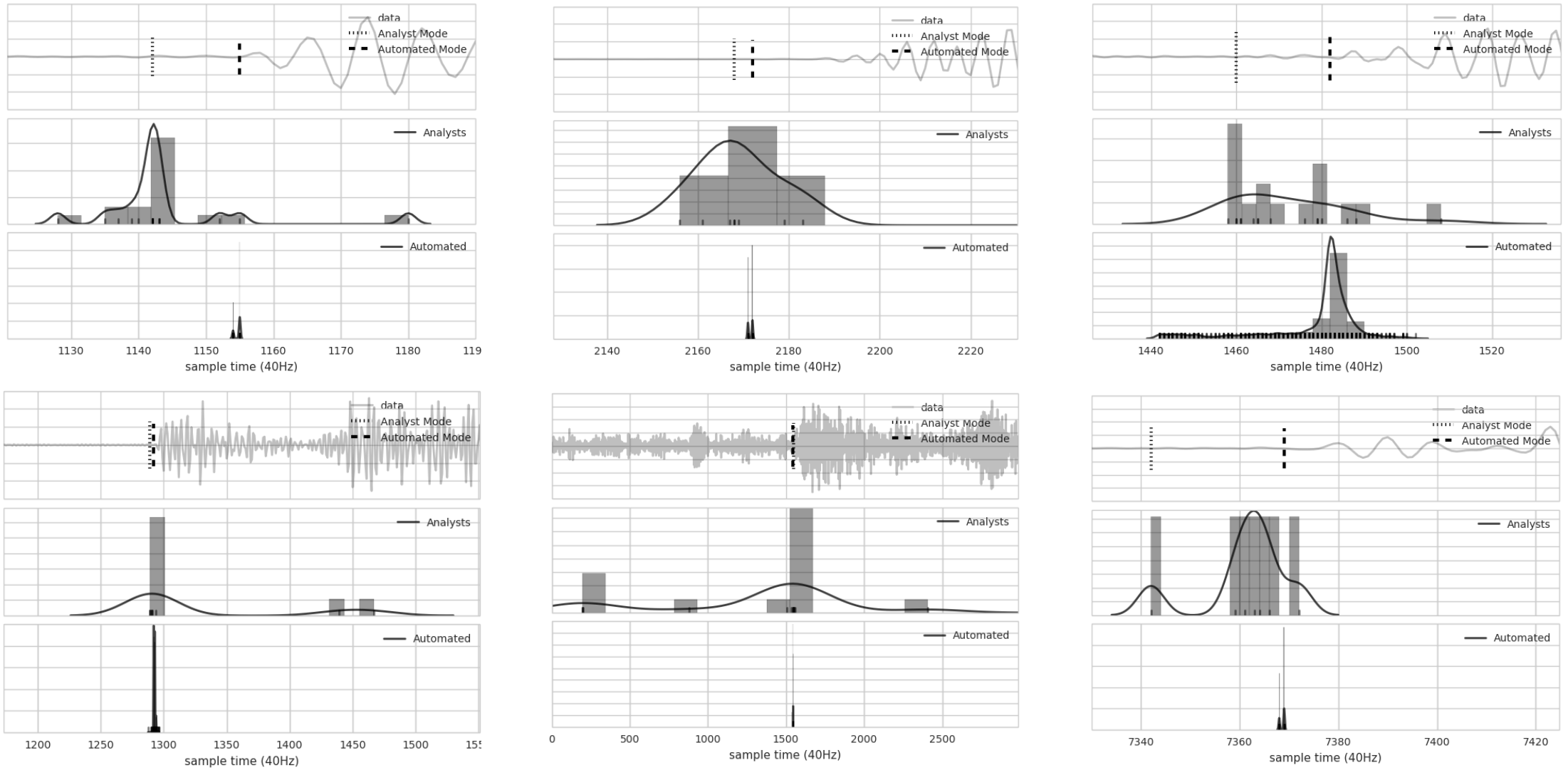
- Use AIC to select *p,q* for $\mathcal{M}_1$ & $\mathcal{M}_2$
- Fit model parameters for each, optimizing *k*
- Sample $\mathcal{M}_1$ & $\mathcal{M}_2$ to create new waveforms
- Fit new models to each sample and record *k*
- Compute the sampling distribution of the estimate k



Results: Analyst vs Automated

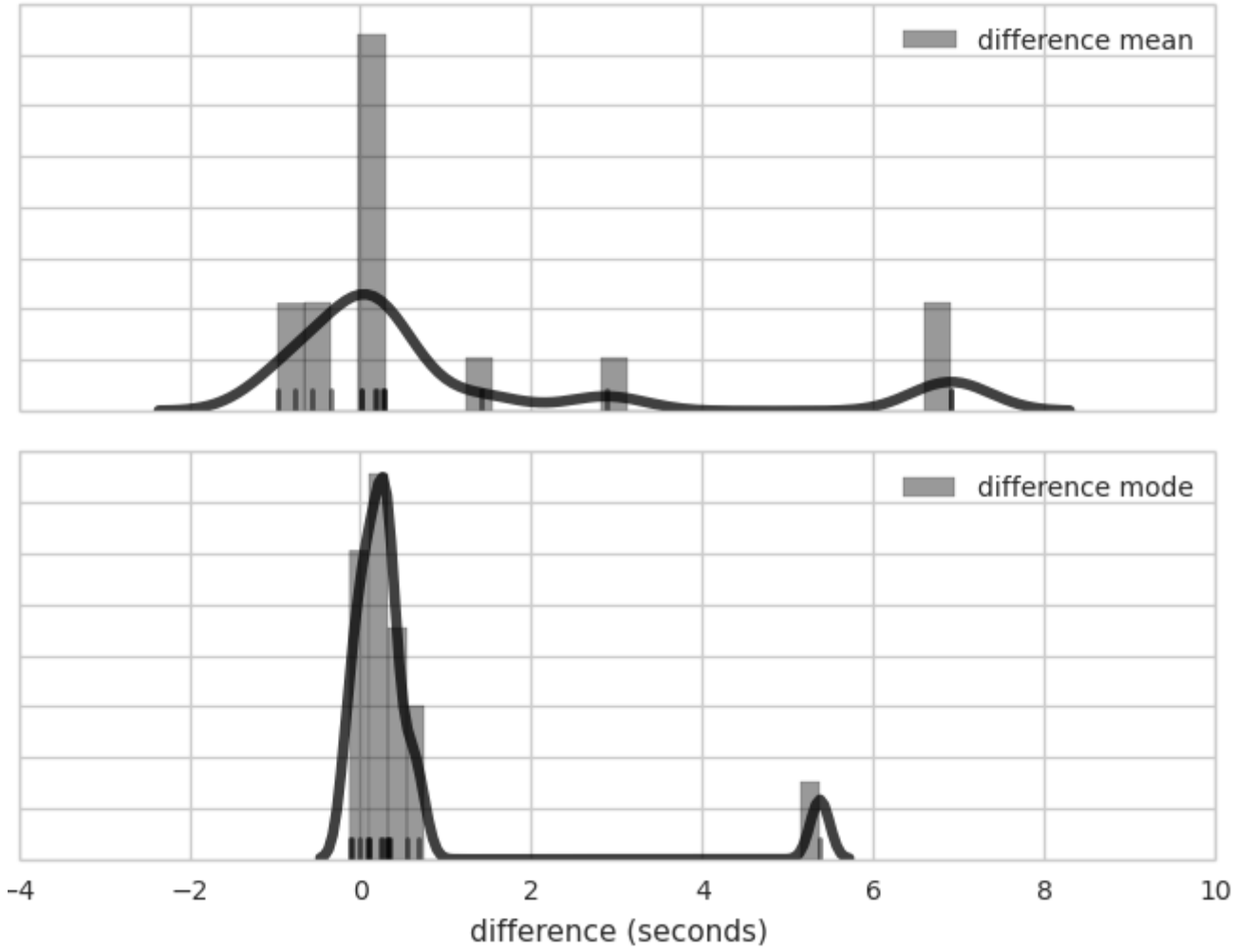


Results: Analyst vs Automated



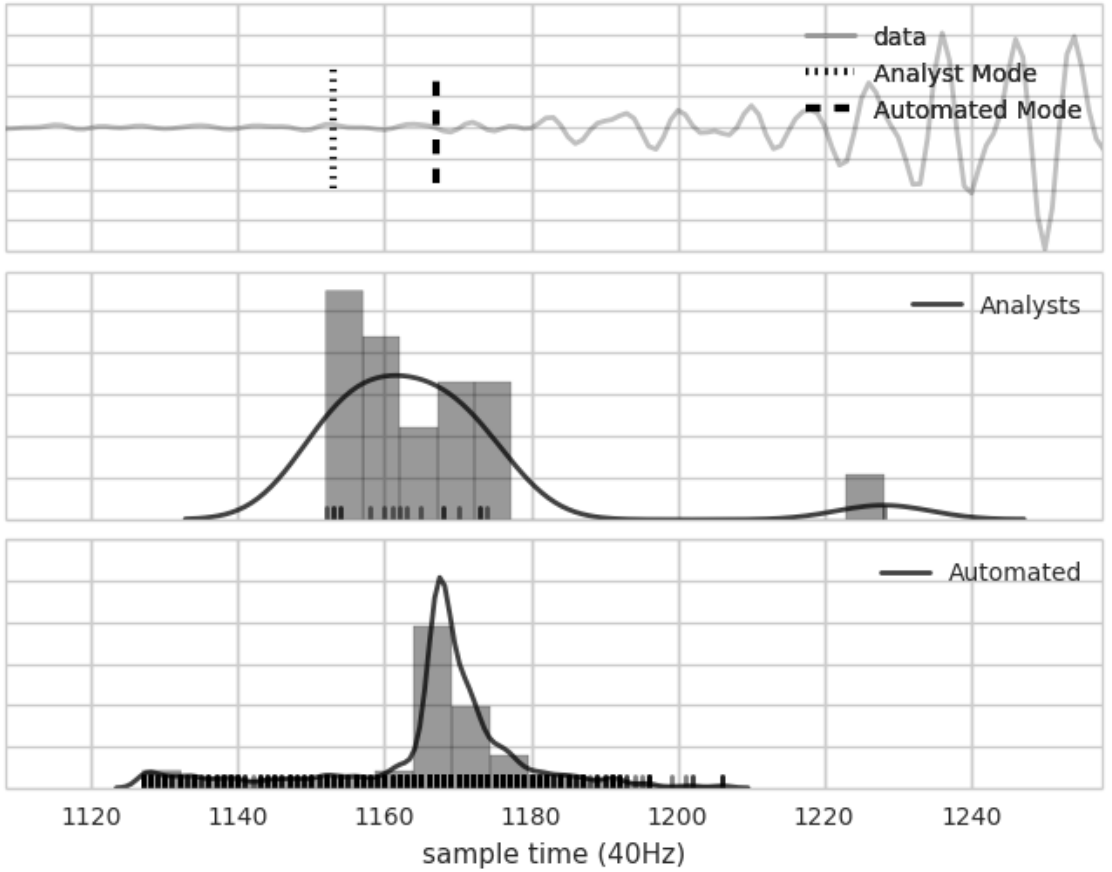
Results: Analyst vs Automated

- 15 waveforms
  - 10-18 analyst picks
  - 1,000 automated picks
- The difference in the means is much more spread out
  - The difference in the modes are tightly packed

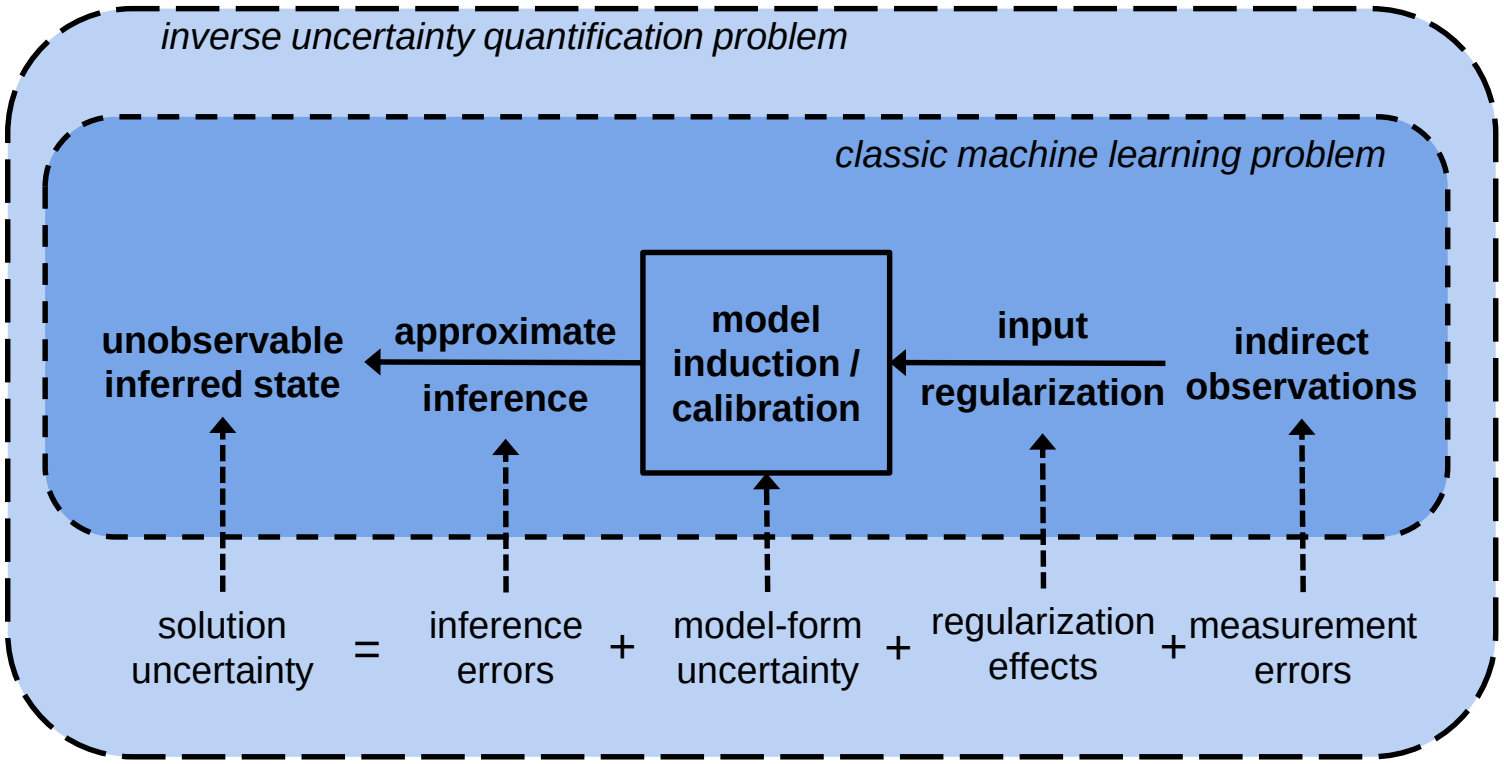


Impact on Decision Making

- UQ provides added information
  - Statistical evidence for the analysts that picked early
  - Alternate (statistical) hypotheses
- Modeling approach reduces assumptions
  - Data specific
  - Not predicated on prior calibration or SNR formulas and cut-offs



Uncertainty Quantification for Statistical Models



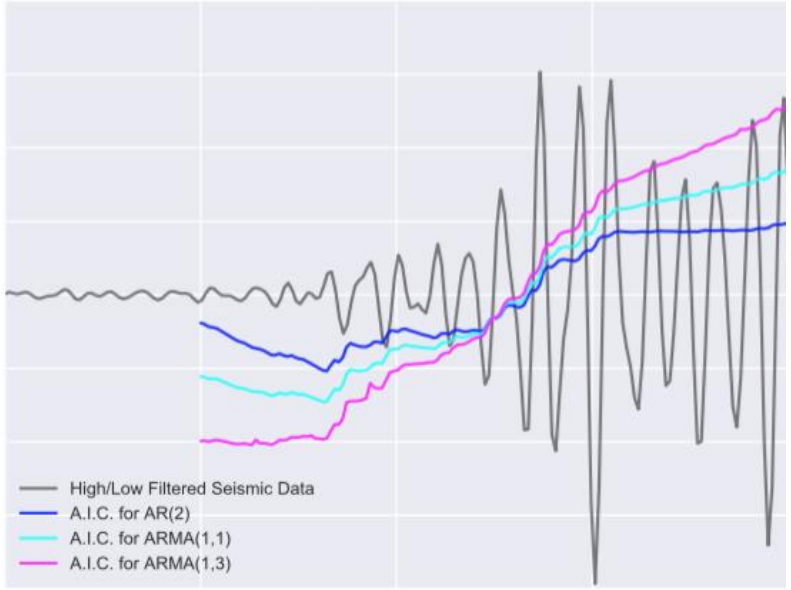
Key Question: How much do sensor observations really tell us about the world?

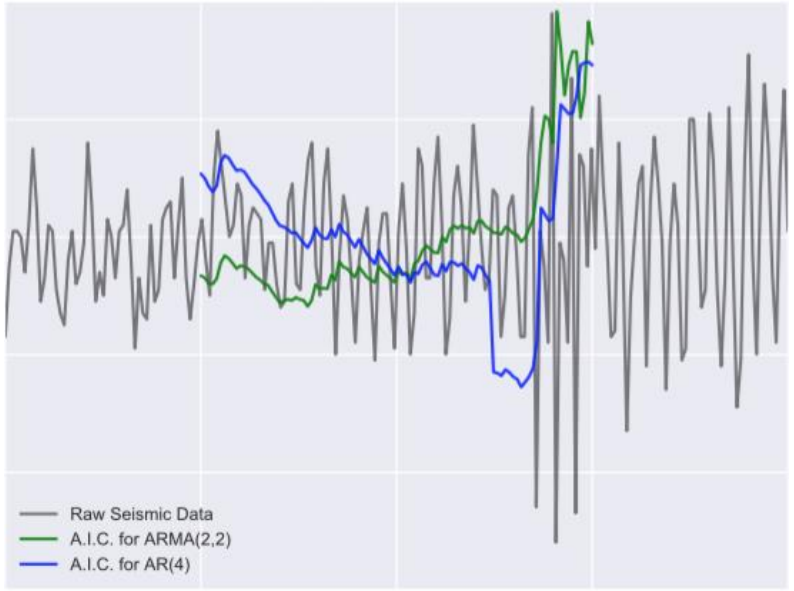
- *Stats / ML research communities do not typically frame questions this way.*
- *Most work focuses on building a better statistical model*
- *"What" the model says emphasized over "how well"*

Related Issues

- Model form uncertainty is tricky
- Depends on definition of “best fit”
  - By what metric?
  - What are we trying to fit?
- Model complexity
- Sampling
- Computational complexity

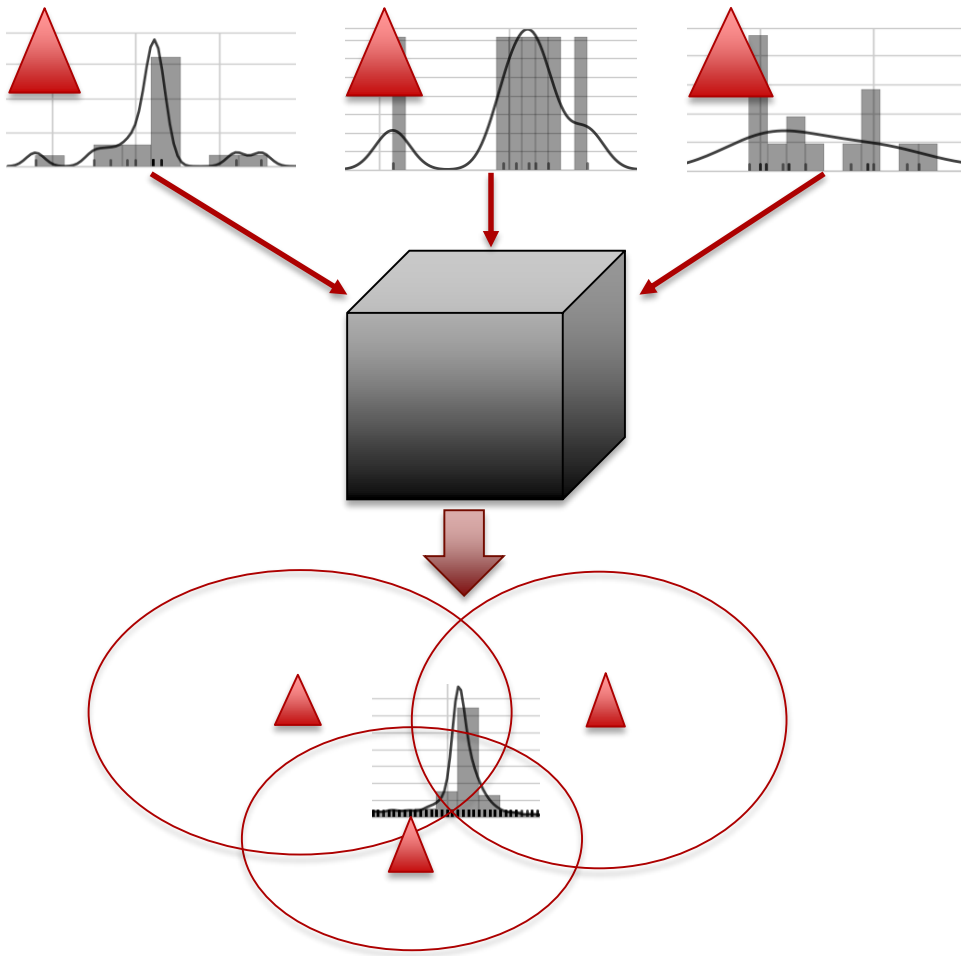
- UQ does not answer these questions, but it might provide some insight
  - A metric that measures something different from Acc, Prec, Recall, ROC





Future Work

- Combine Distributions from multiple sensors for downstream analysis
  - Such as origin location



Method 1: Parametric Bootstrap

- Use AIC to select *p,q* for $\mathcal{M}_1$ & $\mathcal{M}_2$
- Fit model parameters for each, optimizing *k*
- Sample $\mathcal{M}_1$ & $\mathcal{M}_2$ to create new waveforms
- Fit new models to each sample and record *k*
- Compute the sampling distribution of the estimate k

Method 2: Model Sampling

- $\mathcal{M}_1$=ARMA(p_1,q_1); $\mathcal{M}_2$=ARMA(p_2,q_2)
- {p_1, q_1, p_2, q_2} determines *k* and a likelihood
- Sample from {p_1, q_1, p_2, q_2 ,*k*} and fit the data
- Use likelihoods and the prior distributions on {p_1, q_1, p_2, q_2 ,*k*} to construct a posterior distribution for k

Not clear that these produce similar distributions

Summary

- *Uncertainty analysis does not "build a better model"*
  It indicates how well a given model captures the data

- Research is to bridge the theory – application gap
  - Work in space of available data
  - Automatic model selection and calibration
  - Propagating uncertainty through layers of analysis
  - Use UQ to determine if the model is doing what analyst expects (given lack of ground truth)

Coming Soon: PyPickUQ

- Important questions
  - What's the relationship between uncertainties generated by
    - Measurement errors & data sampling (bootstrap)
    - Model selection and induction processes (model sampling)
    - Inference (MCMC)
  - …and how do we combine them? Or should we?
  - What issues arise when we propagate uncertainties from one statistical inverse problem to another?

Sandia National Laboratories

Questions

POC: Matt Peterson
mgpeter@sandia.gov