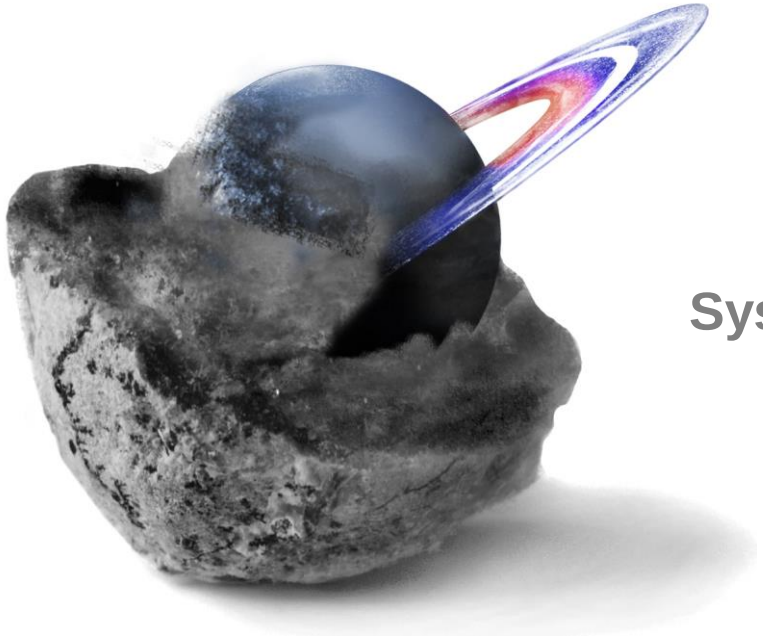
Early Formulation Cost Modeling

Balancing NASA's Vast Engineering Knowledge Base with Hard Data



Michael DiNicola

Systems Modeling, Analysis & Architecture

Jet Propulsion Laboratory

Michael.Dinicola@jpl.nasa.gov

JPL Innovation Foundry

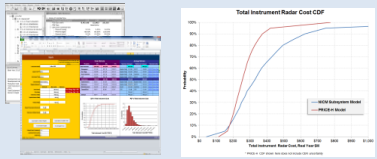
Copyright 2015 California Institute of Technology.
U.S. Government sponsorship acknowledged.

Agenda

- **Motivation**
- **Bayesian Statistics**
  - How can it help us?
  - How is it applicable?
  - What is it?
  - Handling of different kinds of information
- **Example & Discussion**
- **Concluding Remarks**

Motivation – A Cost Estimation Scenario

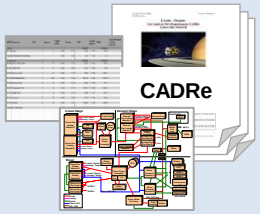
New Mission Concept



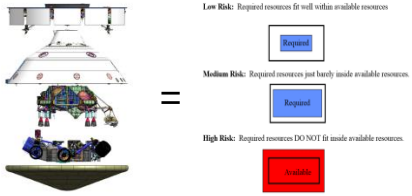
Develop, Calibrate & Run Cost Models

Perform Build-up "Grassroots" Estimate

Gather Data



Review Modeling Assumptions / Results with Experts, Team, Project, etc.



Formulate Cost Recommendation for New Mission Concept

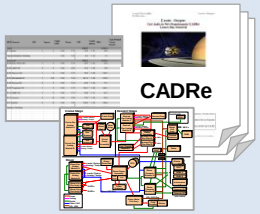
Expertise and actual data, working together, is key to the cost estimation process.

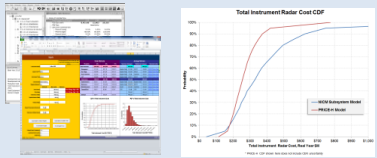
Motivation – A Cost Estimation Scenario

New Mission Concept

Uniqueness of concept can be problematic when using historical cost data

Gather Data

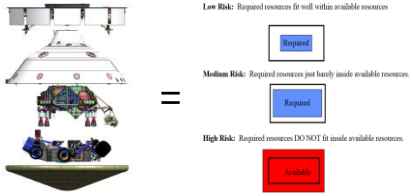




Develop, Calibrate & Run Cost Models

Perform Build-up "Grassroots" Estimate

Review Modeling Assumptions / Results with Experts, Team, Project, etc.



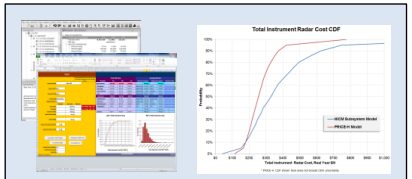
Formulate Cost Recommendation for New Mission Concept

Expertise and actual data, working together, is key to the cost estimation process.

Motivation – A Cost Estimation Scenario

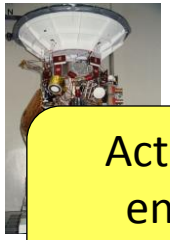
Uniqueness of concept can be problematic when using historical cost data

New Mission Concept



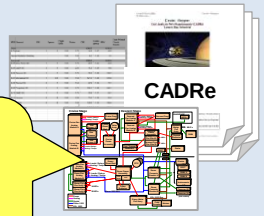
Develop, Calibrate & Run Cost Models

Perform Build-up "Grassroots" Estimate

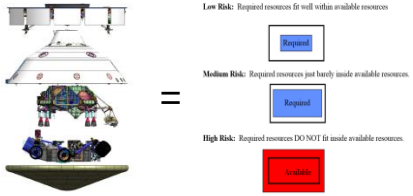


Actual cost data ensures some grounding in reality

Gather Data



Review Modeling Assumptions / Results with Experts, Team, Project, etc.



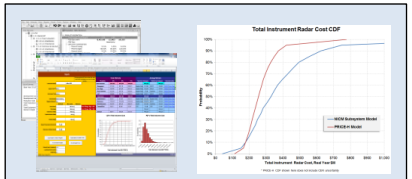
Formulate Cost Recommendation for New Mission Concept

Expertise and actual data, working together, is key to the cost estimation process.

Motivation – A Cost Estimation Scenario

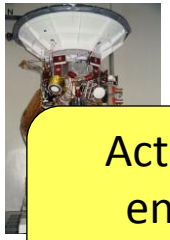
Uniqueness of concept can be problematic when using historical cost data

New Mission Concept



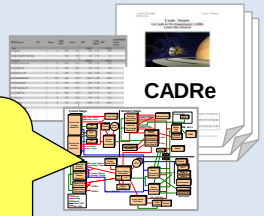
Develop, Calibrate & Run Cost Models

Perform Build-up "Grassroots" Estimate



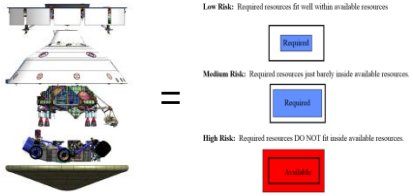
Actual cost data ensures some grounding in reality

Gather Data



Review Modeling Assumptions / Results with Experts, Team, Project, etc.

Tend to be ad-hoc adjustments made to cost modeling

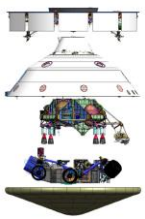


Formulate Cost Recommendation for New Mission Concept

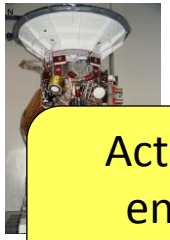
Expertise and actual data, working together, is key to the cost estimation process.

Motivation – A Cost Estimation Scenario

Uniqueness of concept can be problematic when using historical cost data

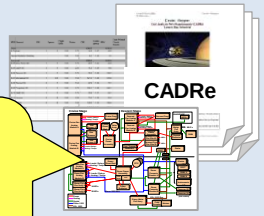


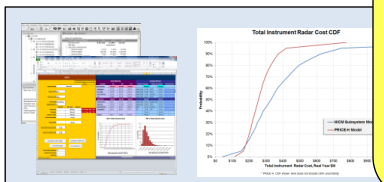
New Mission Concept



Actual cost data ensures some grounding in reality

Gather Data





Develop, Calibrate & Run Cost Models

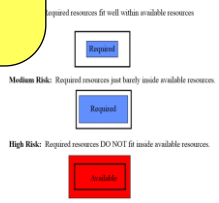
Perform Build-up "Grassroots" Estimate

Review Modeling Assumptions / Results with Experts, Team, Project, etc.

Tend to be ad-hoc adjustments made to cost modeling

In the end, a cost assessment that captures all relevant sources of information is desired

=



Formulate Cost Recommendation for New Mission Concept

Expertise and actual data, working together, is key to the cost estimation process.

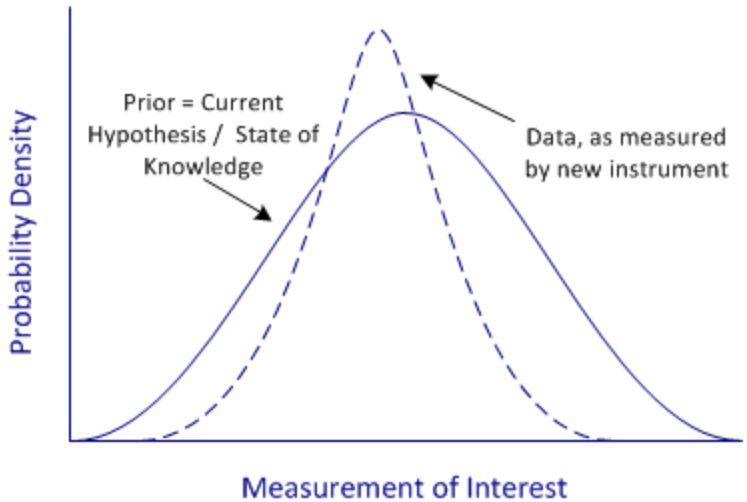
Bayesian Statistics – How Can It Help Us?

- **A "full" probability model**
  - A way to use all available sources of information **from the start of the analysis**
  - A way to incorporate all relevant sources of uncertainty
- **Incorporates engineering, scientific and financial expertise to capture unique aspects of the concept in the probability model; not limited to just the data**
- **Balances expert opinion with the evidence as realized by the data**
- **Avoids ad-hoc adjustments to model output that degrades the interpretation of the probabilistic cost assessment**
- **Small datasets: even one data point can provide useful information!**

The Bayesian approach can provide our probabilistic assessments with a more meaningful interpretation.

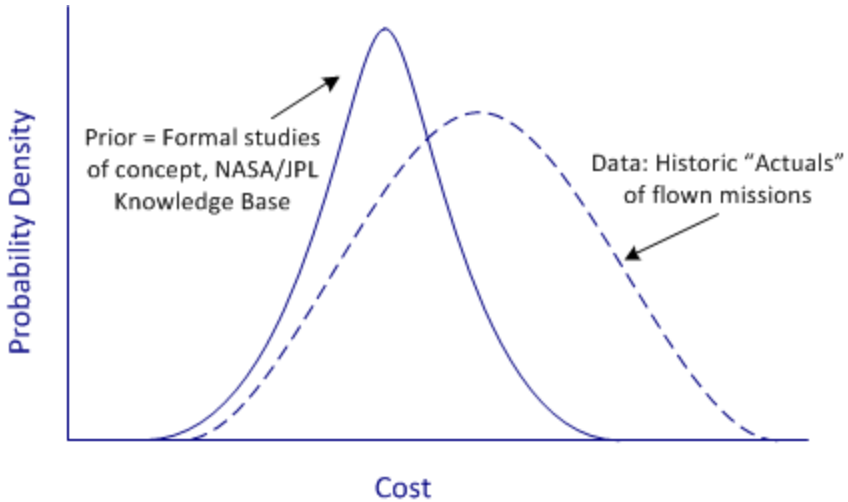
Bayesian Statistics – How Is It Applicable?

Scientific Analysis



- Unlikely that a new instrument will be developed if it does not decrease a measurement's uncertainty enough to draw significant conclusions
- Bayesian analysis will adjust the belief of a prior hypothesis
- Bayesian analysis can reinforce a current hypothesis, entertain other hypotheses, or lead to new hypotheses

Cost Analysis

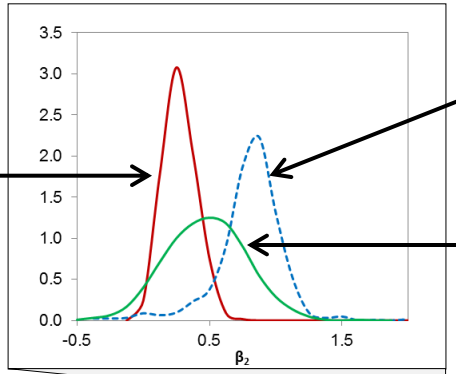


- While expertise can evaluate unique aspects of a mission, they may contain biases (optimistic mean, uncertainty)
- While historic data is grounded in reality, analogies can be rough, producing wider uncertainty than may be appropriate
- Bayesian analysis will **balance – be a compromise between** – the hard data and institutional expertise

Bayesian Statistics – What Is It?

	Bayesian	Classicalist
Data	**Fixed** as observed. **Provides the evidence** for things unknown.	A **random variable**, observed with error.
Distribution Parameters	Unobservables that **require probability distributions**.	**Unknown but have "true" fixed values** that we estimate.

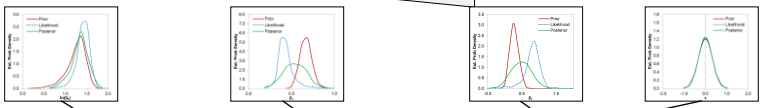
Prior = Pr(β) based on formal studies of concept, NASA/JPL Knowledge Base



Data Likelihood = Pr(Data|β)
How the historical data supports the expert knowledge

Posterior = Pr(β|Data)
= Likelihood x Prior / Pr(Data)
We use this to predict – the Expertise calibrated to the Data

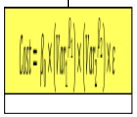
Bayesian



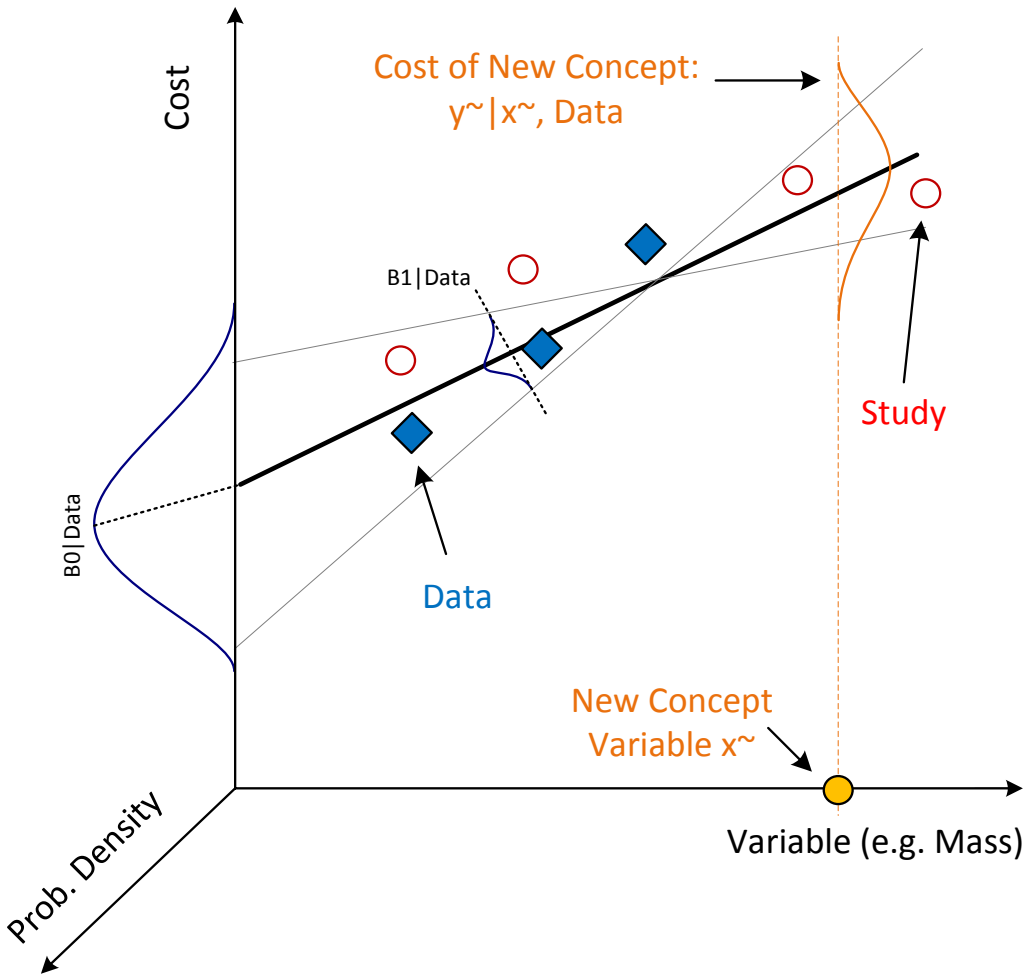
$$Cost = \beta_0 \times \left(Var_1^{\beta_1}\right) \times \left(Var_2^{\beta_2}\right) \times \varepsilon$$

Classicalist

Fit to Data



Bayesian Regression – What Is It?



Graph is notional. Σ, θ not illustrated for presentation purposes.

Bayesian Regression informs the **relationship amongst variables** (the parameters) and how they move with Cost

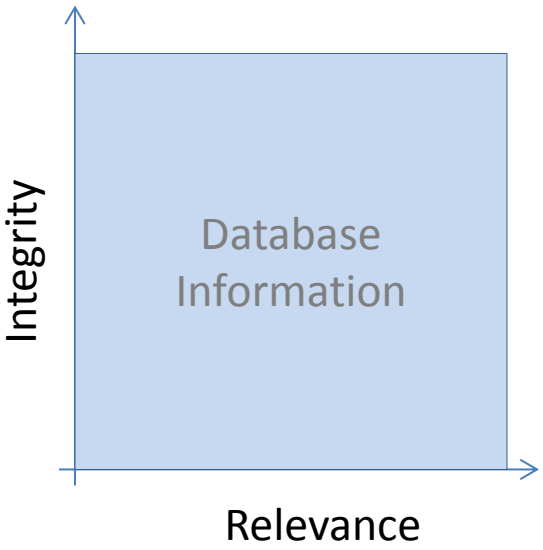
The values of different variables do not matter as much as the relationships they share.

In case you were wondering:

$$p(y\sim \mid X\sim, Data) = \int_\beta \int_\Sigma \int_\theta \left[p(y\sim \mid \beta, \Sigma, \theta, X\sim, Data)\, p(\beta, \Sigma, \theta \mid X\sim, Data) \right] d\theta\, d\Sigma\, d\beta$$

Treating Different Kinds of Information

- **All types of information (historical actual data, studies, expertise) need to be assessed for its proper role in the model**
- **Many ways to do this:**
  - Do nothing: let the data do the work
    - Good, if the data is good and very relevant
    - To much weight given to actuals if they are not extremely similar to what is being estimated
  - Pseudo-observations: Each study that forms the prior distribution is treated as *x* number of actual observations (typically $0<x\leq 1$)
    - This is the path taken in the example to follow (x=0.5)
  - Prediction strength: Weight different pieces of information to maximize prediction strength of model
  - Hierarchical modeling and many other techniques

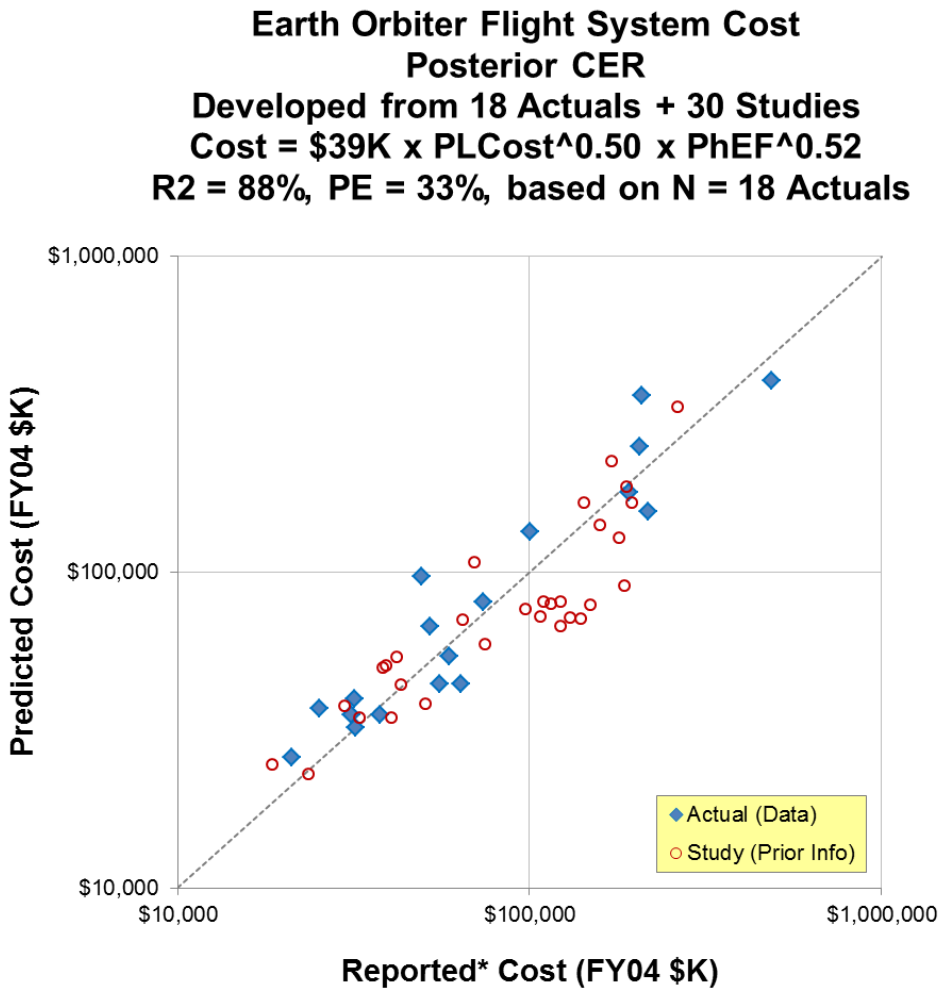


Treat all information going into the model according to its relevance and integrity.

An Example: Early Concept Formulation

- **Goal: Evaluate feasibility of a given mission concept**
  - What cost family does it belong to?
- **JPL has a cost & technical database of actual historic data as well as an inventory of concept studies**
- **Regression models assembled for Spacecraft and I&T costs**
  - Earth Orbiting model example shown here
  - Uses a conjugate prior model:
    - **$(\beta, \Sigma, \theta \mid \text{Data})$ drawn from a Normal-Inverse-Wishart distribution**

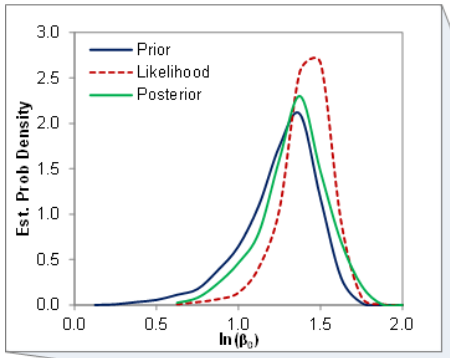
An Example: Earth Orbiting Spacecraft CER

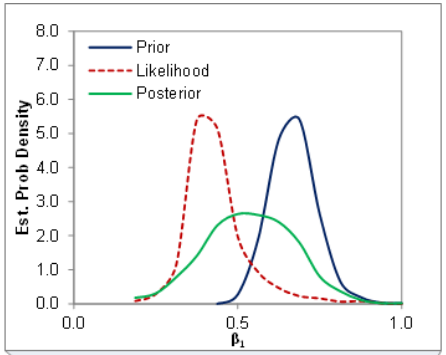


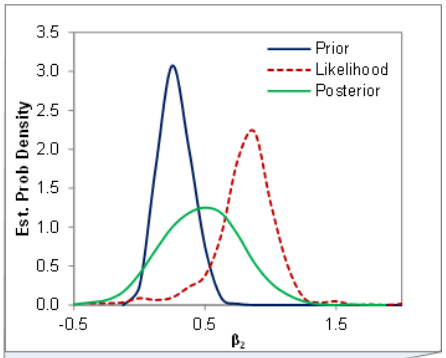
*Reported Cost, Actual: The actual cost documented as of Launch or later.
Reported Cost, Study: The documented cost estimate (w/o reserve).

- **All information (Actuals & Studies) used to build CER**
- **Only actual data used to validate CER**
  - Bootstrap Cross-validation used to derive "PE" = Prediction Error
- **Significant Cost Drivers/Predictors:**
  - Phase E/F Duration (PhEF)
  - Payload Cost (PLCost)
  - Coefficients shown in equation are the means of *posterior* probability distributions (next slide)

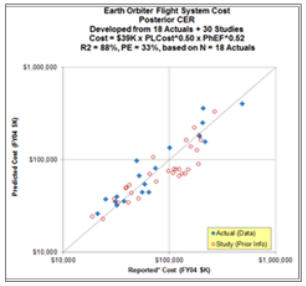
An Example: Earth Orbiting Spacecraft CER







$$S/C\ Cost \sim \boxed{\beta_0} \times \left(PLCost^{\boxed{\beta_1}}\right) \times \left(EFDur^{\boxed{\beta_2}}\right)$$



Bayesian analysis will balance – be a compromise between – the hard data and institutional expertise

Concluding Remarks

- **Bayesian Regression gives a very flexible and versatile framework to capture uncertainty**
  - Informs the relationships between variables
- **Balances, via compromise, differences between "prior information" and data**
- **Can provide our probabilistic models a more meaningful interpretation**
- **The Prior Distribution**
  - Does not need to be feared as a "feeling-based" subjective element that skews the analysis
    - More objective priors can be developed from relevant data and past studies
  - Every credible cost assessment will (and does) rely in part on the injection of expert opinion due to the uniqueness of missions and areas where our data is inadequate
- **Future work:**
  - Generalizing outside the Gaussian framework
  - Explore hierarchical modeling and other ways of weighting information
  - Other Bayesian methods of model checking (e.g. Bayes' Factor)
  - Visualization and Model Adaptability

Bayesian methods provide many new perspectives, allowing us to leverage critical sources of information: expertise and data, in a sound probabilistic framework.

Backup Slides

- Bayes' Theorem and its Statistical Interpretation for this Application
- Covariance structure between Regression Coefficients
- Bayesian Universe: Observables vs. Unobservables
- Bayesian Formulas and the Likelihood Function
- CER Development Process
- CER Conjugate Prior Model
- Abstract

Bayesian Statistics – What Is It (in Math Terms)?

Mathematically:
Bayes' Theorem

For any events A and B in a probability space,

$$\Pr(A \mid B) = \Pr(B \mid A) \times \Pr(A) / \Pr(B)$$

A → set of parameters, **Parms**
B → **Data** = $\{y_1, y_2, \ldots, y_n\}$

Posterior | Likelihood | Prior

Statistically:

$$p(\text{Parms} \mid \text{Data}) = p(\text{Data} \mid \text{Parms}) \times p(\text{Parms}) / p(\text{Data})$$

y = Xβ + θ: Linear model
β: Regression coefficients,
Σ: Covariance matrix of the **β**s
θ: Our fundamental model variability not accounted for by **Xβ**

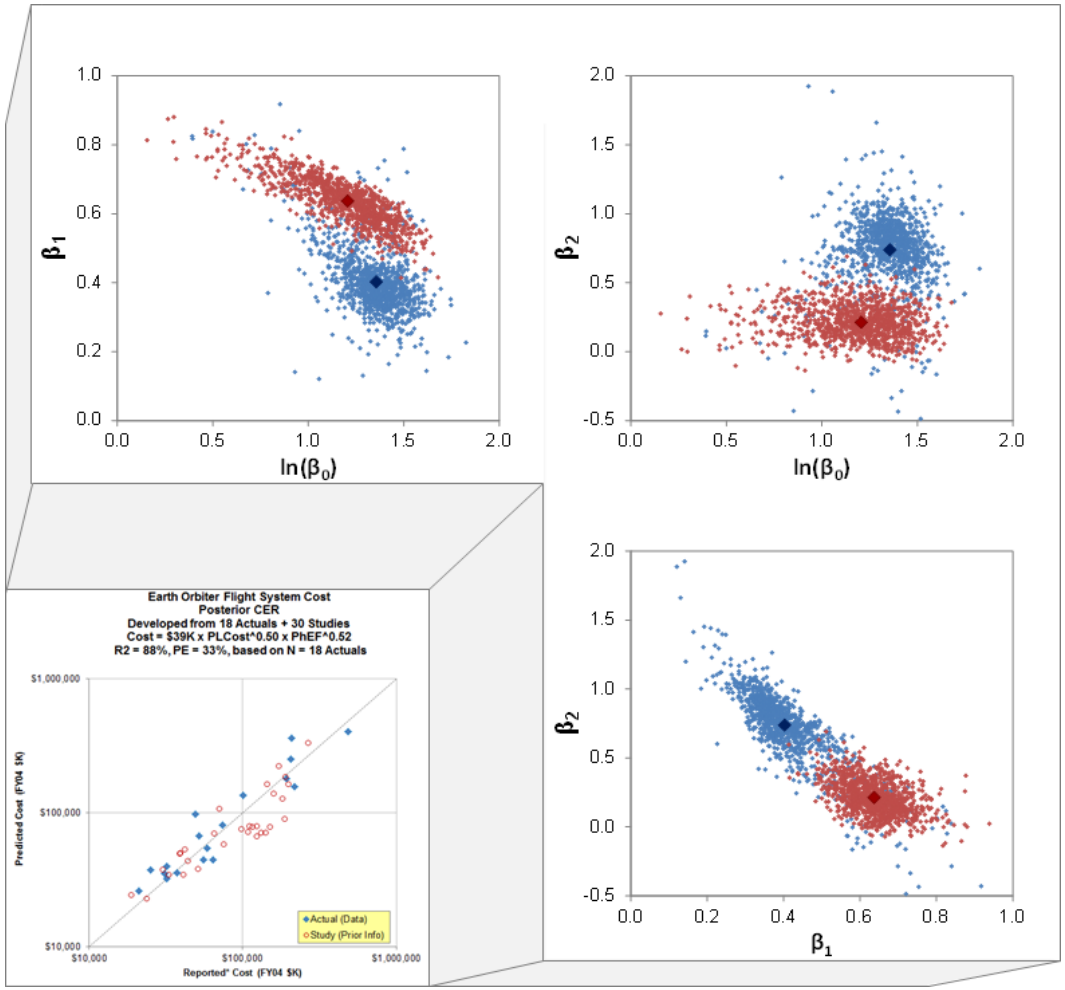
For Our Application:

$$p(\beta, \Sigma, \theta \mid \text{Data}) = p(\text{Data} \mid \beta, \Sigma, \theta) \times p(\beta, \Sigma, \theta) / p(\text{Data})$$

We integrate with this to get ***p*(y~ | X~, Data)**, the probability of a new concept's cost (**y~**), given it's predictors (**X~**) and the Data.

An Example: Earth Orbiting Spacecraft CER

$$S/C\ Cost \sim \beta_0 \times \left(PLCost^{\beta_1}\right) \times \left(EFDur^{\beta_2}\right)$$



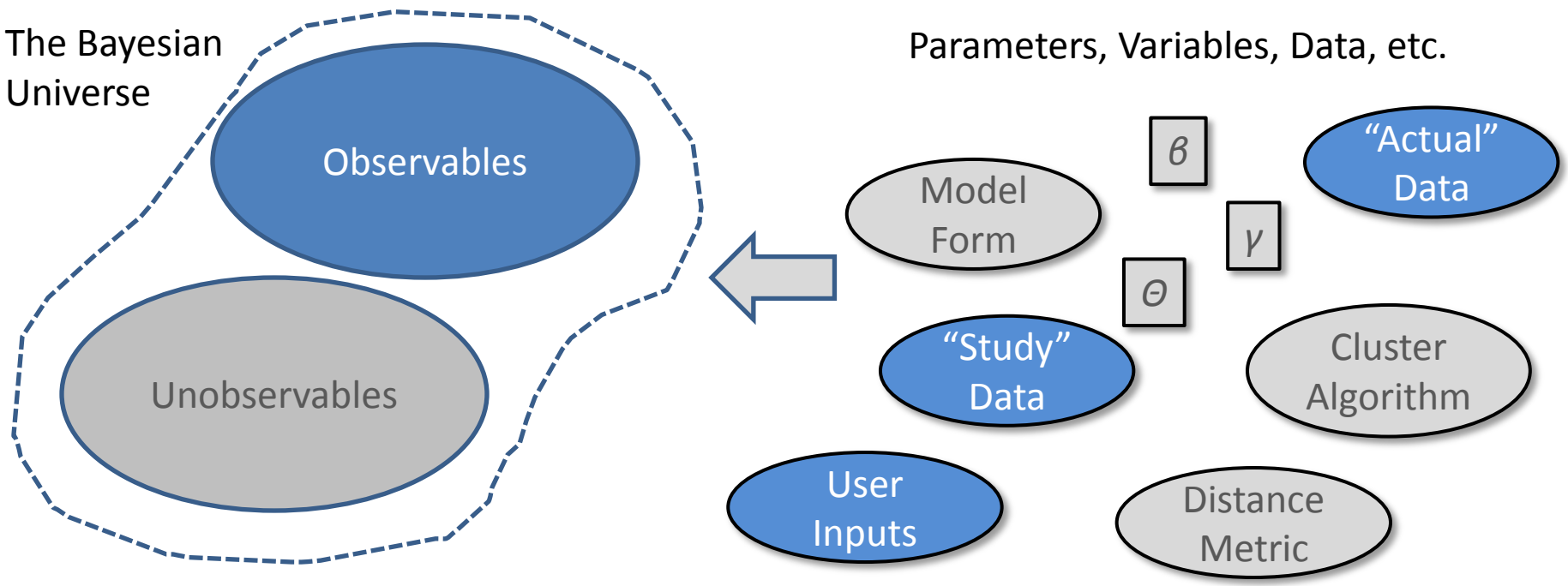
◆ Likelihood (Bootstrapped)

◆ Prior (Bootstrapped)

- **Significant difference between how Studies and Actuals regard mean impact of cost driving variables**
- **Covariance structure similar between Actual data and Studies**
- **Difference in mean values of parameters adds to the total posterior variance**
  - Posterior covariance not shown for presentation purposes

Bayesian Statistics – What Is It?

	Bayesian	Classical
Data	**Fixed** as observed. **Provides the evidence** for things unknown.	A **random variable**, observed with error.
Distribution Parameters	Unobservables that **require probability distributions**.	**Unknown but have "true" fixed values** that we estimate.



Bayesian Statistics – What Is It?

- **Mathematically: Bayes' Theorem**
  - For any events A and B in a probability space,

$$\mathbf{Pr(A \mid B) = Pr(B \mid A) \times Pr(A) / Pr(B)}$$

- **Statistically: In the above, let**
  - Let A be a set of parameters, denoted by β and θ, depending on the context
  - Let B be a set of data, denoted **Data** = $\{\mathbf{y_1, y_2, \ldots, y_n}\}$
  - Use probability density function notation since we are using random variables

$$\boldsymbol{p}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{Data}) = \boldsymbol{p}(\mathbf{Data} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) \times \boldsymbol{p}(\boldsymbol{\beta}, \boldsymbol{\theta}) / \boldsymbol{p}(\mathbf{Data})$$

  - Once the data is observed, we can write p(Data | β, θ) as the "likelihood function", a function of β, θ, denoted L(β, θ | Data)
  - p(Data) is a normalizing factor (constant) in the calculations, so, for this presentation we remove it and write:

$$\boldsymbol{p}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{Data}) \propto \boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{Data}) \times \boldsymbol{p}(\boldsymbol{\beta}, \boldsymbol{\theta})$$

- **Do not mean to downplay the importance of *p*(Data)**
  - Plays a very interesting role in Bayesian analysis, especially for evaluating the evidence when comparing models

An Example: CER Development Process

Architect Model	Assess Info in Database	Group Info (Homogeneity)	Identify Key Variables	Calculate CERs & Validate
Gaussian Conjugate Prior Model*	Available Categorical Fields	Cluster Analysis	Principal Components Analysis	Significance tests, Normality check
	Available Quantitative Fields	Principal Components Analysis	Pairwise OLS Regression	Variogram Assessment
				Bootstrap Cross-validation
				Bayes Factor

- Information – Actual historic data and concept Study information – assessed *individually* and *collectively*
- Model CERs developed with all data, validated on actuals only

* Normal-(Inverse) Wishart conjugate prior model. This model was chosen for simplicity of application and fidelity for cost estimation during early formulation.

Example CER: Conjugate Prior Model

- **Conjugate Prior Model:** Where the Prior Distribution and Posterior Distributions have the same distributional form, just different parameters.
- For our application we assume a Multivariate Normal likelihood model (in log-log space) for simplicity of application and fidelity for cost estimation during early formulation.
- Our conjugate prior model is a Normal-Inverse-Wishart distribution
  - Assumes both mean and covariance structure of regression coefficients are unknown.
  - Use Inverse Wishart distribution to simulate Sum of Squares matrix (which gives covariance matrix)
  - Then simulate regression coefficients using Multivariate Normal Distribution, given the Covariance matrix

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive[note 4]
Multivariate normal	$\boldsymbol{\mu}$ (mean vector) and $\boldsymbol{\Sigma}$ (covariance matrix)	normal-inverse-Wishart	$\boldsymbol{\mu}_0,\ \kappa_0,\ \nu_0,\ \boldsymbol{\Psi}$	$\frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\mathbf{x}}}{\kappa_0 + n},\ \kappa_0 + n,\ \nu_0 + n,$ $\boldsymbol{\Psi} + \mathbf{C} + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T$ • $\bar{\mathbf{x}}$ is the sample mean • $\mathbf{C} = \sum_{i=1}^{n}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T$	mean was estimated from κ_0 observations with sample mean $\boldsymbol{\mu}_0$; covariance matrix was estimated from ν_0 observations with sample mean $\boldsymbol{\mu}_0$ and with sum of pairwise deviation products $\boldsymbol{\Psi} = \nu_0\boldsymbol{\Sigma}_0$	$t_{\nu_0'-p+1}\left(\tilde{\mathbf{x}}\vert\boldsymbol{\mu}_0',\ \frac{\kappa_0'+1}{\kappa_0'(\nu_0'-p+1)}\boldsymbol{\Psi}'\right)$[5]

Source: https://en.wikipedia.org/wiki/Conjugate_prior

Abstract

- Assessing **cost and feasibility of NASA space mission concepts at early design phases** requires not only relevant **flown mission data**, but also **strong engineering, scientific and financial expertise** to guide the concept into what is many times new territory.
- Presented here is a **Bayesian method for developing Cost Estimating Relationships** that leverages both of these critical sources of information (i.e. expertise and data).
- This is done within a **flexible** modeling framework, allowing for **real-time probabilistic cost assessments**.
- We discuss **how this method treats different kinds of information** available and **how to interpret results**.
- **Practical application** of this method is also discussed, within the context of assessing mission feasibility, before commitments are made, proposals submitted and projects implemented.