

Early Formulation Cost Modeling

Balancing NASA's Vast Engineering
Knowledge Base with Hard Data



JPL Innovation Foundry

Michael DiNicola

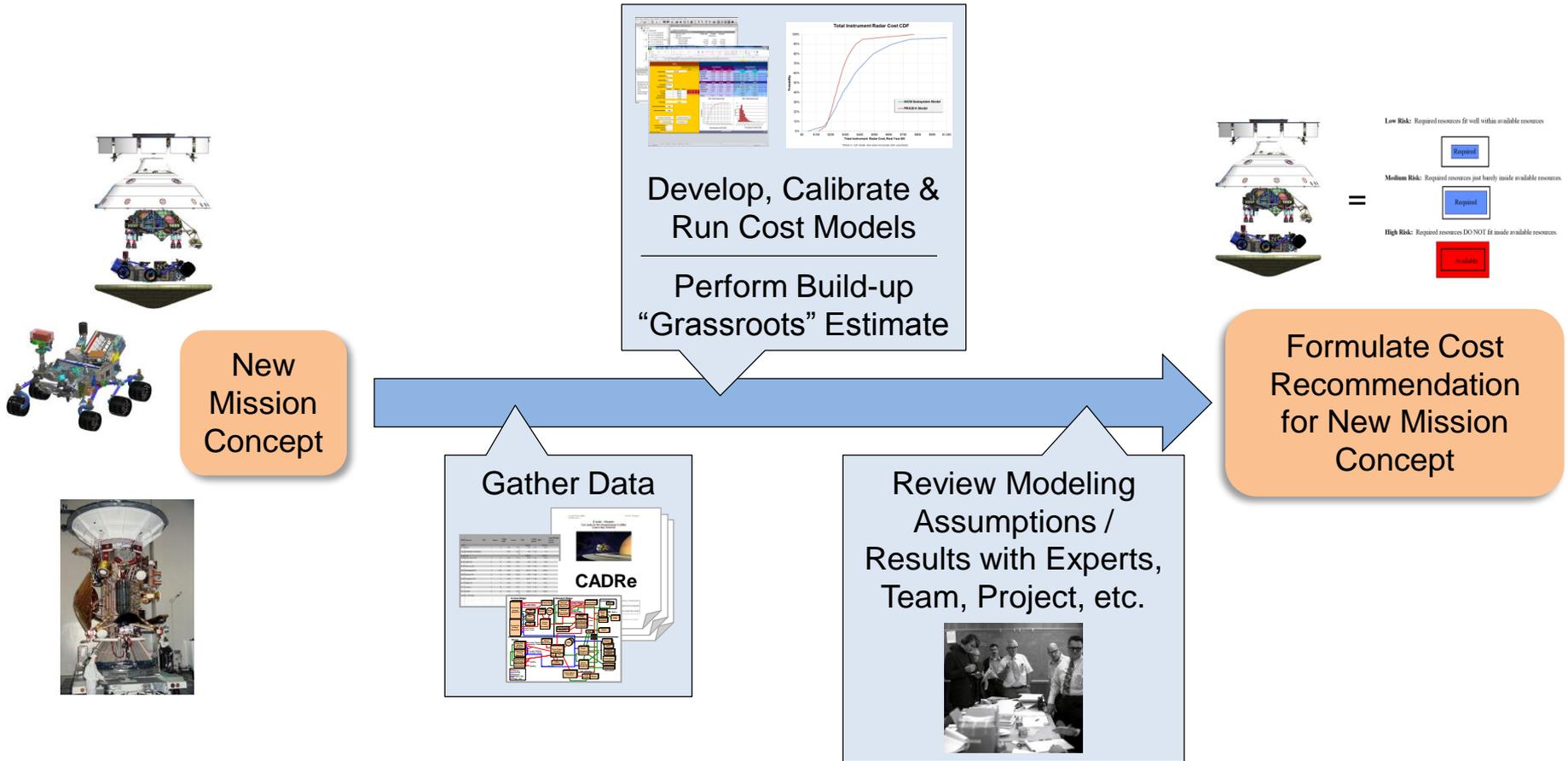
Systems Modeling, Analysis & Architecture

Jet Propulsion Laboratory

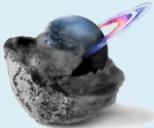
Michael.Dinicola@jpl.nasa.gov

- **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



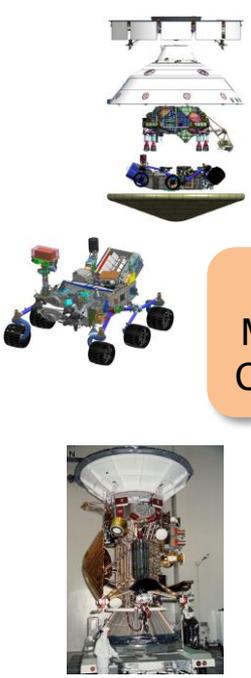
Expertise and actual data, working together, is key to the cost estimation process.



Motivation – A Cost Estimation Scenario

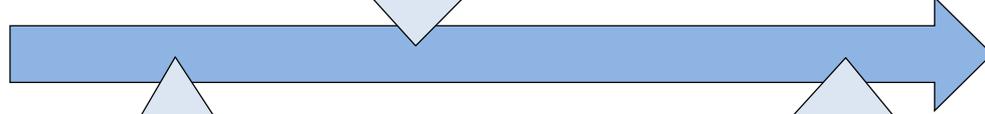


JPL Innovation Foundry



Uniqueness of concept can be problematic when using historical cost data

New Mission Concept



Develop, Calibrate & Run Cost Models

Perform Build-up "Grassroots" Estimate

Gather Data

CADRe

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

Low Risk: Required resources fit well within available resources

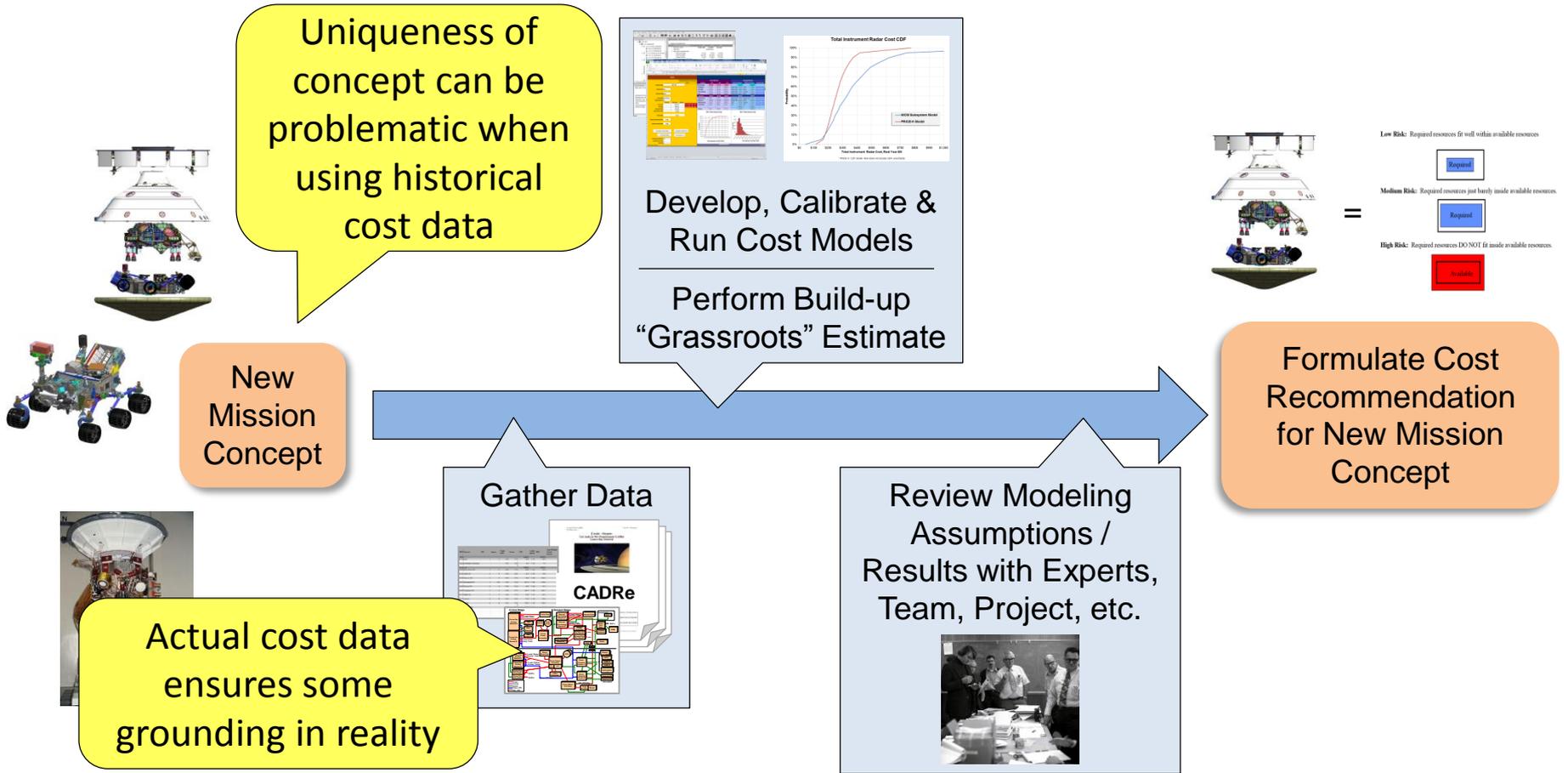
Medium Risk: Required resources just barely inside available resources

High Risk: Required resources DO NOT fit inside available resources

Formulate Cost Recommendation for New Mission Concept

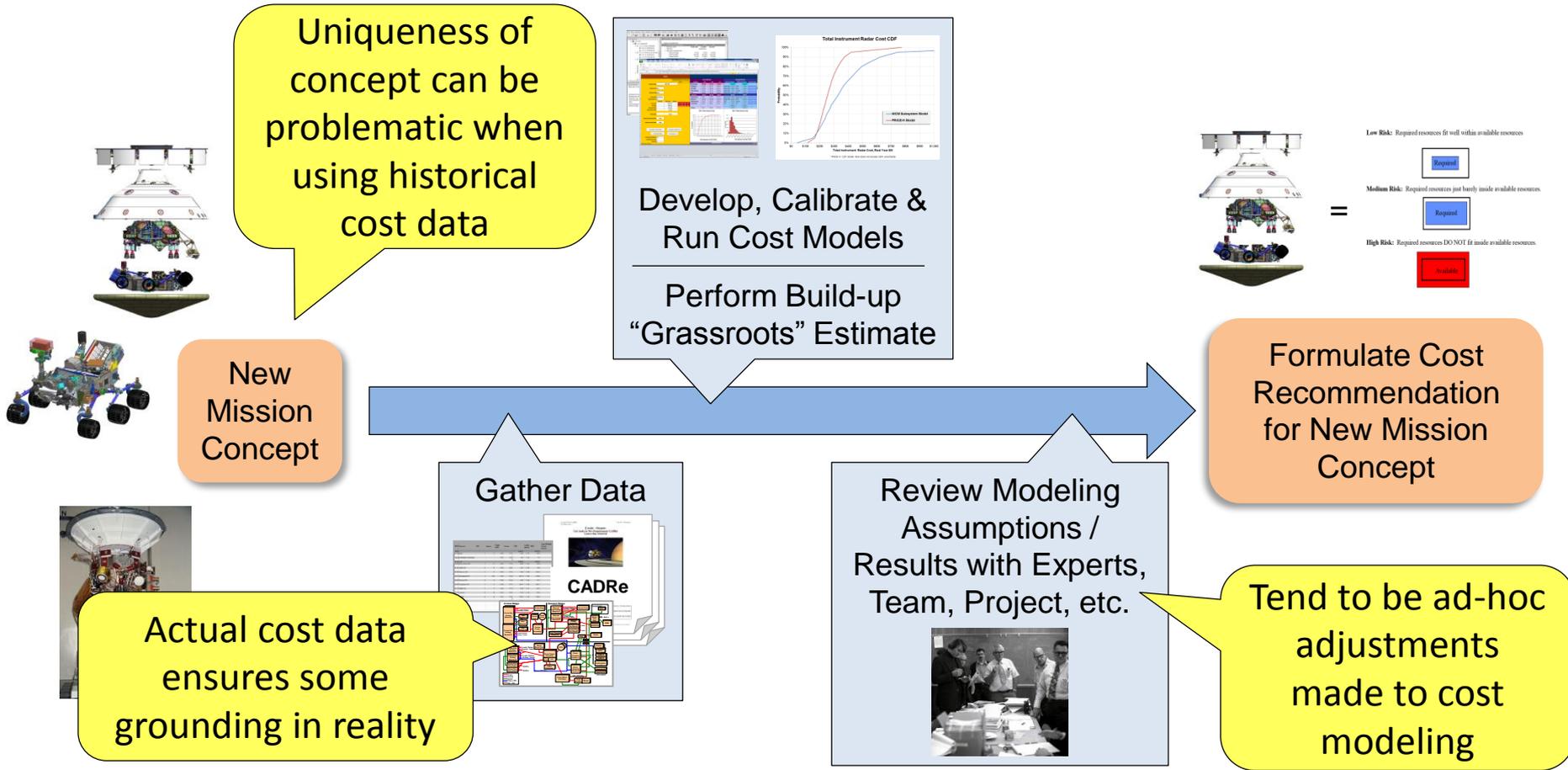
Expertise and actual data, working together, is key to the cost estimation process.

Motivation – A Cost Estimation Scenario



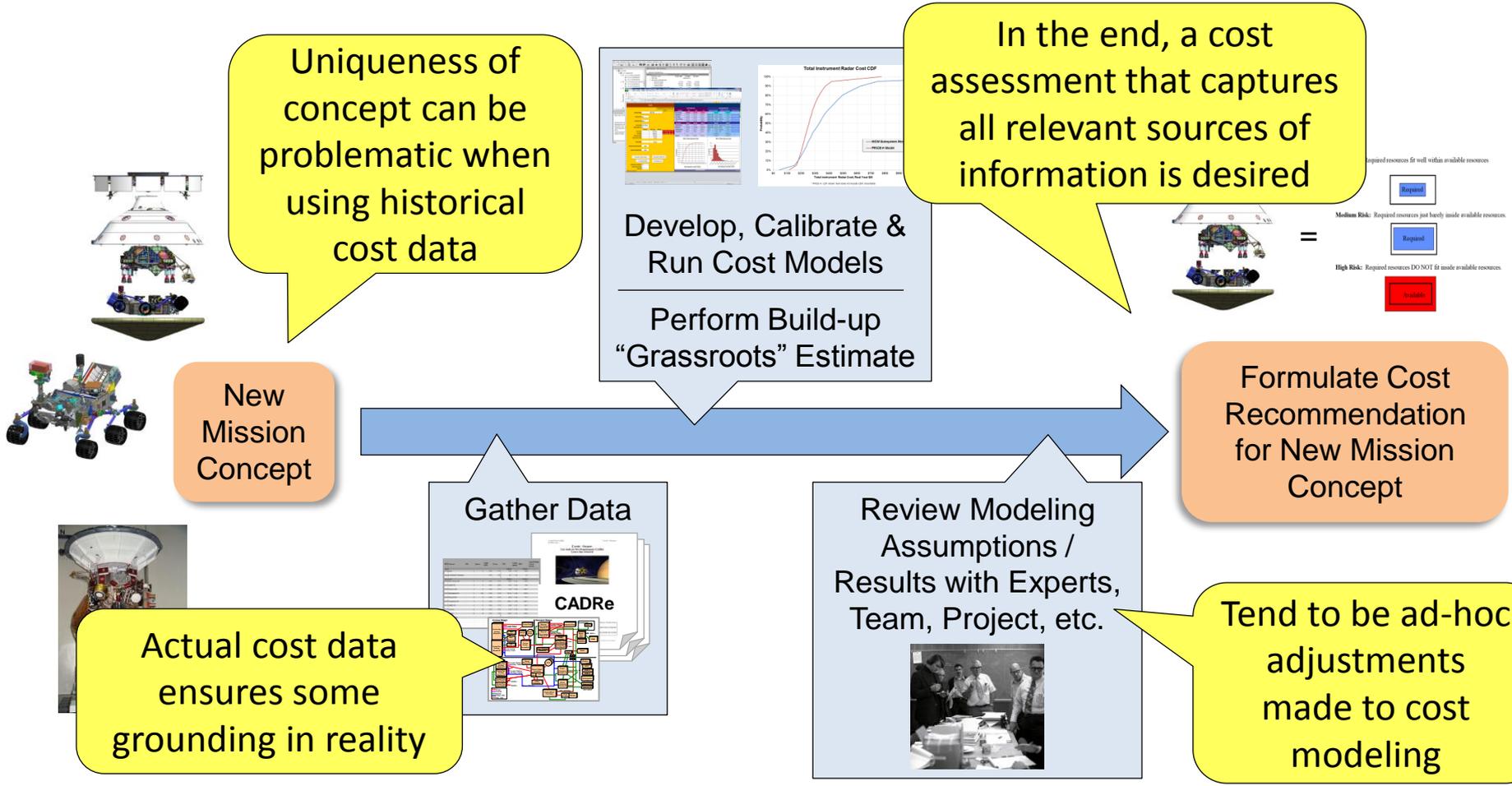
Expertise and actual data, working together, is key to the cost estimation process.

Motivation – A Cost Estimation Scenario



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

Motivation – A Cost Estimation Scenario



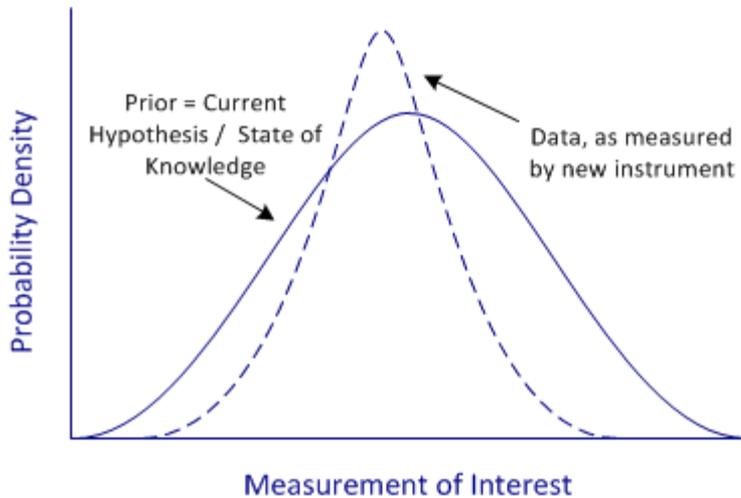
Expertise and actual data, working together, is key to the cost estimation process.



- **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!**

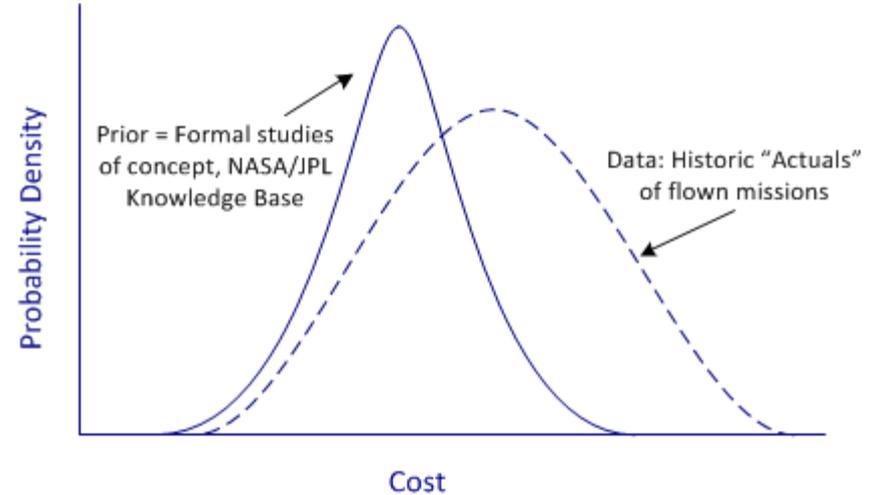
The Bayesian approach can provide our probabilistic assessments with a more meaningful interpretation.

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

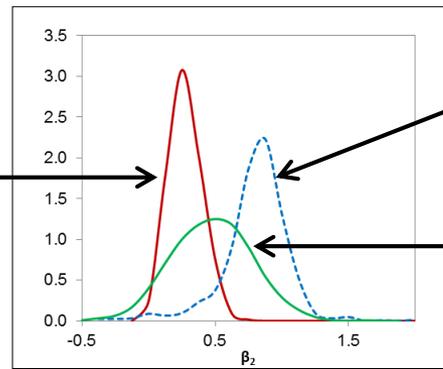


- While expertise can evaluate unique aspects of a mission, they may contain biases (optimistic mean, uncertainty)
- While historic data is grounded in reality, analogies can be rough, producing wider uncertainty than may be appropriate
- Bayesian analysis will **balance – be a compromise between** –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

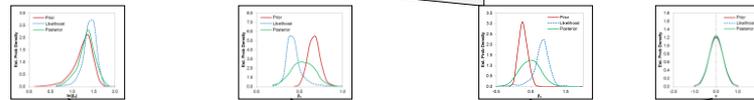
Prior = Pr(β) based on formal studies of concept, NASA/JPL Knowledge Base



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

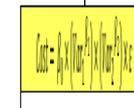


$$Cost = \beta_0 \times (Var_1 \beta_1) \times (Var_2 \beta_2) \times \epsilon$$

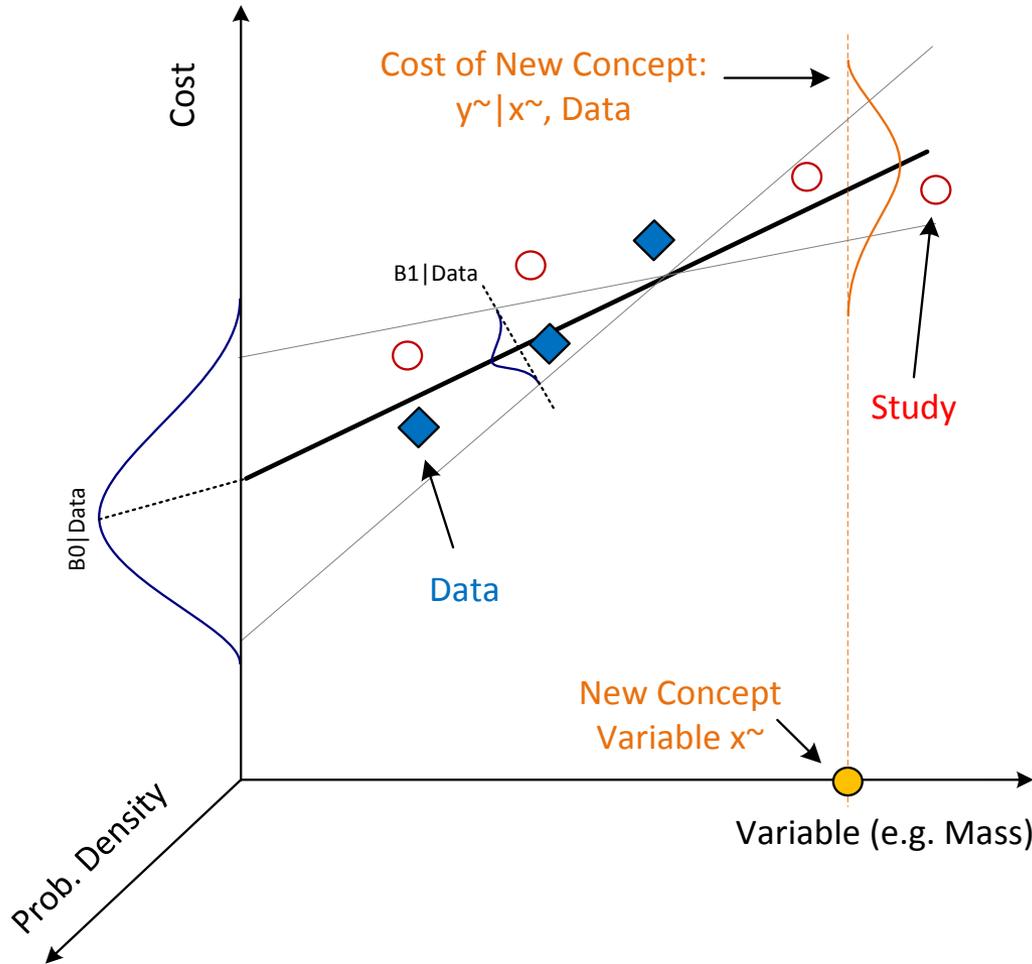
Classicalist



Fit to Data



Bayesian Regression – What Is It?



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

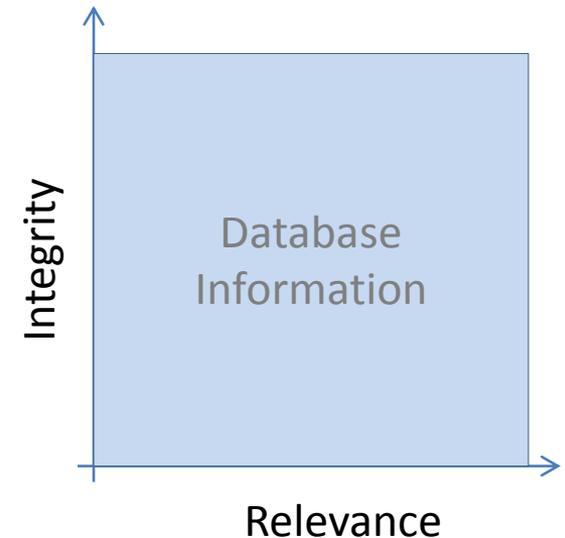
The values of different variables do not matter as much as the relationships they share.

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

In case you were wondering:

$$p(y \sim | X \sim, \text{Data}) = \int_{\beta} \int_{\Sigma} \int_{\theta} [p(y \sim | \beta, \Sigma, \theta, X \sim, \text{Data}) p(\beta, \Sigma, \theta | X \sim, \text{Data})] d\theta d\Sigma d\beta$$

- **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

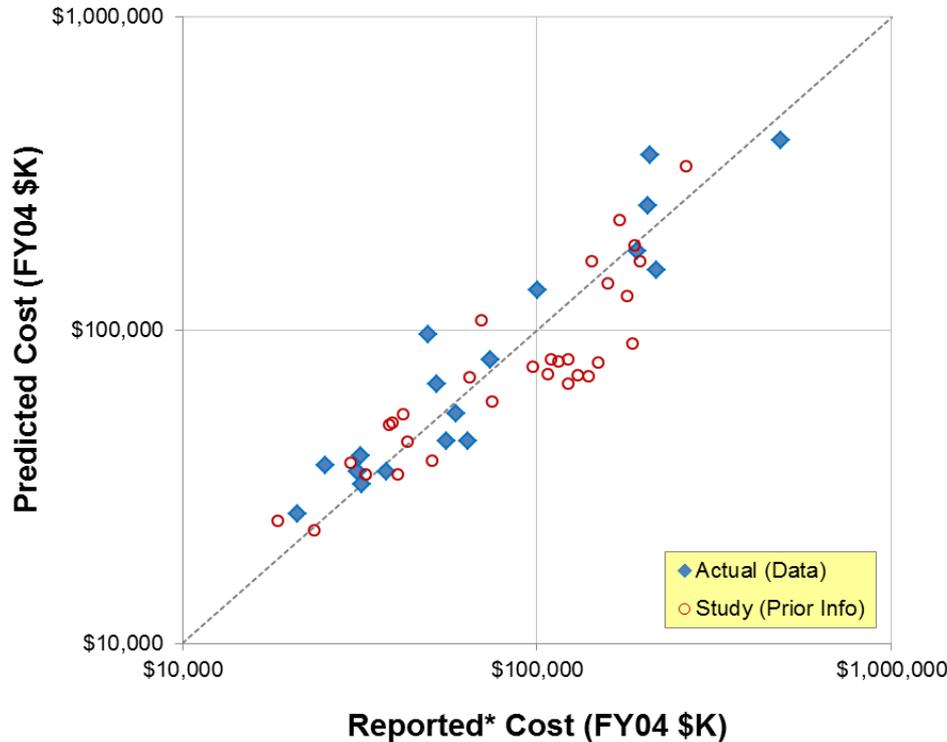


Treat all information going into the model according to its relevance and integrity.



- **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**

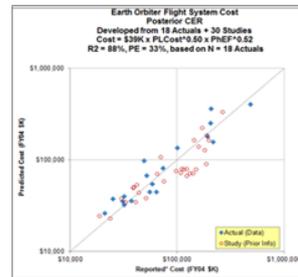
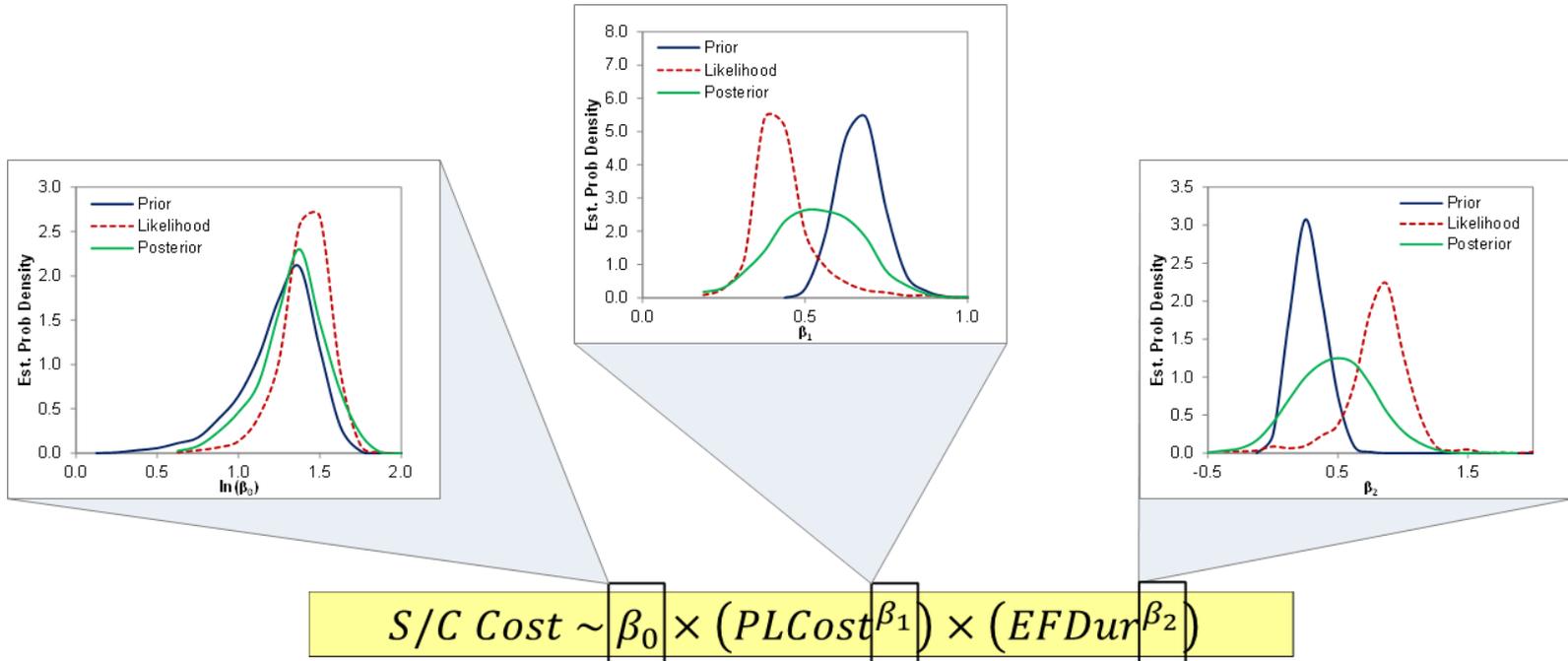
**Earth Orbiter Flight System Cost
Posterior CER**
 Developed from 18 Actuals + 30 Studies
 $\text{Cost} = \$39\text{K} \times \text{PLCost}^{0.50} \times \text{PhEF}^{0.52}$
 $R^2 = 88\%$, $\text{PE} = 33\%$, based on $N = 18$ Actuals



- **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)

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

An Example: Earth Orbiting Spacecraft CER



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.



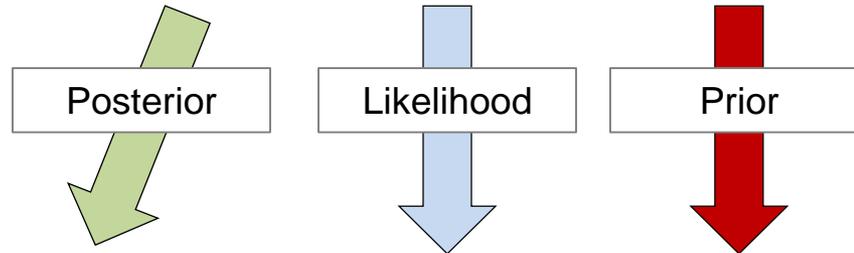
- **Bayes' Theorem and its Statistical Interpretation for this Application**
- **Covariance structure between Regression Coefficients**
- **Bayesian Universe: Observables vs. Unobservables**
- **Bayesian Formulas and the Likelihood Function**
- **CER Development Process**
- **CER Conjugate Prior Model**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

Mathematically:
Bayes' Theorem

For any events A and B in a probability space,

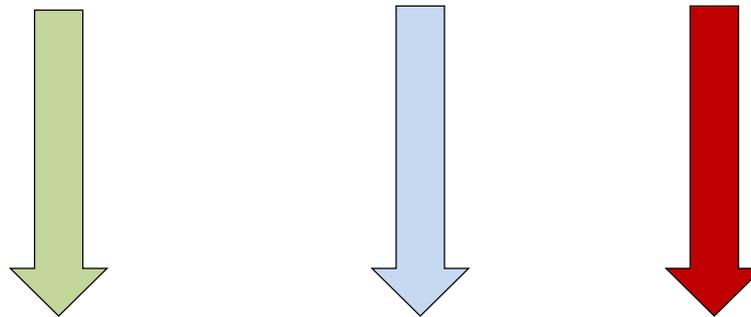
$$\Pr(A | B) = \Pr(B | A) \times \Pr(A) / \Pr(B)$$

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



Statistically: $p(\text{Parms} | \text{Data}) = p(\text{Data} | \text{Parms}) \times p(\text{Parms}) / p(\text{Data})$

$y = X\beta + \theta$: Linear model
 β : Regression coefficients,
 Σ : Covariance matrix of the β s
 θ : Our fundamental model
variability not accounted for by $X\beta$



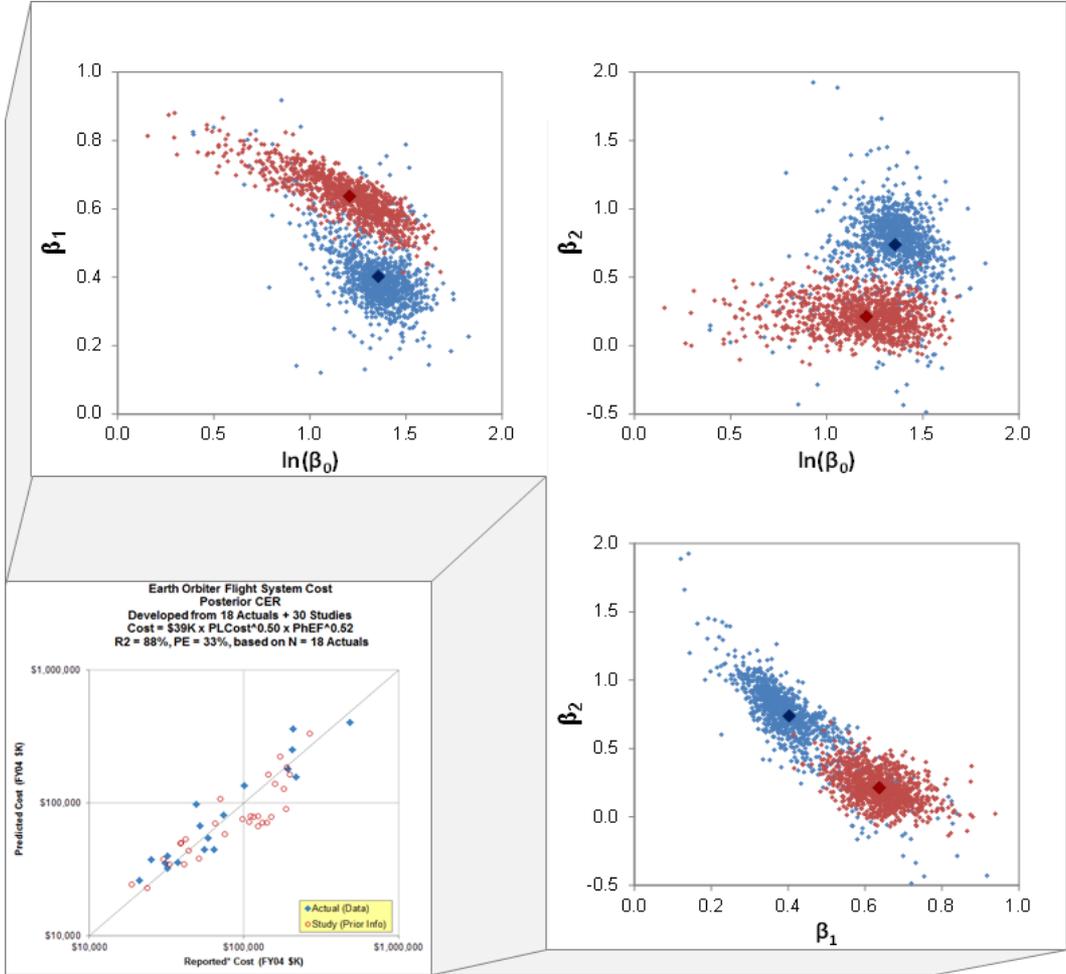
For Our Application: $p(\beta, \Sigma, \theta | \text{Data}) = p(\text{Data} | \beta, \Sigma, \theta) \times p(\beta, \Sigma, \theta) / p(\text{Data})$

We integrate with this to get $p(y_{\sim} | X_{\sim}, \text{Data})$, the probability of a new concept's cost (y_{\sim}), given it's predictors (X_{\sim}) and the Data.

An Example: Earth Orbiting Spacecraft CER

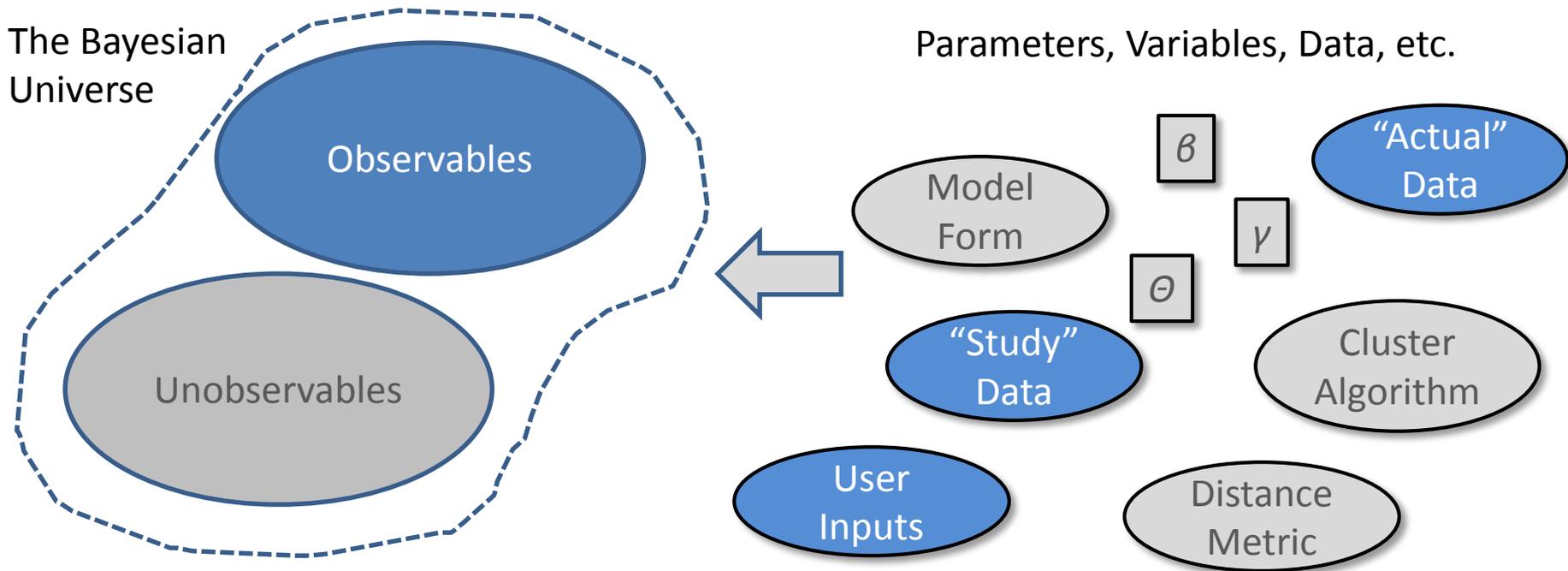


$$S/C \text{ Cost} \sim \beta_0 \times (PLCost^{\beta_1}) \times (EFDur^{\beta_2})$$



- 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	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.



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

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

- **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** = $\{y_1, y_2, \dots, y_n\}$
- Use probability density function notation since we are using random variables

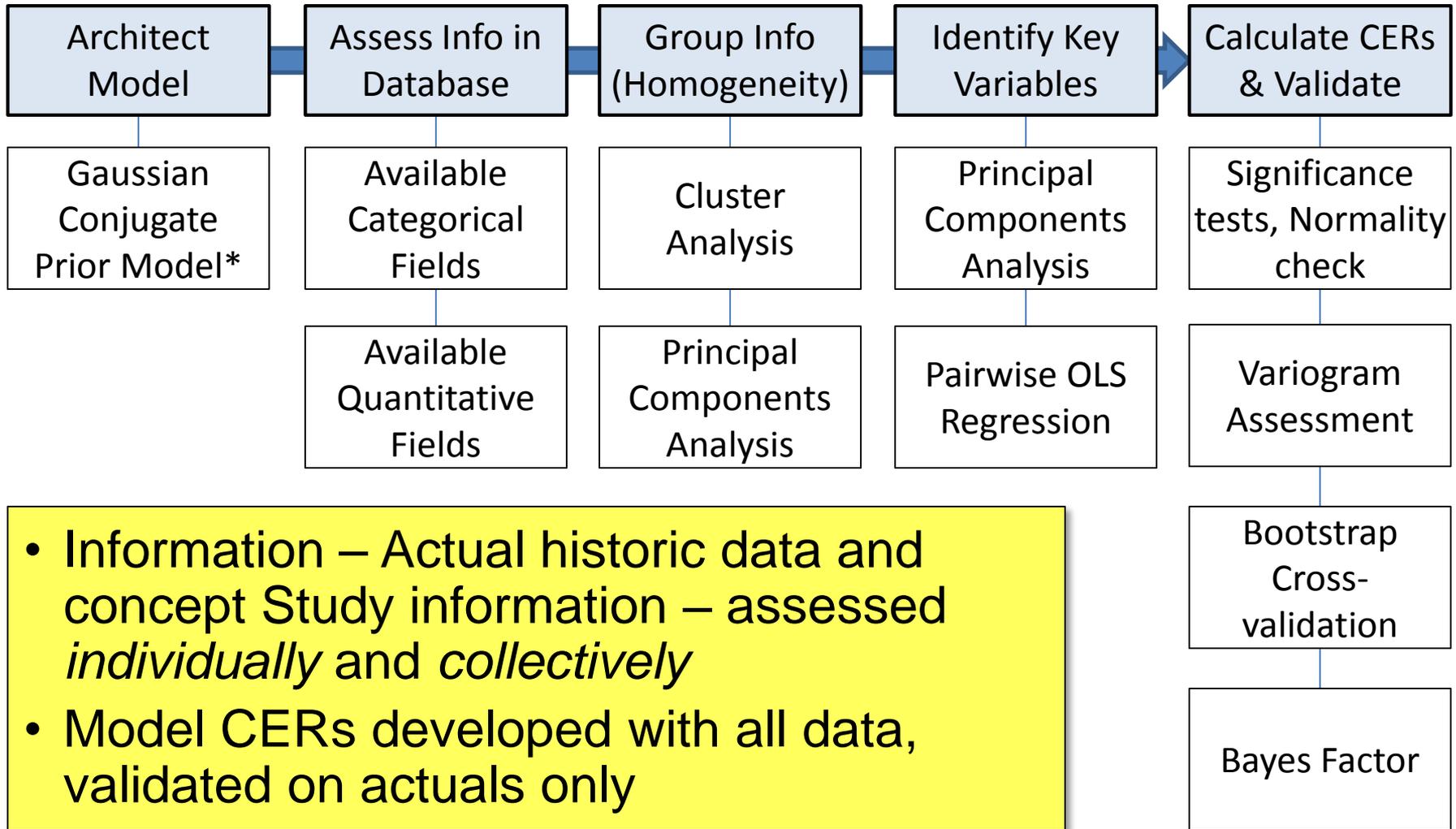
$$p(\beta, \theta | \text{Data}) = p(\text{Data} | \beta, \theta) \times p(\beta, \theta) / p(\text{Data})$$

- Once the data is observed, we can write $p(\text{Data} | \beta, \theta)$ as the “likelihood function”, a function of β, θ , denoted $L(\beta, \theta | \text{Data})$
- $p(\text{Data})$ is a normalizing factor (constant) in the calculations, so, for this presentation we remove it and write:

$$p(\beta, \theta | \text{Data}) \propto L(\beta, \theta | \text{Data}) \times p(\beta, \theta)$$

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

An Example: CER Development Process



- 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	μ (mean vector) and Σ (covariance matrix)	normal-inverse-Wishart	$\mu_0, \kappa_0, \nu_0, \Psi$	$\frac{\kappa_0 \mu_0 + n \bar{x}}{\kappa_0 + n}, \kappa_0 + n, \nu_0 + n,$ $\Psi + C + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{x} - \mu_0)(\bar{x} - \mu_0)^T$ <ul style="list-style-type: none"> • \bar{x} is the sample mean • $C = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$ 	mean was estimated from κ_0 observations with sample mean μ_0 ; covariance matrix was estimated from ν_0 observations with sample mean μ_0 and with sum of pairwise deviation products $\Psi = \nu_0 \Sigma_0$	$t_{\nu_0' - p + 1} \left(\bar{x} \mu_0', \frac{\kappa_0' + 1}{\kappa_0' (\nu_0' - p + 1)} \Psi' \right)^{[5]}$

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



- 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.