Data Fidelity in Formulation

Trevor Beer

Systems Engineering APX, 312C - Systems Modeling, Analysis, and Architectures

June 13, 2024

Jet Propulsion Laboratory
California Institute of Technology

Reviewed and determined not to contain CUI.

© 2024, California Institute of Technology. All Rights Reserved.

About Me

- Completed undergraduate studies at UCSB in Statistics and Data Science
- Graduating with MS in Statistics from UCLA
- Thesis - Generative Data Science: Applications for Early Life Cycle Cost Estimation in the Aerospace Industry
- Joined group in April 2023

The Big Picture

- Group 312C – Systems Modeling, Analysis, and Architectures
- Assist in critical cost estimation tasks during mission formulation activities
- Capabilities range - statistical modeling, data engineering (normalization is a subset of this), web-based tool development
- Dynamic group with members that come from a variety of disciplines – presented many opportunities to broaden my skillset

Reviewed and determined not to contain CUI.

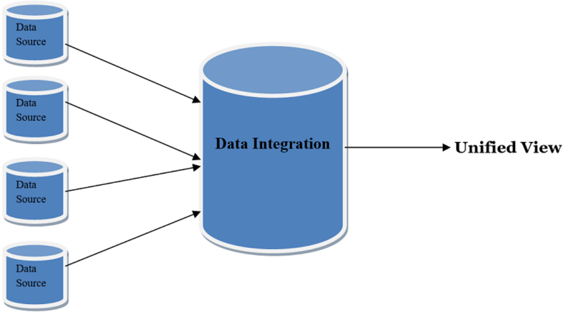
My Contributions to the Group

- Supported, modified and built web-apps using Pythonic software to provide user friendly modeling and data visualization tools
- Harvested, merged and processed raw data into easily ingestible, code-friendly datasets
- Normalized incorrect, outdated and/or inconsistent data into clean datasets to ensure high quality data for our modeling efforts
- Documented all of my code, scripts and work into Github or Jupyter notebook to make all work traceable

Reviewed and determined not to contain CUI.

Relevant Data in Cost Estimation

- Different types of data depending on task
  - FTE (Full-Time Equivalent) – characterizes actual workforce effort
  - Cost – primary variable of interest, includes labor, nonlabor, contributions, and contracts
  - Technical Parameters – by mission, flight element, instrument, include mass, power, data rates
  - Schedule – Stages of lifecycle, viewed in “Phases”
    - Pre-A/A – Study
    - B-D – Development
    - E-F – Operations and Mission Close-Out
- Each comes with unique normalization challenges/constraints



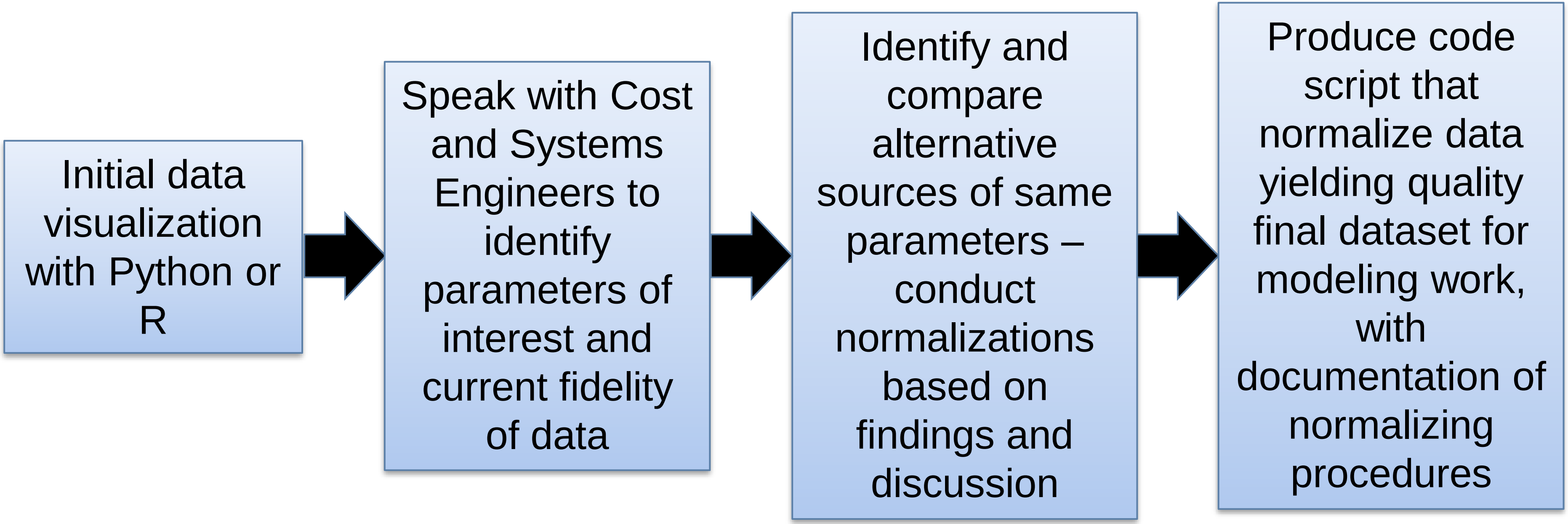
Reviewed and determined not to contain CUI.

Data Fidelity

- Pillar upon which all downstream modeling rests upon
- Relevant questions:
  - Where is data coming from?
  - Does the data agree with other sources? If not, why?
  - What (if any) normalizations need to occur to prepare data for analysis?
- Normalization – refers to process of transforming variables
  - Ex. Consistent Fiscal Year for cost, charge accounts mapped to correct WBS
- Iterative procedure to account for as much uncertainty as possible

Reviewed and determined not to contain CUI.

General Process



Initial data visualization with Python or R

Speak with Cost and Systems Engineers to identify parameters of interest and current fidelity of data

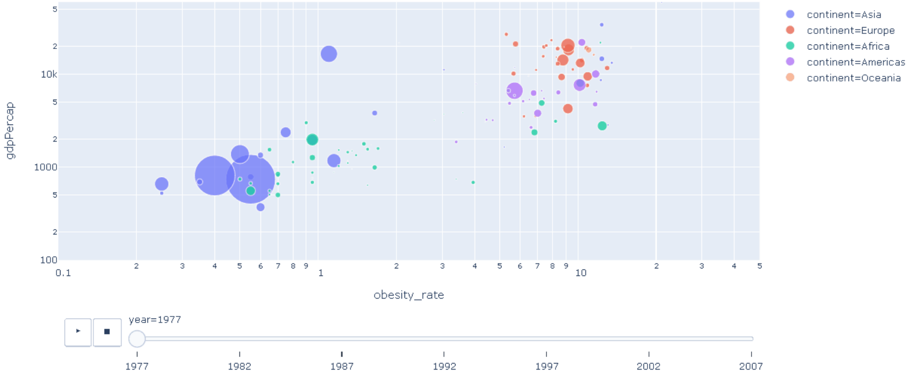
Identify and compare alternative sources of same parameters – conduct normalizations based on findings and discussion

Produce code script that normalize data yielding quality final dataset for modeling work, with documentation of normalizing procedures

Reviewed and determined not to contain CUI.

Processing Tools

- I focused on Python, particularly the Pandas and Plotly packages for data frame manipulation and visualization creation
- Plotly – allows for interactive visuals, elevating users’ ability to control their analysis experience
- Both packages scale well with large datasets in terms of visualization and normalization capabilities



Reviewed and determined not to contain CUI.

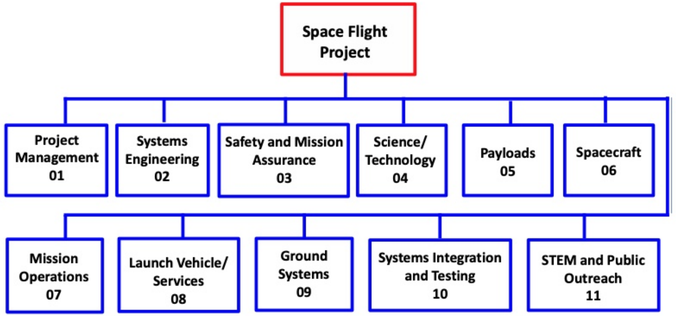
Examples of Sources

- Oracle Database – Provides granular workforce records, monthly FTE’s and cost for various projects
- Work Definition – Project level records and schedule
- CADB – Project level costs by Fiscal year, with information about project phases and WBS mappings
- ZAP – 312C Database intended to serve as warehouse for “gold standard” data based on normalizations conducted
- My work consisted of gathering, merging, and pre-processing data from these sources to generate high-quality datasets for modeling tasks

Reviewed and determined not to contain CUI.

WBS Normalizations

- Work Breakdown Structure provides common framework for projects to bucket work by major mission components
- JPL uses NASA Standardized WBS, with addition of WBS 12 – Mission Design Navigation (MD Nav)
- Older records may be missing, incomplete, or reflect dated WBS standards which make interpreting through present WBS lens more difficult
- Normalizing Procedures:
  - CADB – Records from 2X that define WBS mappings and account for changes in bookkeeping over time
  - Identify keywords, talk to personnel from projects "grassroots normalization"



Reviewed and determined not to contain CUI.

FTE/Cost Normalizations

- FTEs – Monthly or yearly, must determine how to display calculations for interpretable results
- Cost – Standard to inflate to current fiscal year
  - Labor rates do not change 1-1 with inflation, meaning simply inflating may not be sufficient to compare costs temporally
- FTEs are best indicator of raw workforce efforts, but ultimately cost is the variable of interest in most analyses
- Analyzing actuals from formulation lens is difficult:
  - How to back out reserves?
  - Account for slips or other events not typically considered in early life cycle

ICM Modeling Background

- ICM Model – Institutional Cost Model is a section approved model that serves to quantify cost based on general parameters that guide mission development in formulation
  - Such as planned schedule, anticipated FTE breakouts, and broad technical considerations like the intended target
- Must be signed off by institutional manager – imperative that the calculations and results output represent our best and most accurate efforts to provide high-fidelity cost estimates
- My efforts focused on the model to build WBS 05 and 06 estimates, which correspond to Instrument and Spacecraft management and systems engineering

Reviewed and determined not to contain CUI.

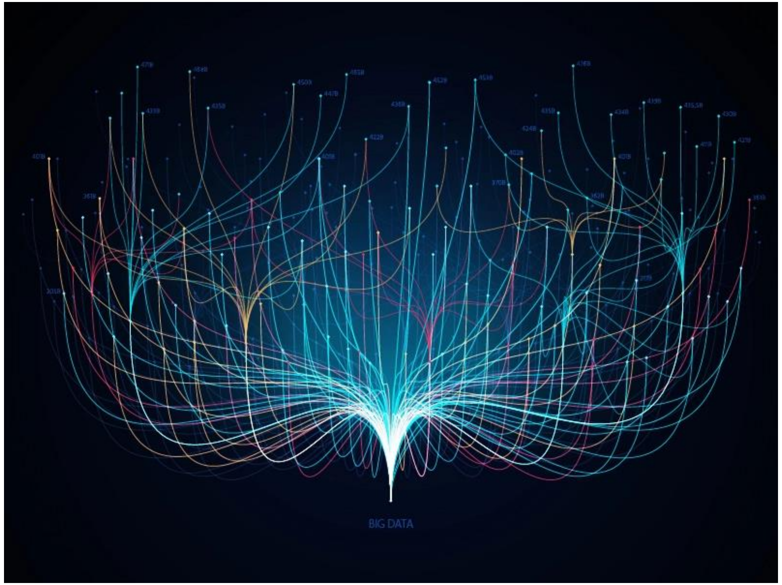
ICM Modeling Background

- Speaking to project leads informs us that bookkeeping processes are not perfect during development as indicated in Oracle, so the CADB exists in order to identify and amend bookkeeping errors after they are initially recorded
- This process is critical because the ICM Model must be informed by actual data, and incorrect bookkeeping could lead to over or underestimation of the target WBS

Reviewed and determined not to contain CUI.

ICM Modeling Background

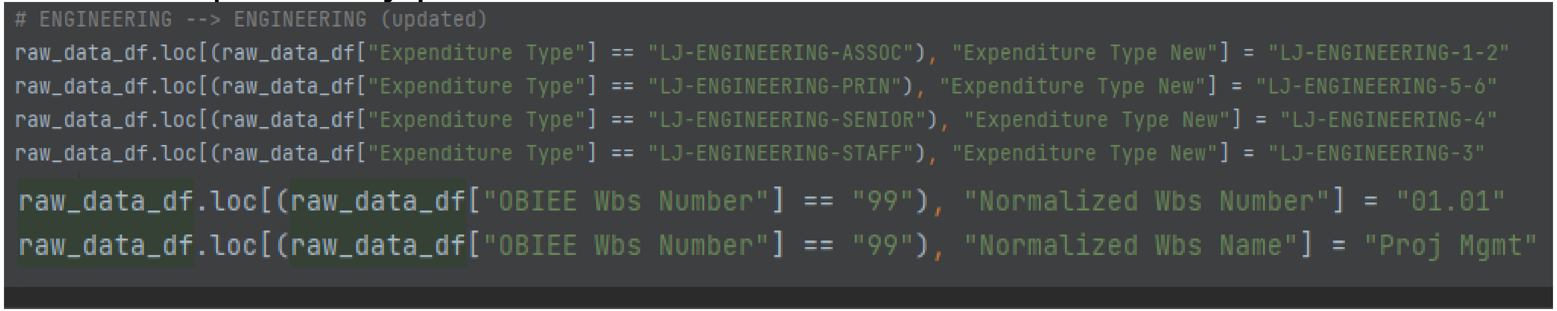
- I conducted various normalization procedures on the raw data of over 1.6 million entries, including:
  - Merging CADB mapping of WBS to raw Oracle data to correct known bookkeeping errors
  - Manual WBS normalization based on WBS keywords (such as Payload SE, Flight System SE, etc) to catch unknown mappings
  - Merged mission level characteristics such as Cost Category to allow visualizing groups of missions easily
  - Mapping entries to the proper phase by Project Number and date with CADB and Work Definition information



Reviewed and determined not to contain CUI.

ICM Modeling Example

- The procedures from the previous slide were conducted on both Cost and FTE data to allow further analysis of labor rate calculations and assist in model validation
- I created a Python script to automate as much of this process as possible, merging the unique relational data frames and generating an output CSV that is ready for immediate analysis
- Results stemming from validation with data I normalized serve as key motivator to update model in order to best reflect how JPL conducts business in higher level of detail than previously possible



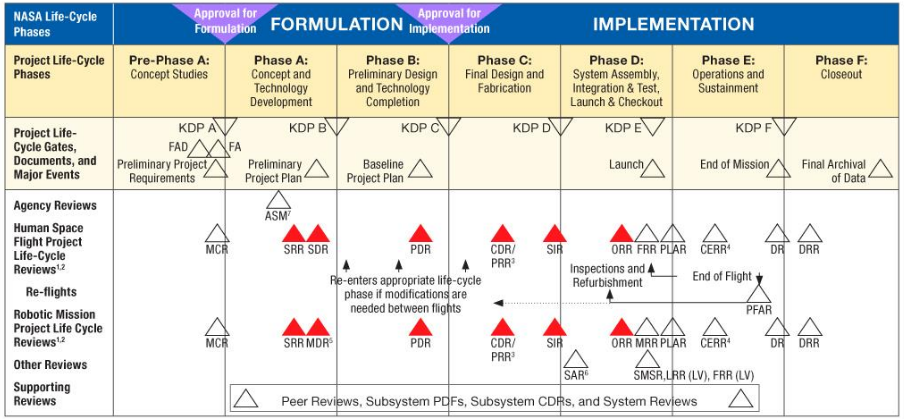
```
# ENGINEERING --> ENGINEERING (updated)
raw_data_df.loc[(raw_data_df["Expenditure Type"] == "LJ-ENGINEERING-ASSOC"), "Expenditure Type New"] = "LJ-ENGINEERING-1-2"
raw_data_df.loc[(raw_data_df["Expenditure Type"] == "LJ-ENGINEERING-PRIN"), "Expenditure Type New"] = "LJ-ENGINEERING-5-6"
raw_data_df.loc[(raw_data_df["Expenditure Type"] == "LJ-ENGINEERING-SENIOR"), "Expenditure Type New"] = "LJ-ENGINEERING-4"
raw_data_df.loc[(raw_data_df["Expenditure Type"] == "LJ-ENGINEERING-STAFF"), "Expenditure Type New"] = "LJ-ENGINEERING-3"

raw_data_df.loc[(raw_data_df["OBIEE Wbs Number"] == "99"), "Normalized Wbs Number"] = "01.01"
raw_data_df.loc[(raw_data_df["OBIEE Wbs Number"] == "99"), "Normalized Wbs Name"] = "Proj Mgmt"
```

Reviewed and determined not to contain CUI.

Schedule Normalizations

- Schedule and Phases are projected at different checkpoints, such as PDR, SIR, and EoM
- Differences in schedules based on various factors – launch windows, internal/external slips, funding availability
- Which form of schedule will best inform models?


NASA Life-Cycle Phases
Approval for Formulation
FORMULATION
Approval for Implementation
IMPLEMENTATION
Project Life-Cycle Phases
Pre-Phase A: Concept Studies
Phase A: Concept and Technology Development
Phase B: Preliminary Design and Technology Completion
Phase C: Final Design and Fabrication
Phase D: System Assembly, Integration & Test, Launch & Checkout
Phase E: Operations and Sustainment
Phase F: Closeout
Project Life-Cycle Gates, Documents, and Major Events
KDP A
KDP B
KDP C
KDP D
KDP E
KDP F
FAD
FA
Preliminary Project Requirements
Preliminary Project Plan
Baseline Project Plan
Launch
End of Mission
Final Archival of Data
Agency Reviews
ASM[7]
Human Space Flight Project Life-Cycle Reviews[1,2]
MCR
SRR
SDR
PDR
CDR/PRR[3]
SIR
ORR
FRR
PLAR
CERR[4]
DR
DRR
Re-flights
Re-enters appropriate life-cycle phase if modifications are needed between flights
Inspections and Refurbishment
End of Flight
PFAR
Robotic Mission Project Life Cycle Reviews[1,2]
MCR
SRR
MDR[5]
PDR
CDR/PRR[3]
SIR
ORR
MRR
PLAR
CERR[4]
DR
DRR
Other Reviews
SAR[6]
SMSR, LRR (LV), FRR (LV)
Supporting Reviews
Peer Reviews, Subsystem PDFs, Subsystem CDRs, and System Reviews

FOOTNOTES

1. Flexibility is allowed as to the timing, number, and content of reviews as long as the equivalent information is provided at each KDP and the approach is fully documented in the Project Plan.
2. Life-cycle review objectives and expected maturity states for these reviews and the attendant KDPs are contained in Table 2-5 and Appendix D Table D-3 of this handbook
3. PRR is needed only when there are multiple copies of systems. It does not require an SRB. Timing is notional.
4. CERRs are established at the discretion of program .
5. For robotic missions, the SRR and the MDR may be combined.
6. SAR generally applies to human space flight.
7. Timing of the ASM is determined by the MDAA. It may take place at any time during Phase A.

▲ Red triangles represent life-cycle reviews that require SRBs. The Decision Authority, Administrator, MDAA, or Center Director may request the SRB to conduct other reviews.

ACRONYMS

ASM – Acquisition Strategy Meeting
CDR – Critical Design Review
CERR – Critical Events Readiness Review
DR – Decommissioning Review
DRR – Disposal Readiness Review
FA – Formulation Agreement
FAD – Formulation Authorization Document
FRR – Flight Readiness Review
KDP – Key Decision Point
LRR – Launch Readiness Review
LV – Launch Vehicle
MCR – Mission Concept Review
MDR – Mission Definition Review
MRR – Mission Readiness Review
ORR – Operational Readiness Review
PDR – Preliminary Design Review
PFAR – Post-Flight Assessment Review
PLAR – Post-Launch Assessment Review
PRR – Production Readiness Review
SAR – System Acceptance Review
SDR – System Definition Review
SIR – System Integration Review
SMSR – Safety and Mission Success Review
SRB – Standing Review Board
SRR – System Requirements Review

FIGURE 3.0-1 NASA Space Flight Project Life Cycle from NPR 7120.5E

Reviewed and determined not to contain CUI.

Schedule Normalization Example

- ICM model uses phase lengths to build mission profile to better understand how time drives cost
- Phases B, C, and D (development) correspond to the highest degree of work on Payload and Spacecraft components
- I aided in efforts to document these schedule discrepancies from plans and actuals and identify rules of thumb to breakout sub-schedule durations that are used in mission planning
- Identified source of schedule slips (internal or external) and removed months pertaining the delays caused by non-JPL factors
- Delivered processed and documented schedule data to incorporate in ZAP, the 312C visualization and data warehouse

Reviewed and determined not to contain CUI.

Technical Data Normalizations

- May wish to utilize technical parameters from flown missions and studies to deduce cost estimating relationships (CERs)
- Parameters may be given on instrument, flight element, or mission basis
- Studies can include multiple design scenarios with different values – must choose what to accept for analysis

Reviewed and determined not to contain CUI.

Technical Data Normalizations Example

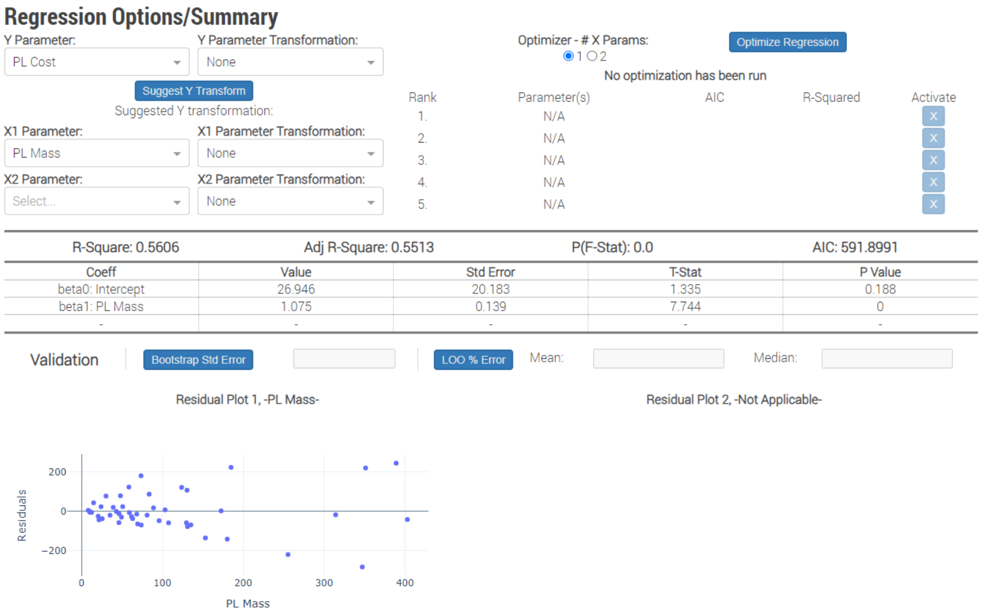
- MCDB – Mission and Cost Database
- MCDB task aims to unify actual flown data and study to create comprehensive “gold standard” technical data for missions, flight elements, and instruments
- I read and documented study reports and workbooks to intimately understand mission objectives and identify parameters of interest
- Provided data to personnel responsible for updating database

Mission	Mass (kg)	Power (kW)
A	500	26
B	2000	45
C	1500	30

Reviewed and determined not to contain CUI.

Technical Data Normalizations Example

- Technical data is used to model relationships between physical parameters and cost, such as in the A-Team Regression (ART) tool
- Quantifying these relationships with regression works well within our generally low-data paradigm (more complex models may overfit)
- Allows the user to explore strength of different CERs, such as mass vs cost, data rate vs cost



Regression Options/Summary

Y Parameter: PL Cost

Y Parameter Transformation: None

Suggested Y transformation:

X1 Parameter: PL Mass

X1 Parameter Transformation: None

X2 Parameter: Select...

X2 Parameter Transformation: None

Optimizer - # X Params: ● 1 ○ 2

No optimization has been run

Rank	Parameter(s)	AIC	R-Squared	Activate
1.	N/A			X
2.	N/A			X
3.	N/A			X
4.	N/A			X
5.	N/A			X

R-Square: 0.5606 Adj R-Square: 0.5513 P(F-Stat): 0.0 AIC: 591.8991

Coeff	Value	Std Error	T-Stat	P Value
beta0: Intercept	26.946	20.183	1.335	0.188
beta1: PL Mass	1.075	0.139	7.744	0
-	-	-	-	-

Validation

Mean:

Median:

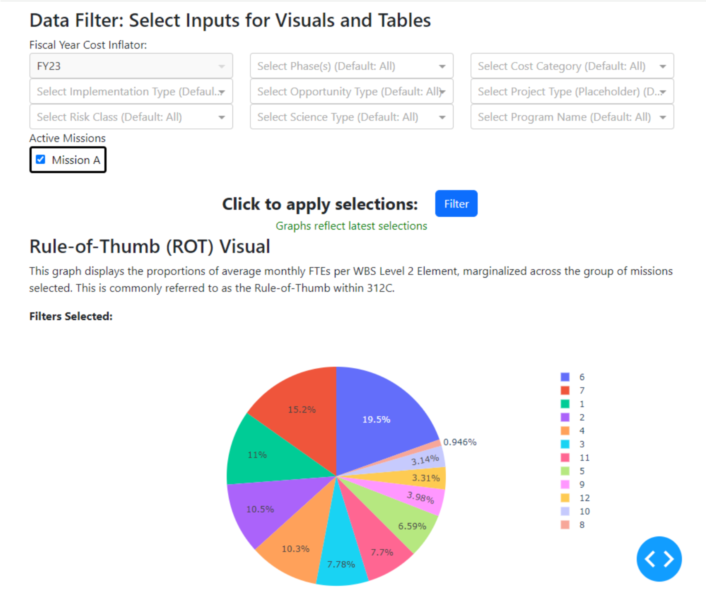
Residual Plot 1, -PL Mass-

Residual Plot 2, -Not Applicable-

Reviewed and determined not to contain CUI.

Workforce Analysis Visualization and Exploration Tool

- Otherwise known as WAVE, part of 312C Suite of tools
  - https://costing.jpl.nasa.gov/dash/wave/
- I created using Dash – a Pythonic framework for creating powerful data science oriented web apps
- Integrates elements of WBS, FTE, Cost, and Schedule normalizations I conducted in backend data with simple filters and user interface on frontend
  - This data is local right now, with plans to integrate backend directly to ZAP
- Use case: quickly identify most prevalent expenditure categories in WBS 05 and 06 to inform ICM model



Data Filter: Select Inputs for Visuals and Tables

Fiscal Year Cost Inflator: FY23

Active Missions

☑ Mission A

Click to apply selections:

Graphs reflect latest selections

Rule-of-Thumb (ROT) Visual

This graph displays the proportions of average monthly FTEs per WBS Level 2 Element, marginalized across the group of missions selected. This is commonly referred to as the Rule-of-Thumb within 312C.

Filters Selected:

Select Phase(s) (Default: All)
Select Cost Category (Default: All)
Select Implementation Type (Defaul...
Select Opportunity Type (Default: All)
Select Project Type (Placeholder) (D...
Select Risk Class (Default: All)
Select Science Type (Default: All)
Select Program Name (Default: All)
Filter
19.5%
15.2%
11%
10.5%
10.3%
7.78%
7.7%
6.59%
3.98%
3.31%
3.14%
0.946%
6
7
1
2
4
3
11
5
9
12
10
8

Reviewed and determined not to contain CUI.

Summary

- Led data normalization procedures for data used in various critical tasks including ICM Model updates and MCDB data collection
- Expanded the capabilities of 312C web presence with the WAVE tool
- Created scripts and templates in Python and R to provide standardized and traceable normalization procedures on various data, namely cost, FTE, schedule, and technical parameters
- Learned about fundamental JPL engineering and business practices and integrated lessons from cost and systems engineers into data normalization process

Reviewed and determined not to contain CUI.

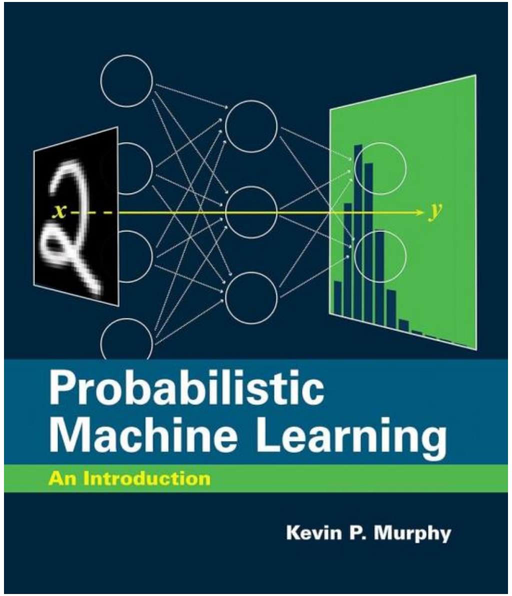
Acknowledgements

- Michael Saing and Anto Kolanjian
  - Served as my primary mentors for the duration of my internship
  - Provided invaluable domain knowledge, guidance on tasks, and freedom to explore projects of interest
- Patrick Bjornstad and Patricia Gallagher
  - Collaborated with on a number of projects including web development, ICM, DAG, MCDB

Reviewed and determined not to contain CUI.

Additional Acknowledgements

- Weekly Math/Stats reading group – engaging discussions on statistical/ML techniques with eye towards applications in our daily work
- Weekly Tag-Ups – presented opportunity to engage in discussions surrounding other activities performed by group members and understand broader scope of group impact



Reviewed and determined not to contain CUI.

Thank You!

Reviewed and determined not to contain CUI.

Jet Propulsion Laboratory
California Institute of Technology

jpl.nasa.gov