The Silent *S* in NICM: NICM Schedule Capabilities

Presenters:
Joe Mrozinski
Mike DiNicola

NASA Cost Symposium, August 2017

Jet Propulsion Laboratory
California Institute of Technology

NICM Team

- Gary Ball
- Luther Beegle
- Justin Boland
- Bob Cesarone
- Mike DiNicola
- Michael Fong
- Hamid Habib-Agahi
- Joe Mrozinski
- Al Nash
- Michael Saing
- Marc Walch

What is NICM?

- NASA Instrument Cost Model
  - Probabilistic Cost Estimates for Space Flight Instruments
  - Used by all NASA Centers
    - And any organization proposing instruments for NASA Instruments
    - And proposal evaluators
  - Version I Released in 2007
  - Version VII Rev 2 Released 2016

What is NICM?

- NICM also:
  - Estimates schedule
  - Estimates cost and schedule phase breakdowns
  - Supports JCL
  - Contains an normalized instrument database (for civil servants)

Yes – you can get a copy of NICM

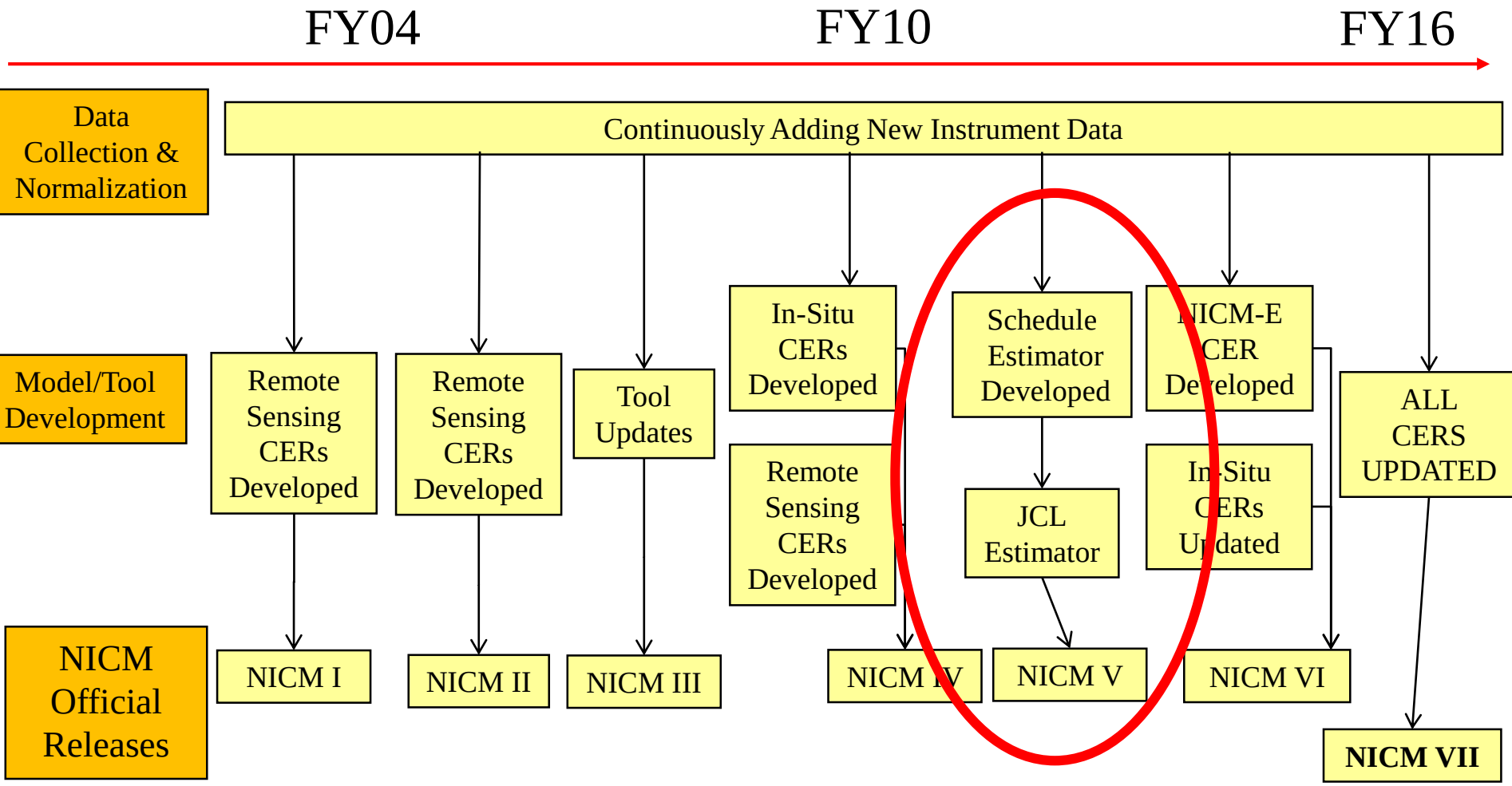
- RSVP for only training at:

Joseph.J.Mrozinski@jpl.nasa.gov

Just kidding, you'll never remember that

NICM@jpl.nasa.gov

Today's Story: Schedule Estimation



NICM Schedule Estimation Relationships (SERs)

Mike DiNicola

NASA Cost Symposium, August 2017

Agenda

- Data Exploration:
  - Histograms, Box Plots
  - Cluster Analysis
  - Principal Components Analysis (PCA)
- Draft SERs for NICM VIII
- Future Work
- Feedback
  - Especially from our new schedule friends!

*In what follows:

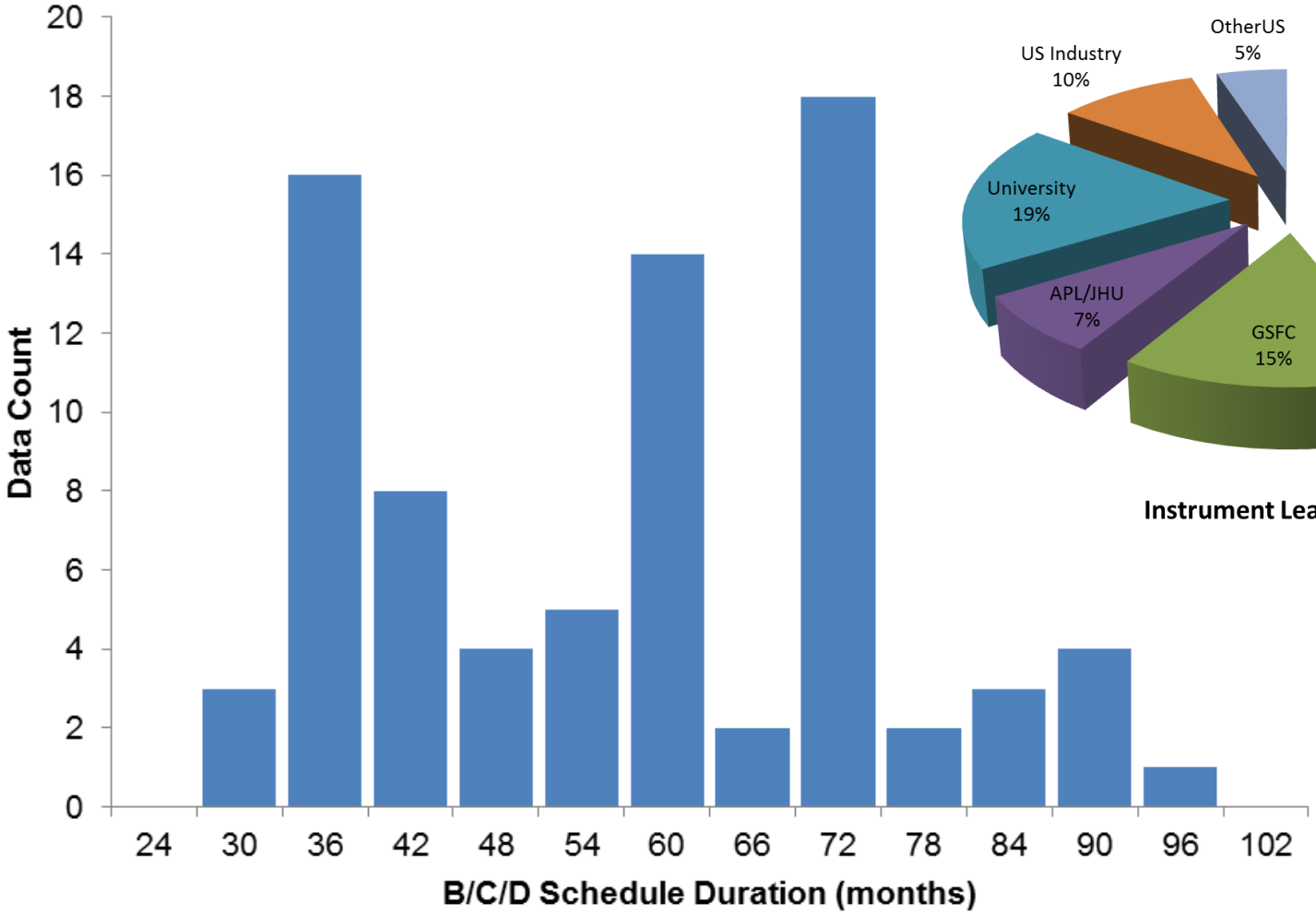
TIC = Total Instrument B/C/D Cost (FY04 $M)

Agenda

- **Data Exploration:**
  - Histograms, Box Plots
  - Cluster Analysis
  - Principal Components Analysis (PCA)
- Draft SERs for NICM VIII
- Future Work
- Feedback
  - Especially from our new schedule friends!

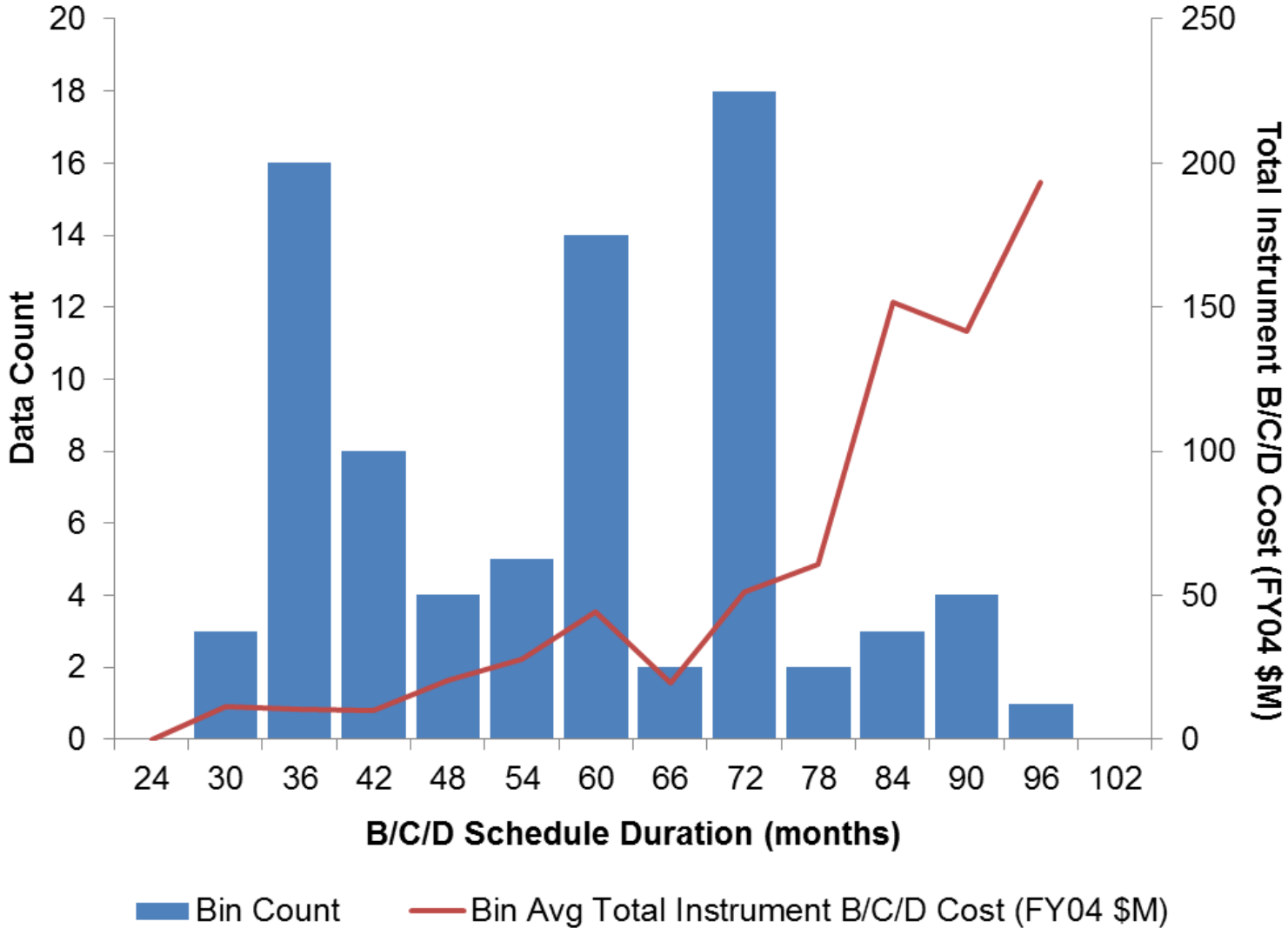
*In what follows:
TIC = Total Instrument B/C/D Cost (FY04 $M)

80 Data Points analyzed from across the NASA Community



Instrument Lead Organization

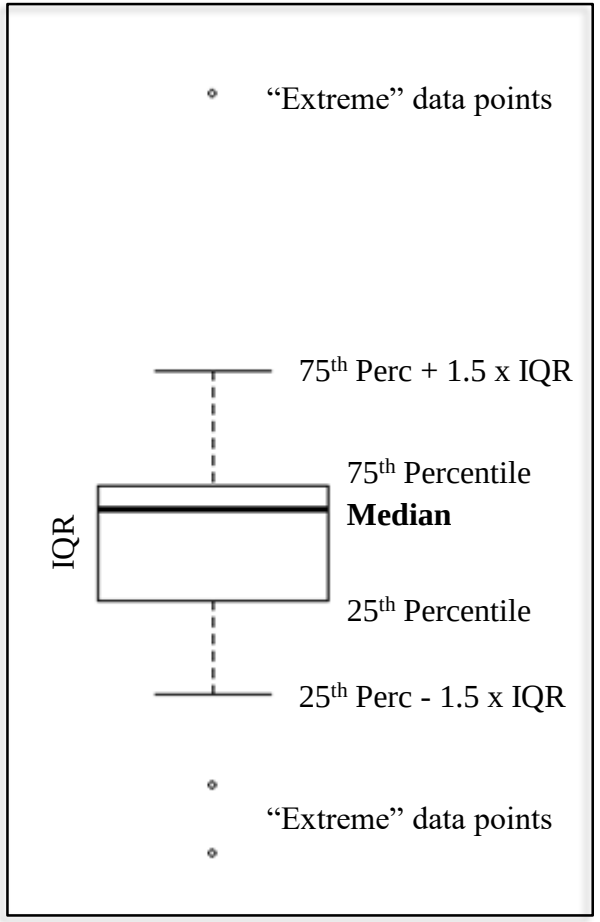
Cost & Schedule are Correlated



Using Box Plots to Find Important Attributes in Data

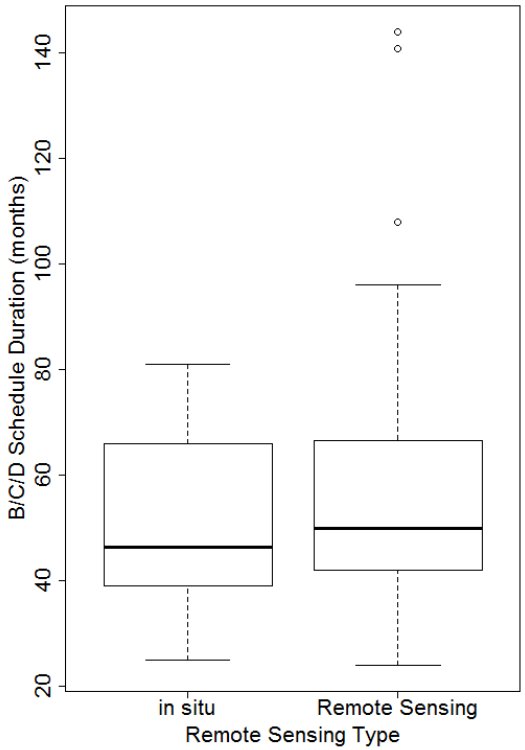
- Great way to find important categorical attributes in the data
  - Box plots give a feel for the distributions in the data without having to make any assumption on the distributional form
  - Informs regression analysis

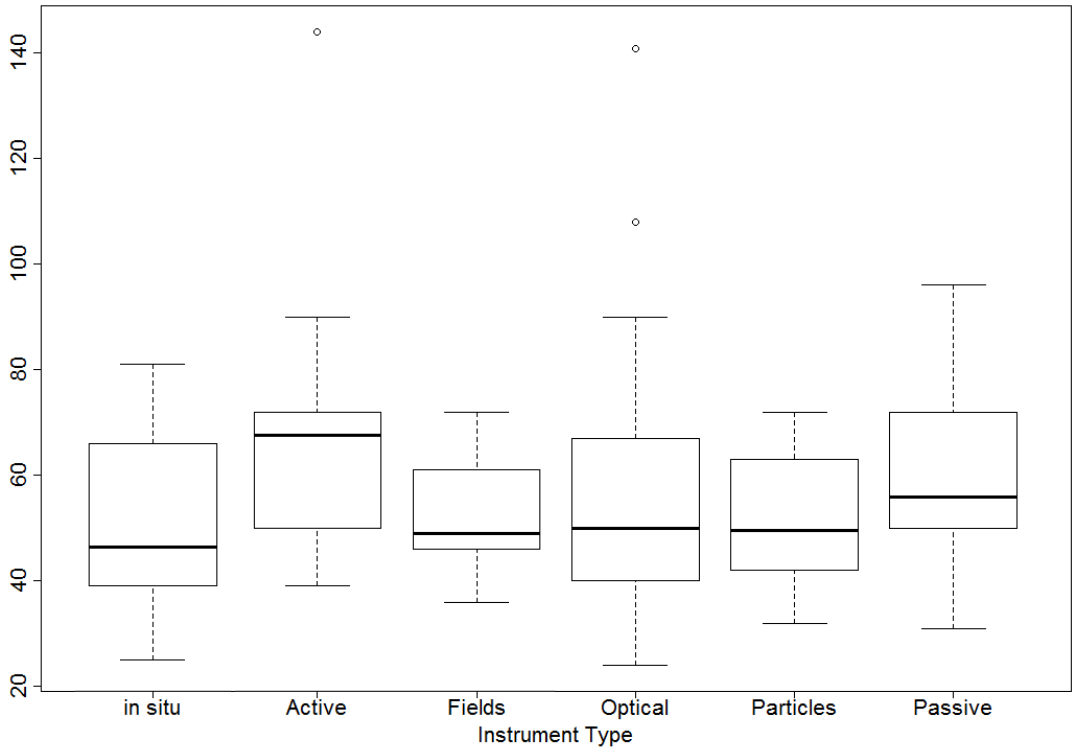
- Box plots are defined according to the picture to the right
  - IQR = Inter-Quartile Range = 75th percentile minus 25th percentile



Schedule Duration by Instrument & Sensing Type

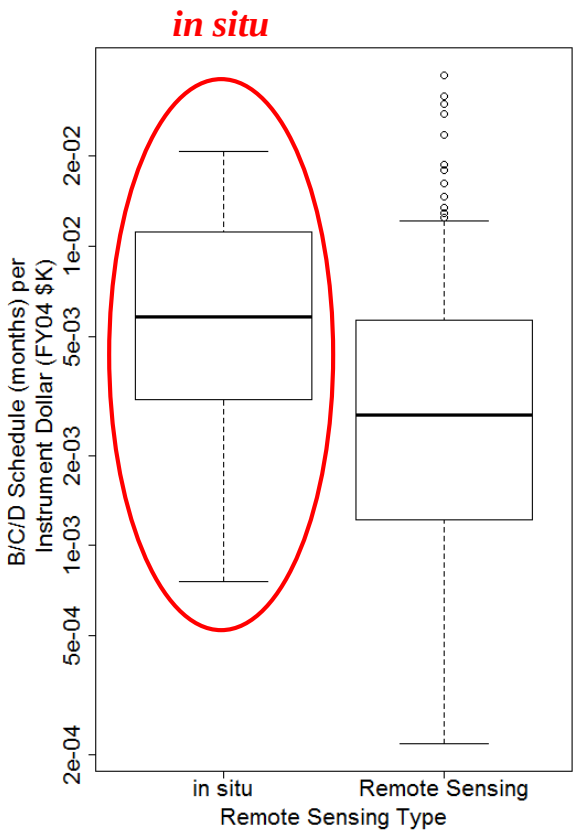
Data by instrument and sensing type do not look much different when looking at *absolute* schedule duration…

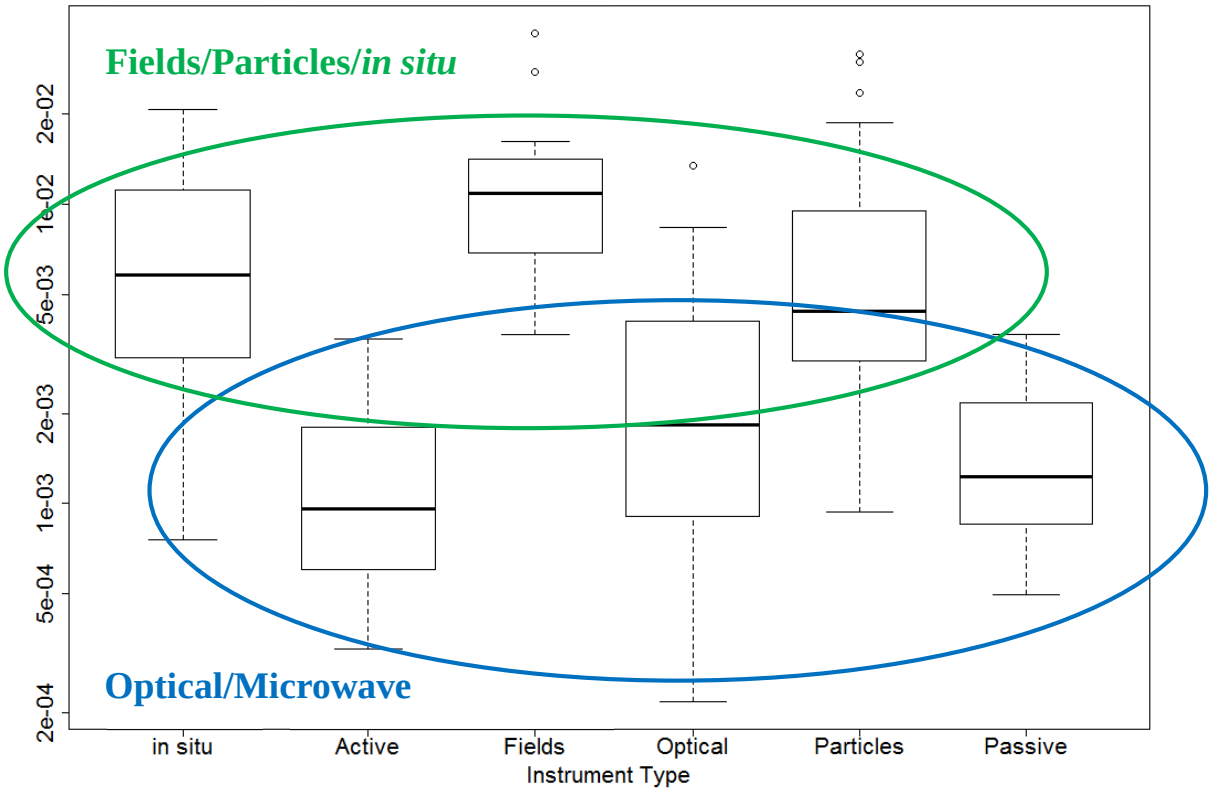




Schedule per Dollar by Instrument & Sensing Type

…but Schedule Duration **per Dollar** shows potential group differentiation for *in situ*, Fields/Particles and Optical/Microwave instruments

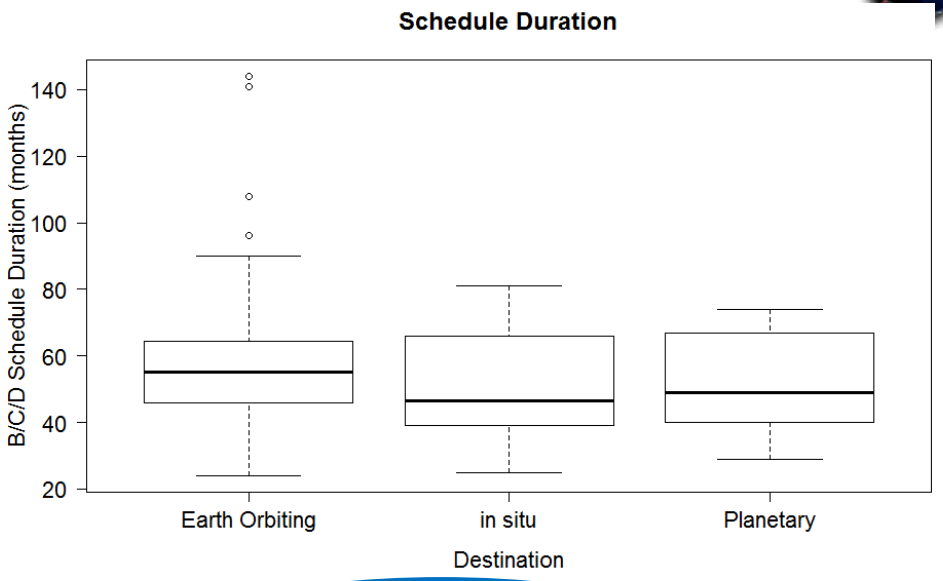


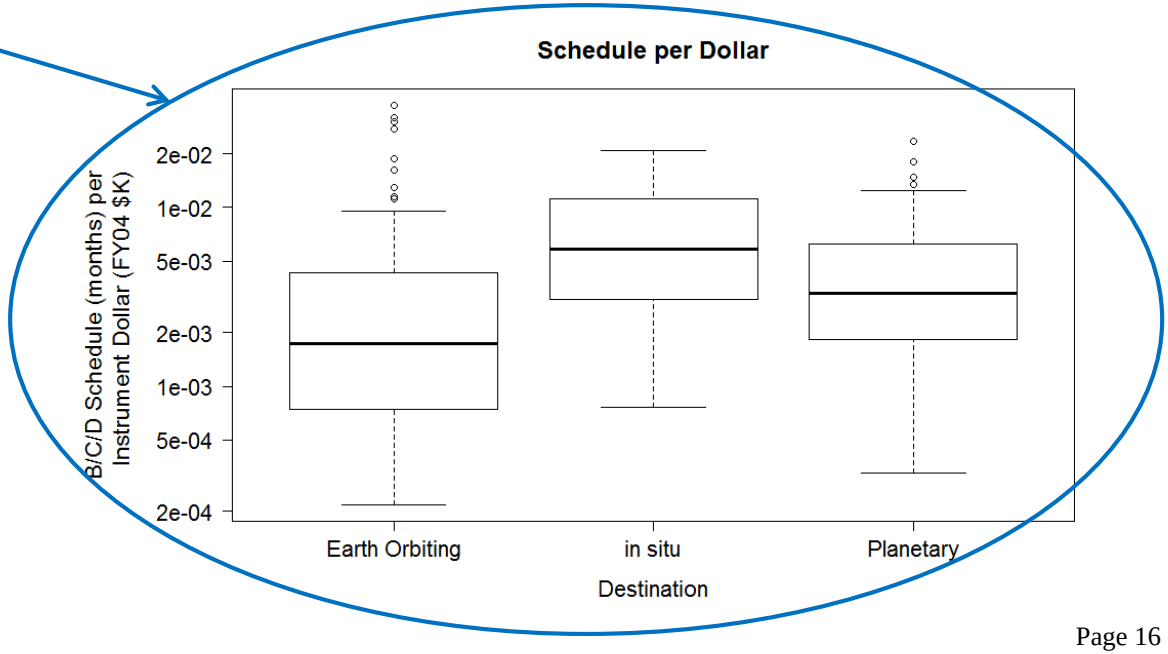


Schedule Metrics by Destination Type

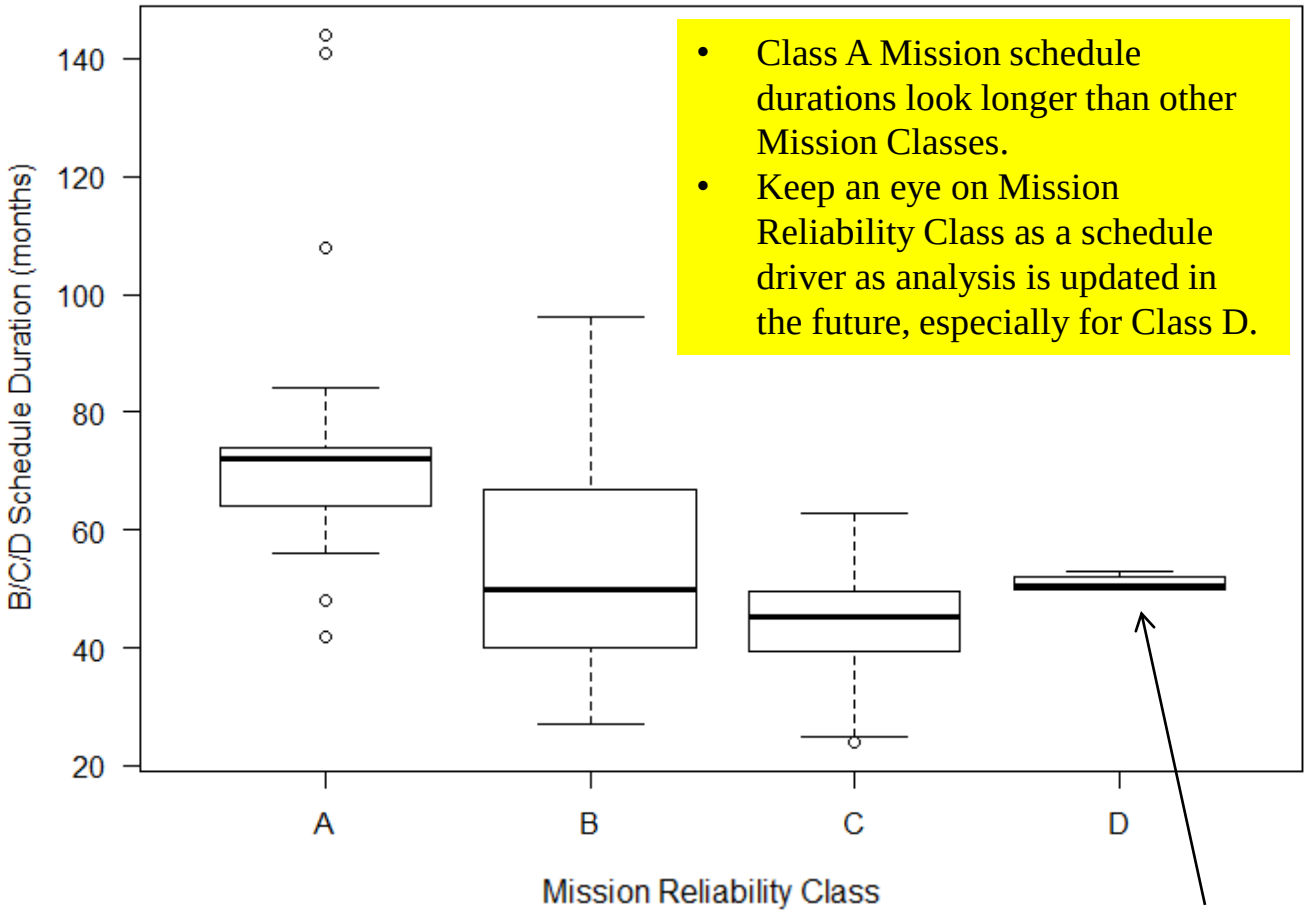
Similar to the previous slide, schedule duration per dollar shows potential groups for:

- Earth Orbiting
- Planetary Remote Sensing
- Planetary *in situ*





Schedule Duration by Mission Reliability Class

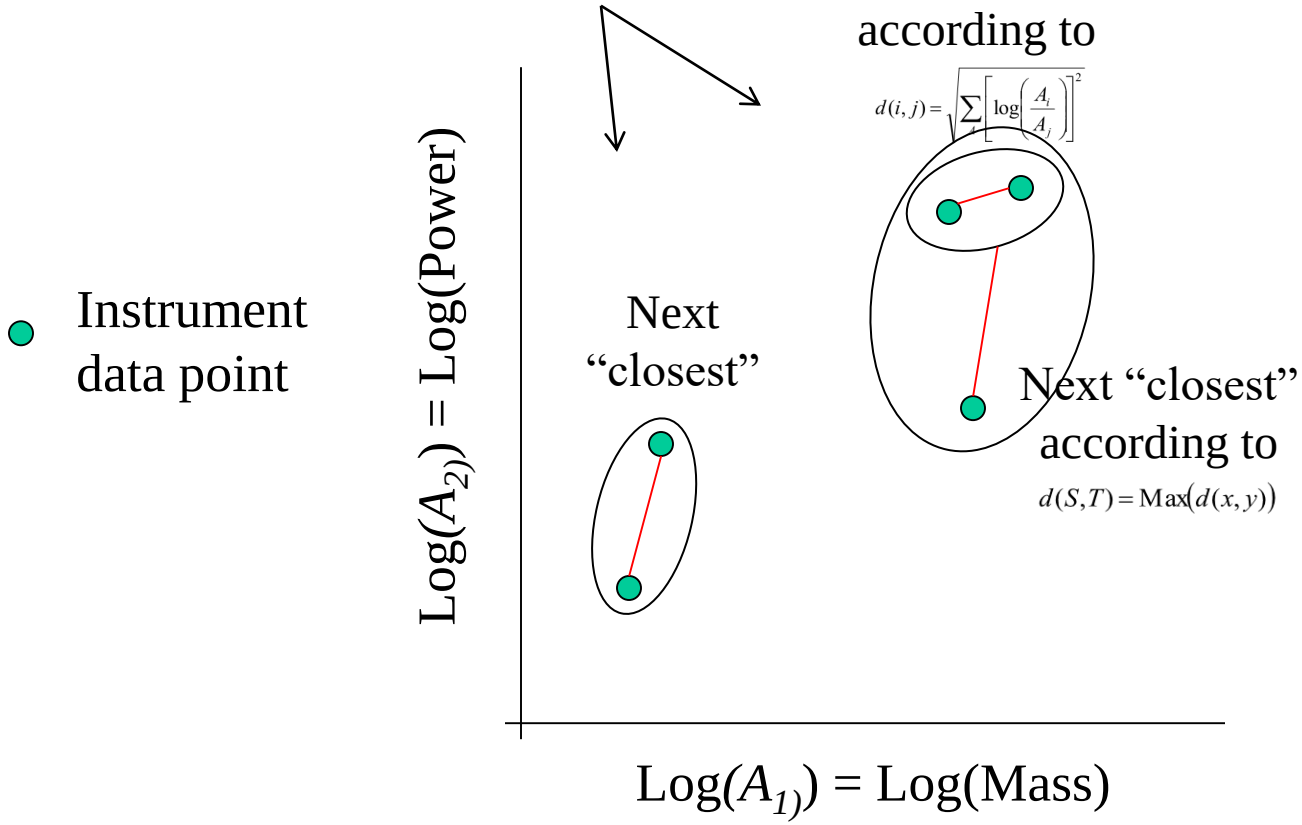


- Class A Mission schedule durations look longer than other Mission Classes.
- Keep an eye on Mission Reliability Class as a schedule driver as analysis is updated in the future, especially for Class D.

When LCROSS data are included (36 month development schedule), there will be a more significant difference between Classes C and D

Cluster Analysis – Quick Example

We end up with two main clusters.
One of these has a "sub-cluster".



Note that we have distances between points and clusters. How do we tell distance between these?

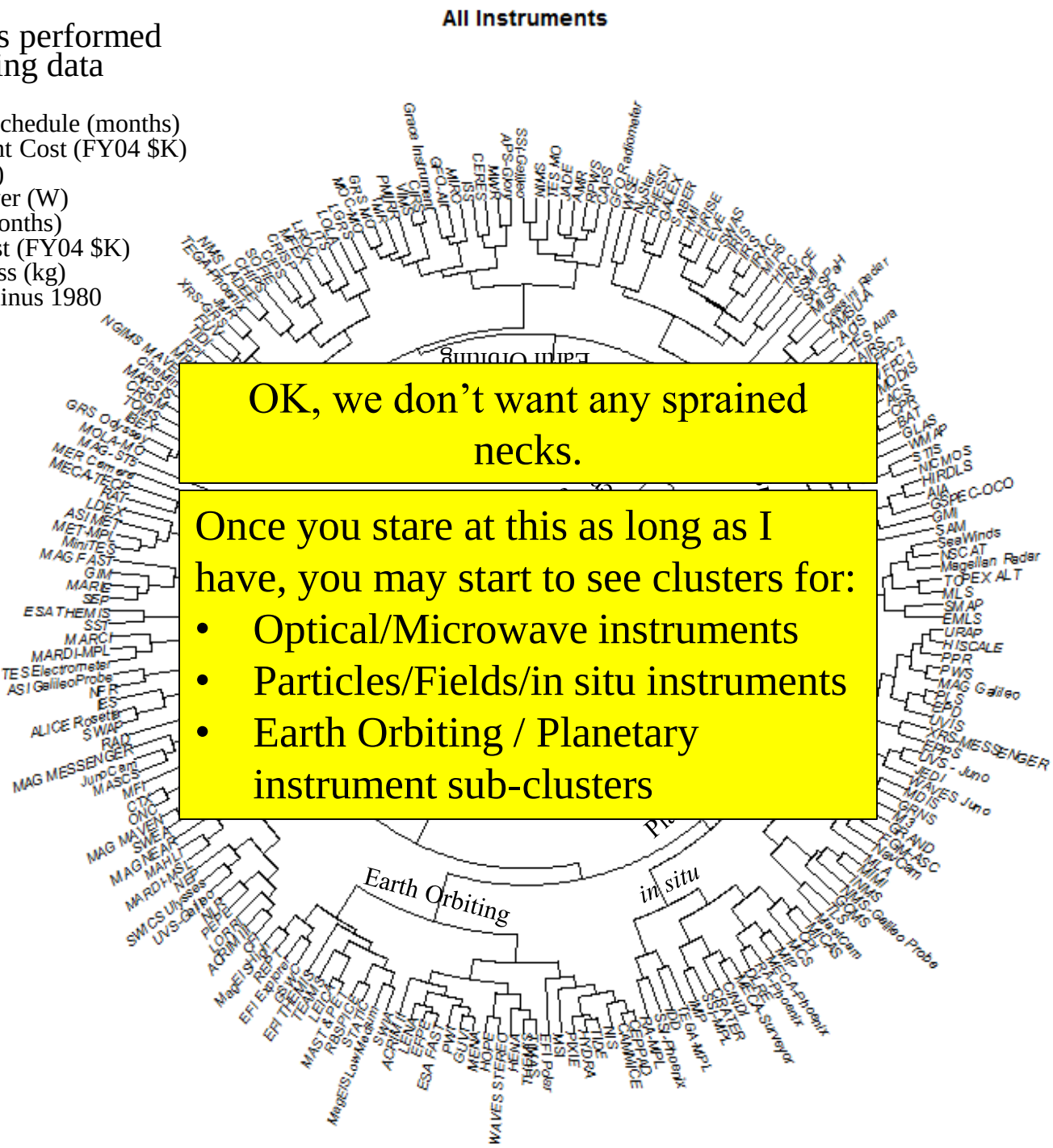
$$d(S,T) = \mathrm{Max}(d(x,y))$$

where S and T are clusters (maybe with only one element), x and y are points in the clusters

You may have been able to see that from the beginning, but what if you had to deal with many more variables and much more data? (beware: "eye chart" to follow)

Cluster Analysis performed used the following data attributes:

- Development Schedule (months)
- Total Instrument Cost (FY04 $K)
- Total Mass (kg)
- Total Max Power (W)
- Design Life (months)
- Electronics Cost (FY04 $K)
- Electronics Mass (kg)
- Launch Year minus 1980

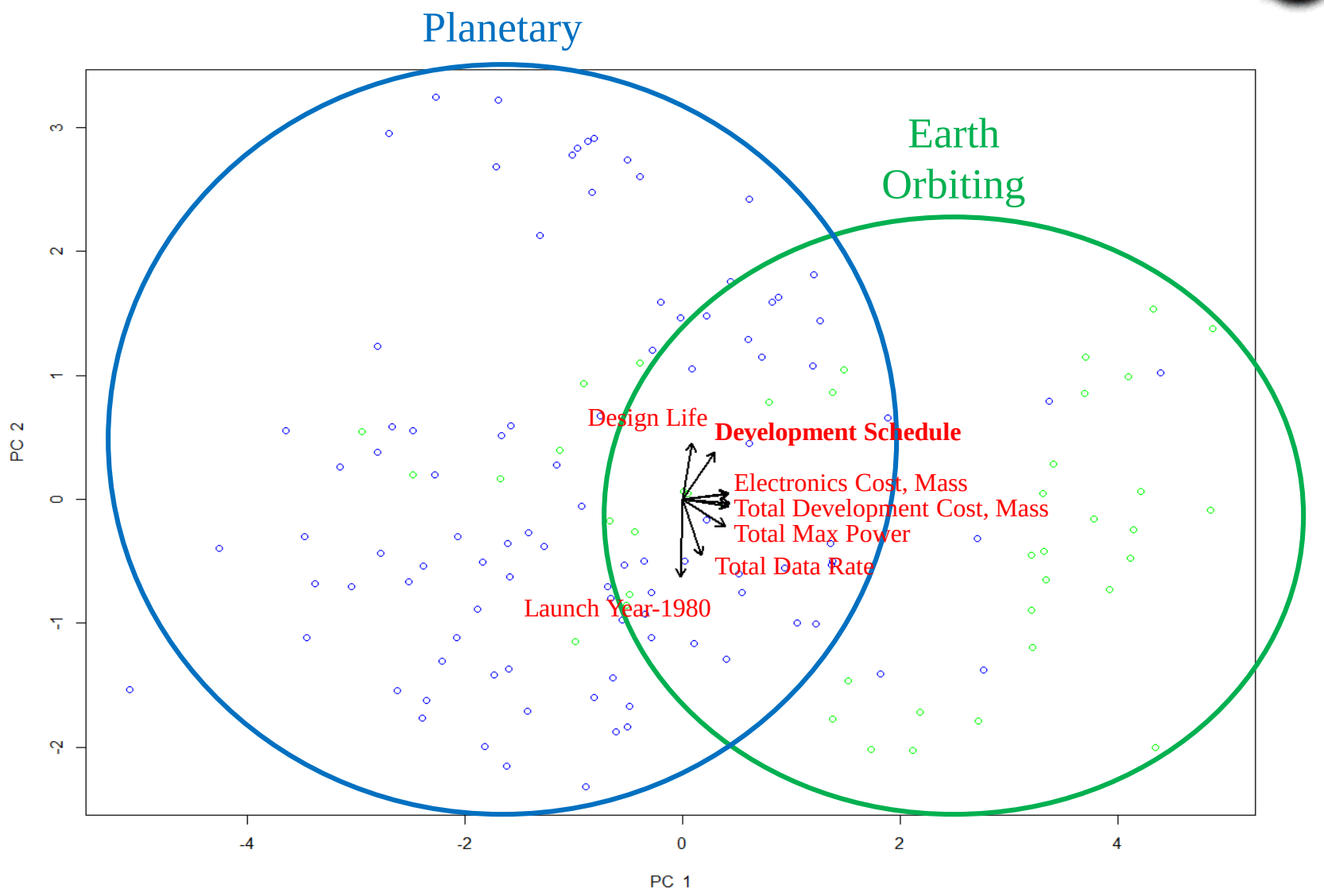


Principal Components Analysis (PCA)

- What if we could somehow look at all variables at once and determine how they are correlated?
  - Specifically, what is correlated with schedule duration?
- What if we could identify combinations of variables that explain the most variation in the data
  - This could help us develop a regression relationship
- What if we saw the data projected onto the primary sources of variation in the data?
  - This is another way to see how our data might be clustering
  - Different than the previous clustering technique because it factors in correlation

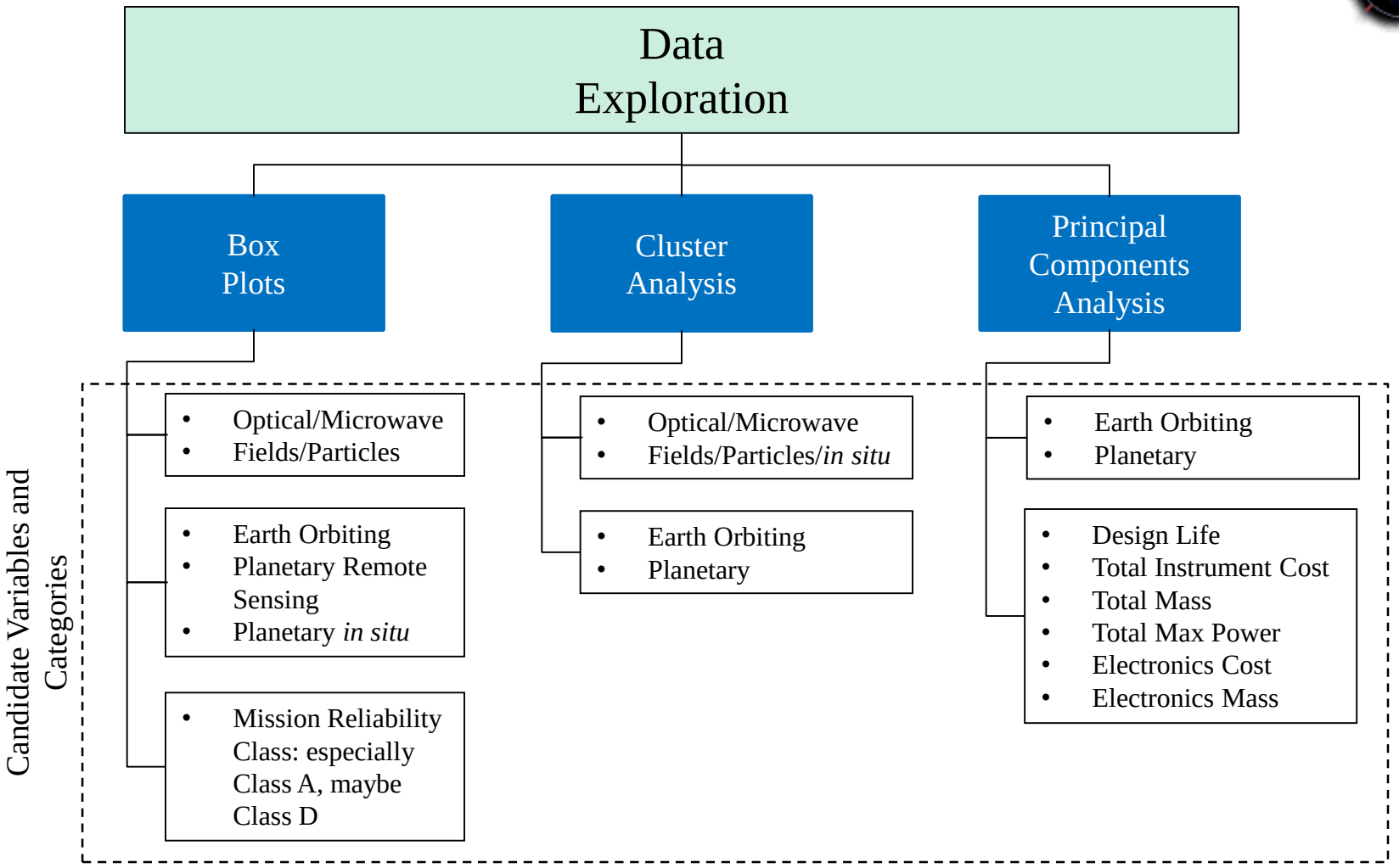
These are some of the many benefits of PCA.

PCA color-coded for Destination Type



UD ○ Planetary ○ Earth Orbiting **V^T →**

Summary up to this Point of the Presentation



Next step… → Model Selection, Fit & Validation

Agenda

- Data Exploration:
  - Histograms, Box Plots
  - Cluster Analysis
  - Principal Components Analysis (PCA)
- **Draft SERs for NICM VIII**
- Future Work
- Feedback
  - Especially from our new schedule friends!

*In what follows:
TIC = Total Instrument B/C/D Cost (FY04 $M)

Current NICM VII SER

- Current NICM VII SER*:
  - Schedule = A(Mission Type, Instrument Type) x $TIC^{0.11}$
  - R^2 = 66%, PE = 20%, N = 148
  - Power model form is reasonable
  - Updated data allowed our Team to relook at variables used in model and their significance

A =

Instrument Type	non-Flagship Planetary	EO & Flagship Planetary
optical	31.3	43.1
active microwave	34.1	46.9
passive microwave	30.9	42.6
particle	34.0	46.7
fields	35.8	49.3
body	31.3	43.1
probe	39.4	54.1
arm/mast	33.4	45.9

Table 5-2 Instrument SER Parameter Matrix (= *A*)

*TIC = Total Instrument Cost (FY04 $K)
PE = Prediction Error of SER (1-sigma); N = # of data points used on SER
"Mission Type" can take the values Earth Orbiting (EO), Flagship Planetary, or non-Flagship Planetary

Draft SERs for NICM VIII

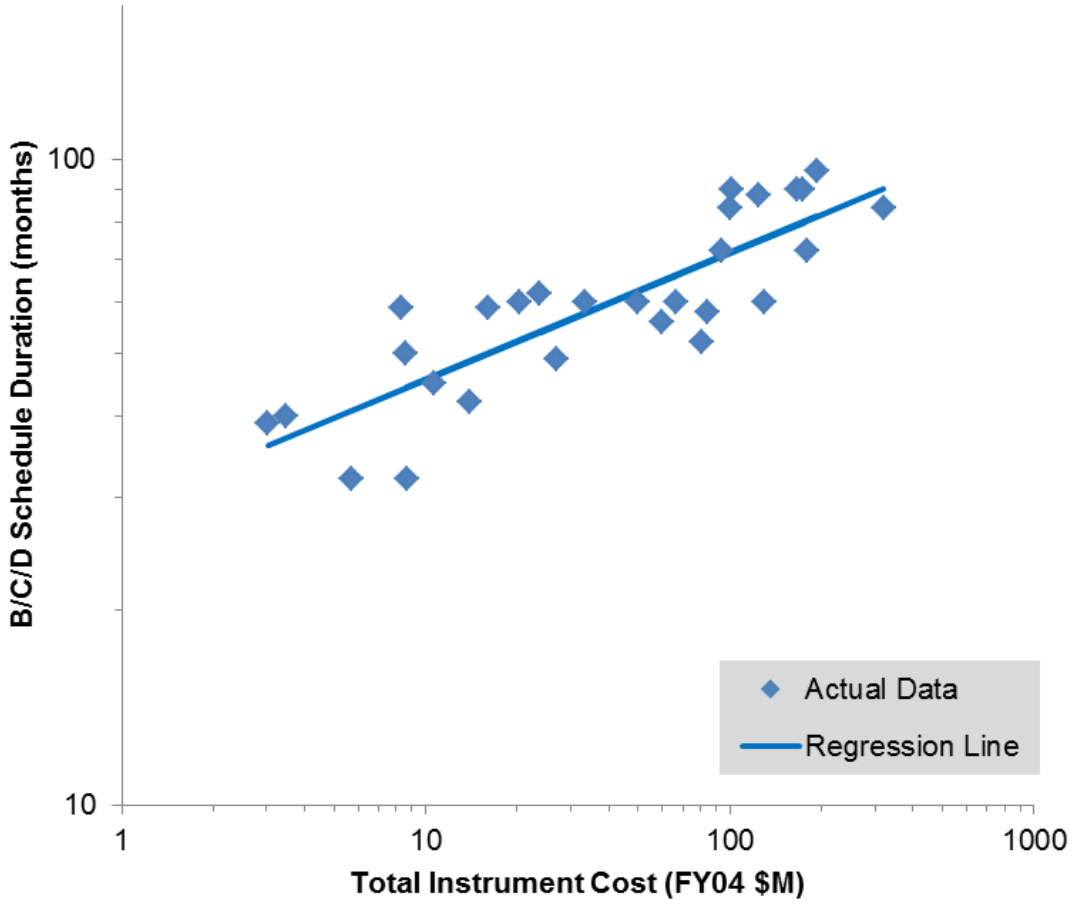
Group	Equation*	Statistics of Merit
Earth Orbiting	$\text{Schedule} = 29 \times \text{TIC}^{0.20}$	R^2 = 71% PE = 18% N = 28
Planetary, Remote Sensing	$\text{Schedule} = \begin{cases} 17 \times \text{TIC}^{0.34} \text{ if Optical or MW} \\ 22 \times \text{TIC}^{0.34} \text{ if Fields or Particles} \end{cases}$	R^2 = 55% PE = 23% N = 36
Planetary, *in situ*	$\text{Schedule} = \begin{cases} 41 \times \text{TIC}^{0.14} \text{ if Mission Reliability Class A} \\ 25 \times \text{TIC}^{0.14} \text{ otherwise} \end{cases}$	R^2 = 90% PE = 16% N = 12

*TIC = Total Instrument Cost (FY04 $K)
PE = Prediction Error of SER (1-sigma); N = # of data points used on SER

Earth Orbiting SER

$$\text{Schedule(months)} = 29 \times \text{TIC}^{0.20}$$

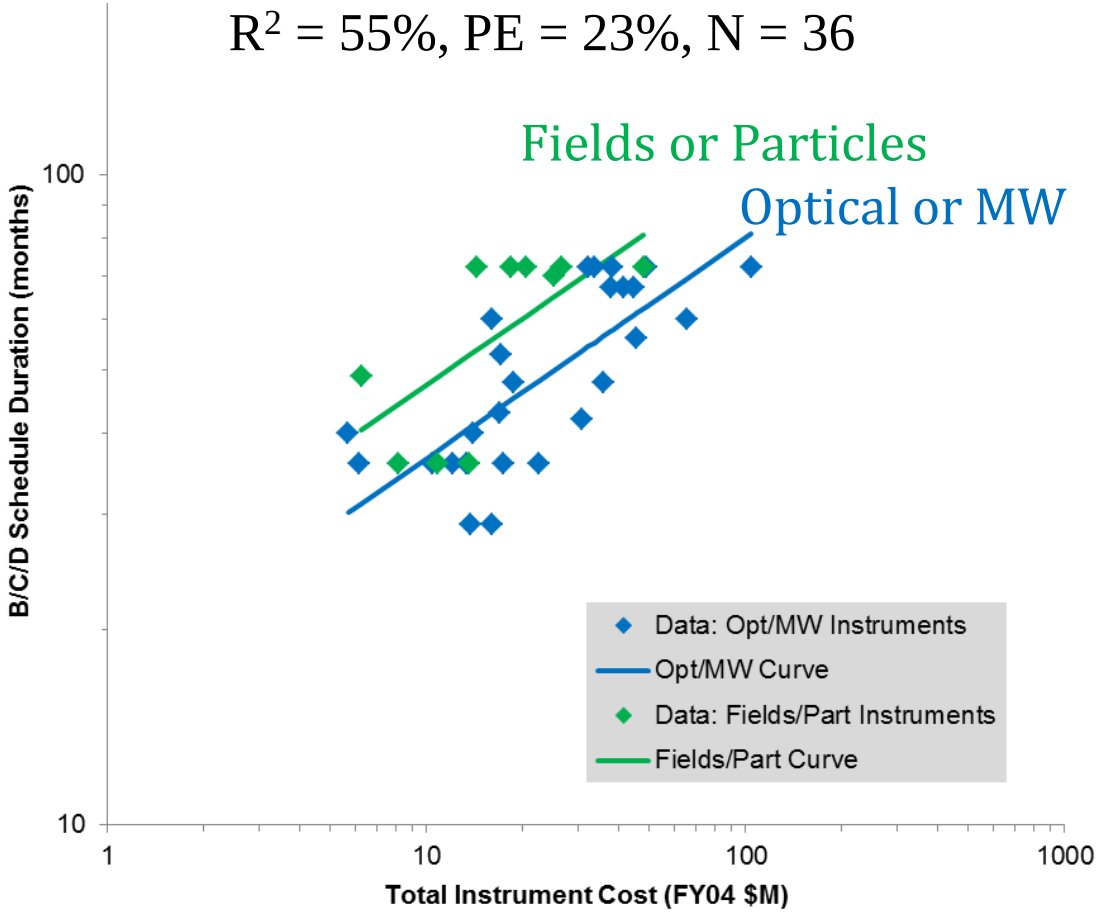
R^2 = 71%, PE = 18%, N = 28



Planetary, Remote Sensing SER

$$\text{Schedule (months)} = \begin{cases} 17 \times \text{TIC}^{0.34} \text{ if Optical or MW} \\ 22 \times \text{TIC}^{0.34} \text{ if Fields or Particles} \end{cases}$$

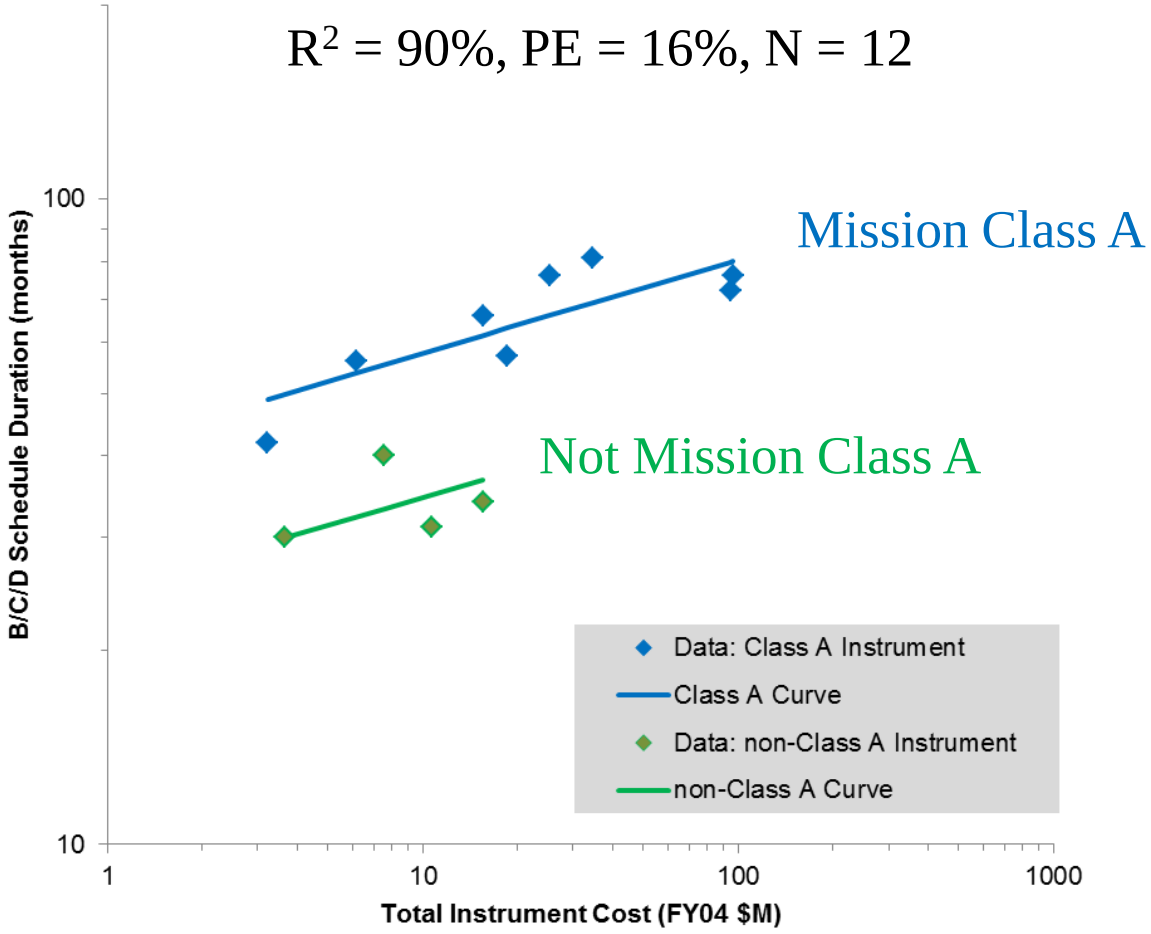
$R^2 = 55\%$, PE = 23%, N = 36



Planetary *in situ* SER

$$\text{Schedule (months)} = \begin{cases} 41 \times \text{TIC}^{0.14} \text{ if Mission Reliability Class A} \\ 25 \times \text{TIC}^{0.14} \text{ otherwise} \end{cases}$$

$R^2 = 90\%$, PE = 16%, N = 12



Agenda

- Data Exploration:
  - Histograms, Box Plots
  - Cluster Analysis
  - Principal Components Analysis (PCA)
- Draft SERs for NICM VIII
- **Future Work**
- Feedback
  - Especially from our new schedule friends!

*In what follows:
TIC = Total Instrument B/C/D Cost (FY04 $M)

Concluding Remarks & Future Work

- We have shown in this presentation:
  - Candidate SERs for NICM VIII
  - The analysis results that steered us to them
- From here we would like to:
  - Hear what you have to say!
  - Incorporate NASA cost/schedule community feedback into our modeling
- Other work we plan on doing:
  - Update analysis with NICM VIII dataset
    - Keep an eye on Mission Class and Design Life as potential parameters for SERs
    - Look more closely at *in situ* – Probe Mounted data
  - Incorporate updated SERs into the NICM VIII Tool

Agenda

- Data Exploration:
  - Histograms, Box Plots
  - Cluster Analysis
  - Principal Components Analysis (PCA)
- Draft SERs for NICM VIII
- Future Work
- **Feedback**
  - Especially from our new schedule friends!

*In what follows:
TIC = Total Instrument B/C/D Cost (FY04 $M)

Backup

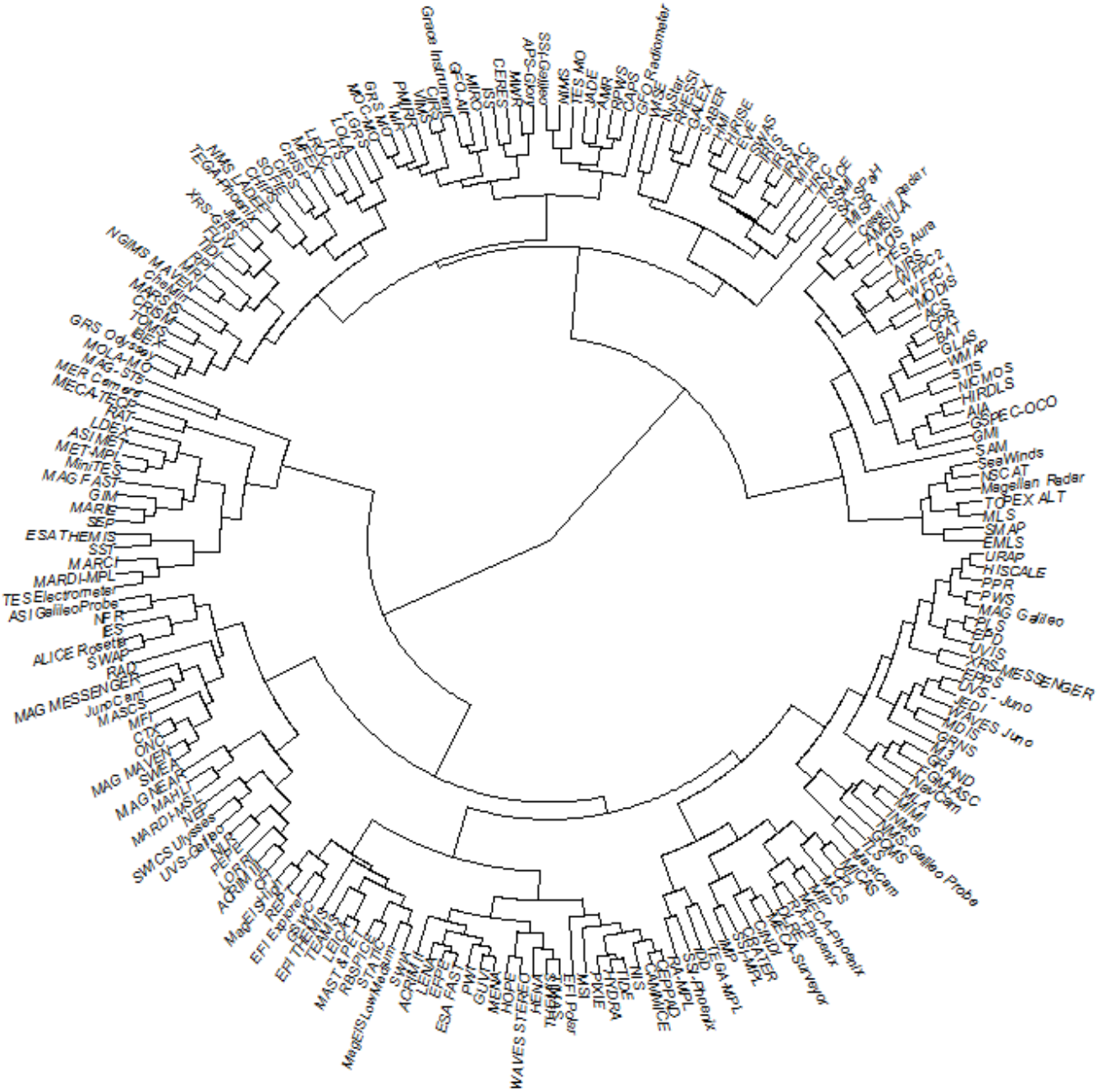
Data used in this Analysis

- Utilized NICM data normalized for modeling, as of November 9, 2016

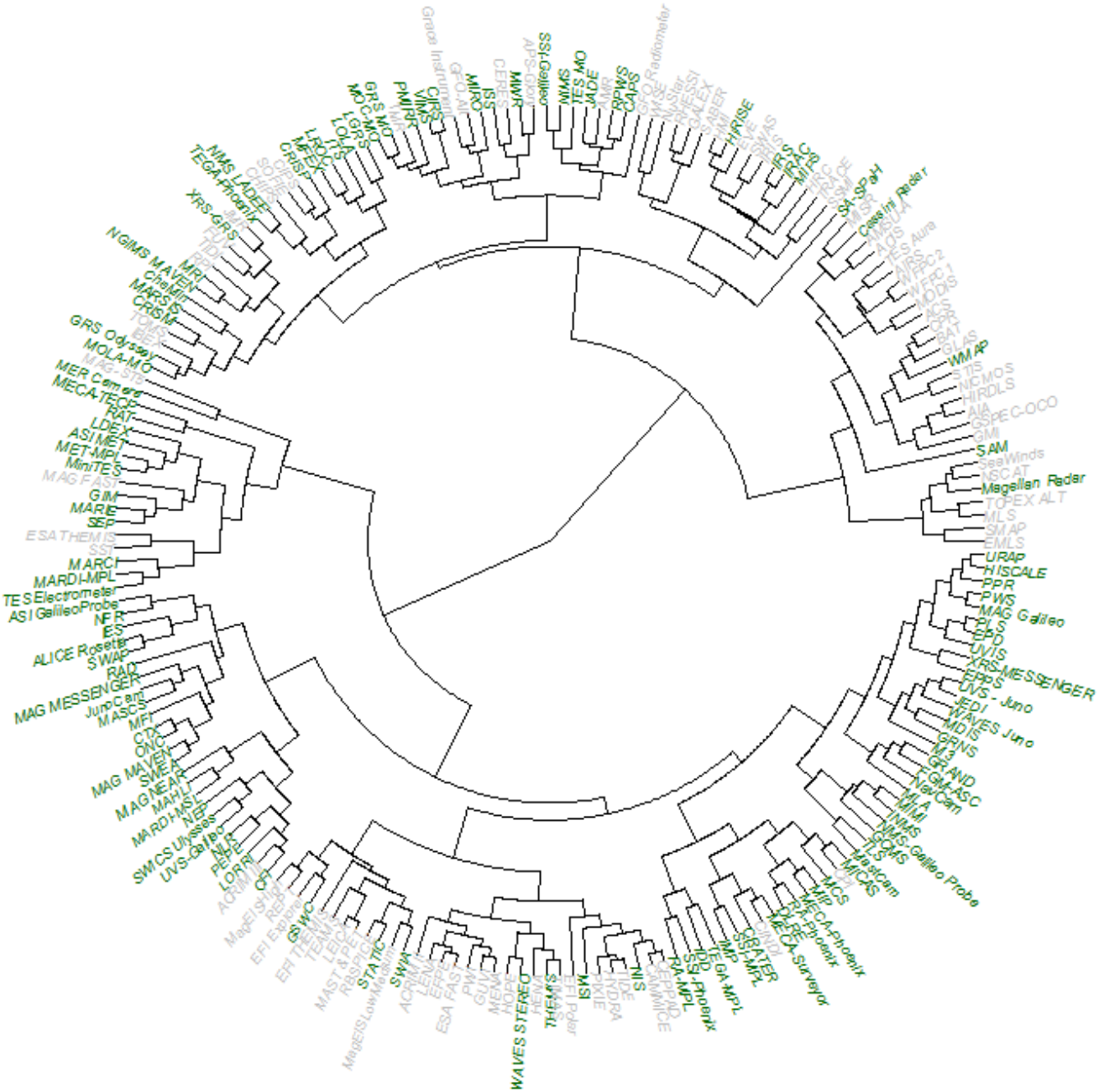
- Filtered these data for the following when performing regression analysis
  - Include only observations where the B/C/D schedule duration was reviewed and documented.
  - "Faster-Better-Cheaper" (FBC) data were not included
  - Instruments launched prior to 1990 were excluded
    - Only exception to this are *in situ* Probe mounted instruments, due to small sample size
  - 76 of 80 total data points used to develop SERs

- NICM-E instruments are included in SERs

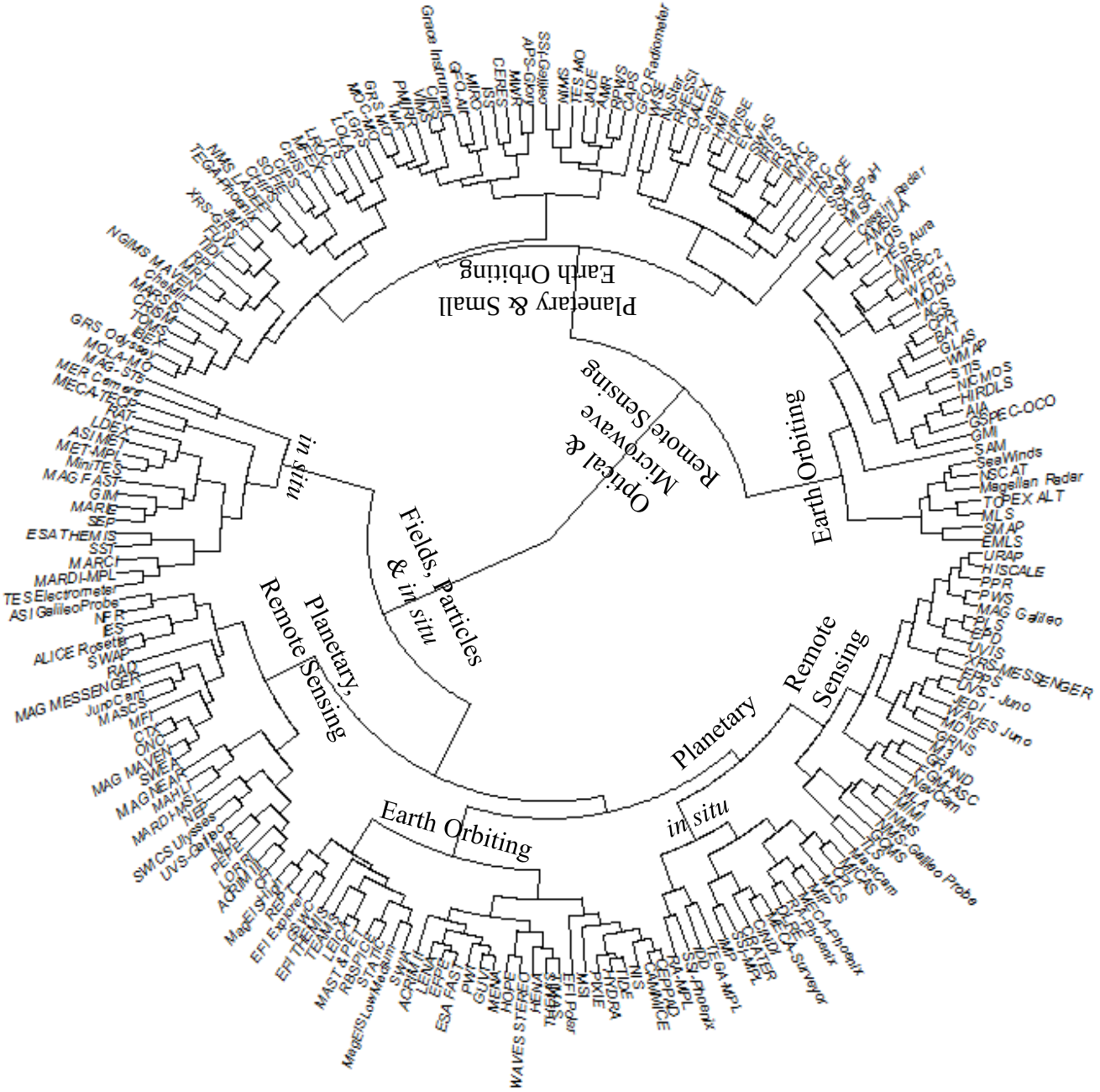
All Instruments



Planetary Instruments



All Instruments



Principal Components Analysis

- Decomposes our data into matrices describing the covariance structure and driving dimensions of variance
  - Utilizes the Singular Value Decomposition (SVD)
- $X = UDV^T$ where
  - U, V are unitary matrices containing the eigenvectors of XX^T and X^TX, respectively
  - D is a diagonal matrix of square roots of eigenvalues of X^TX, in decreasing order
- Helps us:
  - Find uncorrelated dimensions of the data
  - Identify candidates for regression variables
  - Look for clusters in a way that considers correlation
  - Assess multicollinearity

Decomposing Data with Principal Components

$$\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V^T}$$

X — Original data matrix of *n* observations and *p* variables

U — Data re-written according to principal components (unitary)

D — Magnitude of Principle Components

$\mathbf{V^T}$ — Variables re-written according to principal components (unitary)

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{np} \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_p \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \ddots & \vdots \\ v_{p1} & \cdots & v_{pp} \end{bmatrix}$$

UD

This product shows us our data re-written according to principal components (sometimes called "scores")

$\mathbf{V^T}$

Shows us the projection of our variables as vectors in our principal components space

What structure is in our data?

(with an eye to schedule)

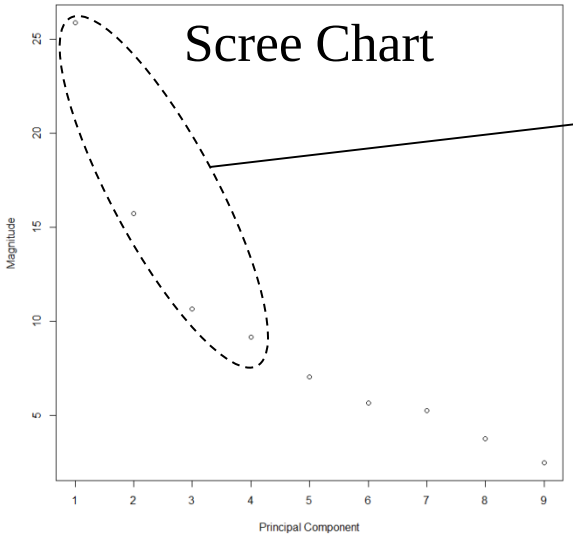
Non-predictive collinearities

SVD - V Matrix	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
log.DevelopmentSchedule.	0.30	0.38	-0.03	-0.39	-0.74	0.06	0.22	-0.02	0.05
log.TotalDevelopmentCosts.	0.43	-0.03	0.02	0.08	-0.06	0.02	-0.43	0.49	-0.61
log.DesignLife.	0.09	0.45	-0.81	0.28	0.17	0.18	0.01	-0.05	-0.01
log.TotalMass.	0.42	-0.05	0.11	0.07	0.26	0.22	0.42	0.55	0.45
log.TotalMaxPower.	0.39	-0.22	0.12	0.12	-0.03	0.65	-0.28	-0.50	0.12
log.TotalDataRate.	0.17	-0.45	-0.45	-0.71	0.22	-0.07	-0.03	-0.04	-0.01
log.ElectronicsCost.	0.42	0.04	-0.02	0.16	-0.03	-0.63	-0.40	-0.14	0.47
log.ElectronicsMass.	0.42	0.05	0.14	0.10	0.25	-0.28	0.53	-0.42	-0.43
log.LaunchYear...1980.	-0.02	-0.63	-0.32	0.44	-0.48	-0.10	0.26	0.05	-0.03
Magnitude	25.88	15.70	10.64	9.13	7.03	5.65	5.25	3.75	2.46

Potentially Significant Schedule Impact on Data Structure

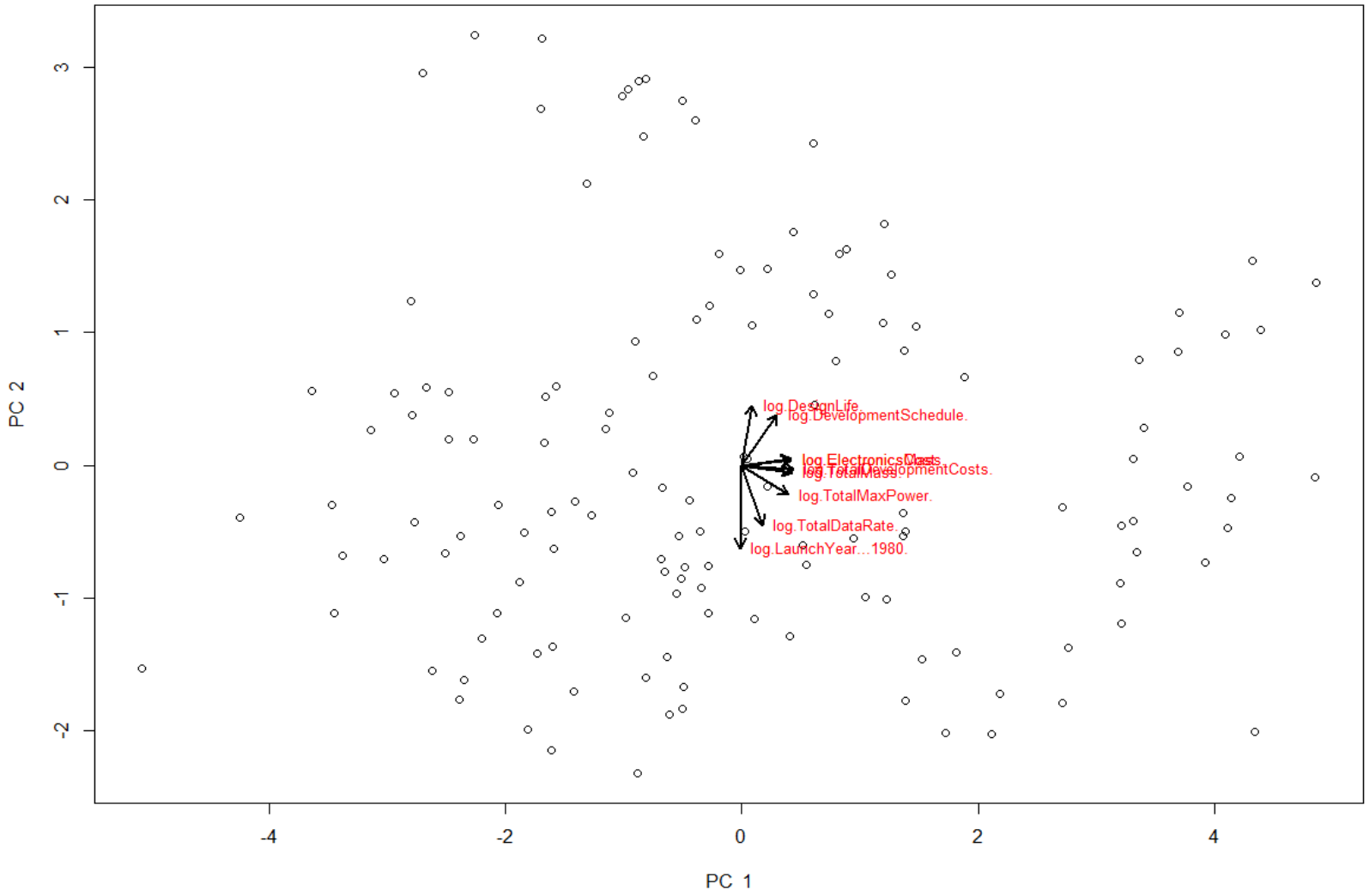
Potentially Significant Correlation with Schedule on Predictive Collinearities

Absolute value of 0.30 judgmentally selected as a threshold for correlates above.



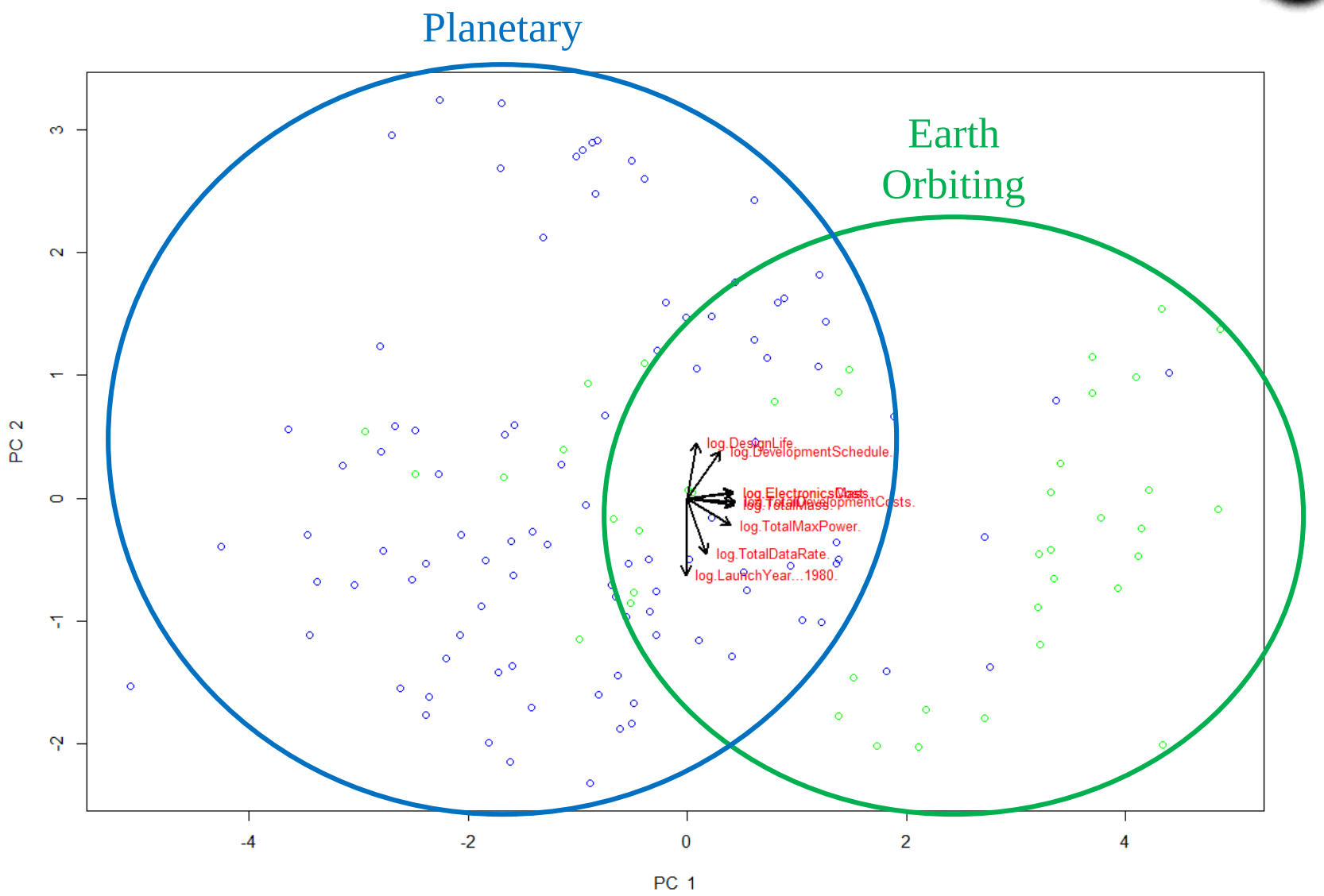
1-4 Principal Components seem reasonable to explain most of the data structure (may be able to refine this with some bootstrapping)

NICM Data Decomposed using PCA
(as seen from the first two principal dimensions of variance)



$$\mathbf{UD} \circ \mathbf{V}^{\mathbf{T}} \rightarrow$$

Previous slide color-coated for Destination Type

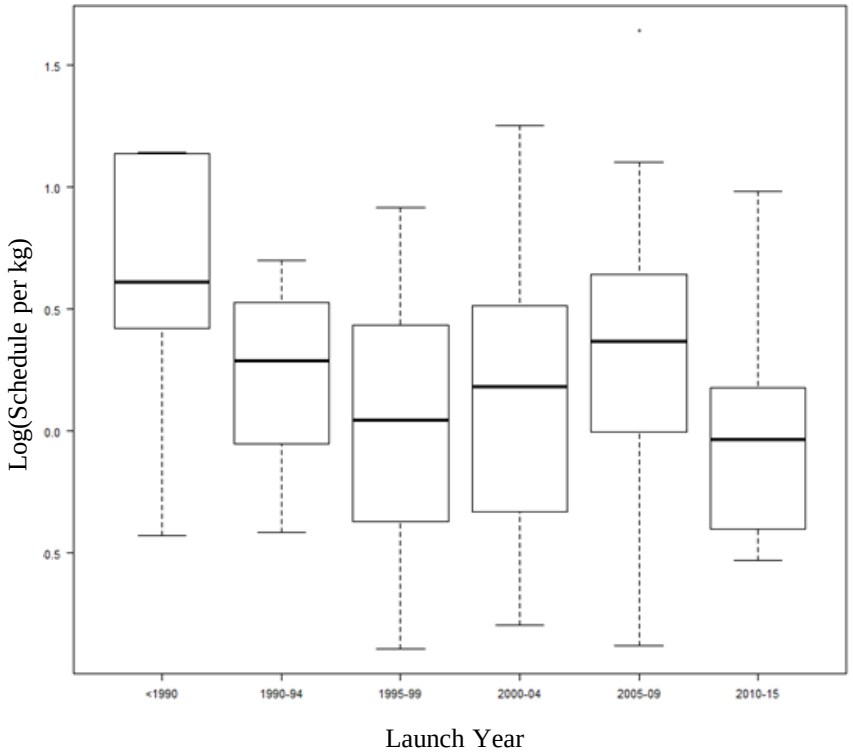


UD ○ Planetary ○ Earth Orbiting **$V^T \rightarrow$**

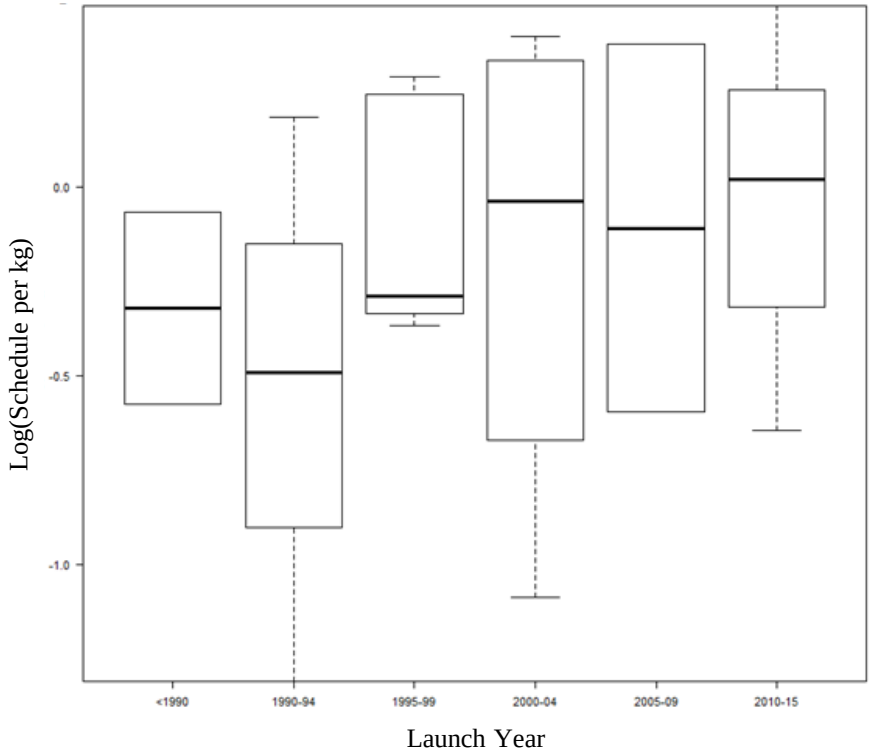
Schedule per kg: Median decrease for Optical Instruments; increase/same for other instrument types

- Not possible to identify future trends in the data
- Removed instruments launched before 1990 to use more relevant data in the model

Optical Instruments

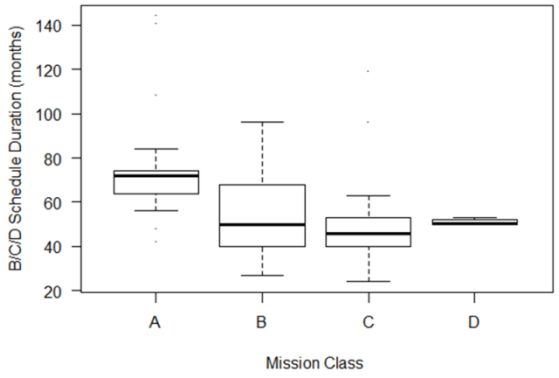


Microwave Instruments

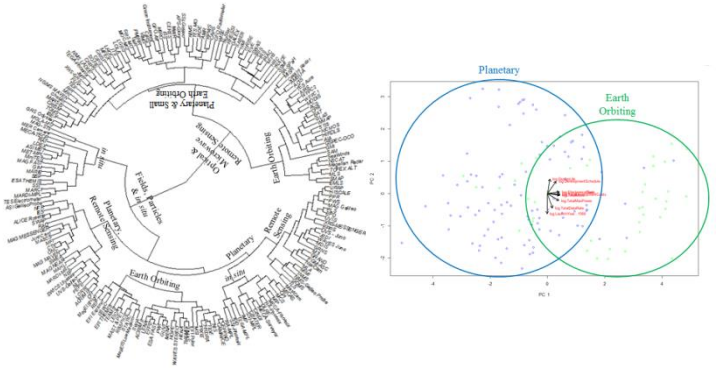


Analysis Process

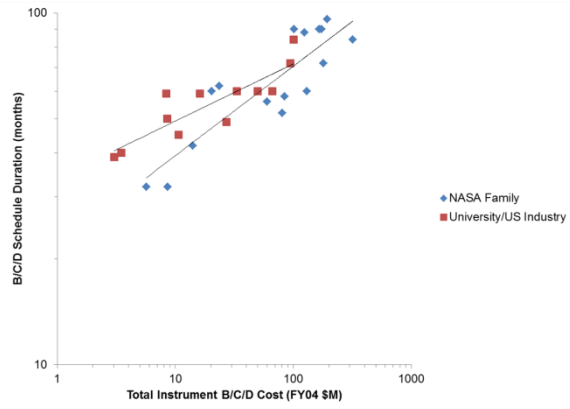
1 Data Exploration w/ Box plots



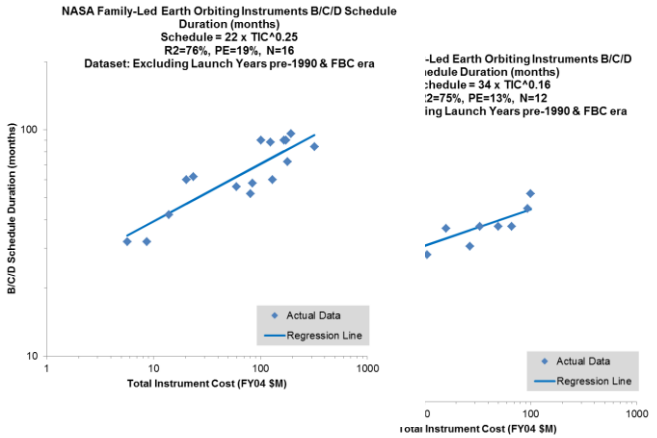
2 Data Exploration w/Cluster Analysis & PCA



3 Regression Analysis



4 Candidate SERs for NICM VIII



Cluster Analysis

- Cluster Analysis performed used the following data fields:
  - Development Schedule (months)
  - Total Instrument Cost (FY04 $K)
  - Total Mass (kg)
  - Total Max Power (W)
  - Design Life (months)
  - Electronics Cost (FY04 $K)
  - Electronics Mass (kg)
  - Launch Year - 1980

$$d(i,j) = \sqrt{\sum_{A}\left[\log\left(\frac{A_i}{A_j}\right)\right]^2}$$

Distance between instruments *i* and *j*. A_i is an attribute of instrument *i* (e.g. Cost, Schedule)

$$d(S,T) = \mathrm{Max}\left(d(x,y)\right)$$

Distance between clusters of instruments, *S* and *T*.