Data Mining Methods and Cost Estimation Models

Why is it so hard to infuse new ideas?

Jairus Hihn
Jet Propulsion Laboratory, California Institute of Technology

Tim Menzies
North Carolina State University

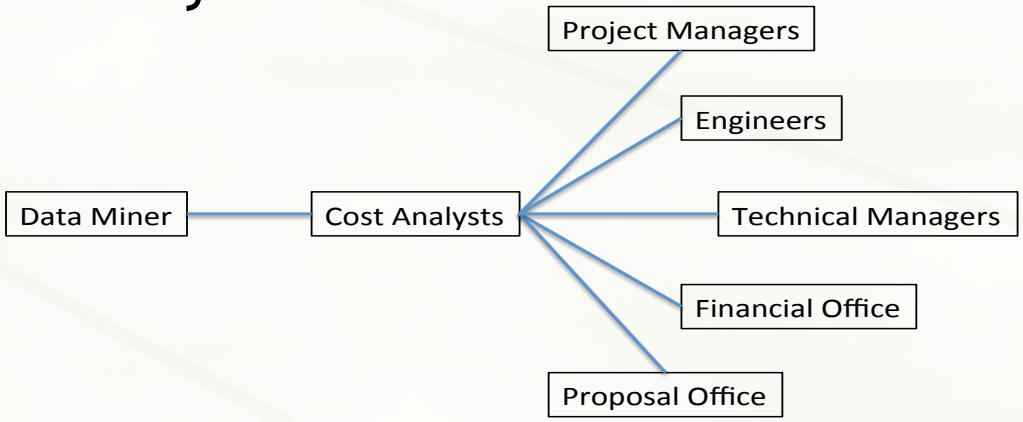
ACTION15: Actionable Analytics for Software Engineering
Lincoln, Ne.

Introduction

- In this talk we will describe our experiences and lessons learned from our ten year journey in conducting and infusing data mining methods into the world of cost estimation
  - 10 years ago with 2CEE we failed to make it to the launch pad
  - This time we have launched and even have a few early users but will we make it to a stable earth orbit

Background

- Everyone in an organization knows just enough about cost to be dangerous
- Current cost methods have been in use for over 30 years and even then there is still some resistance to the use of cost models over expert judgment bottom up methods
- Infusing new ways of doing things is hard and takes time
- There is minimal to no cross fertilization of ideas between academia and practitioners in industry
- Stakeholder Communication is layered



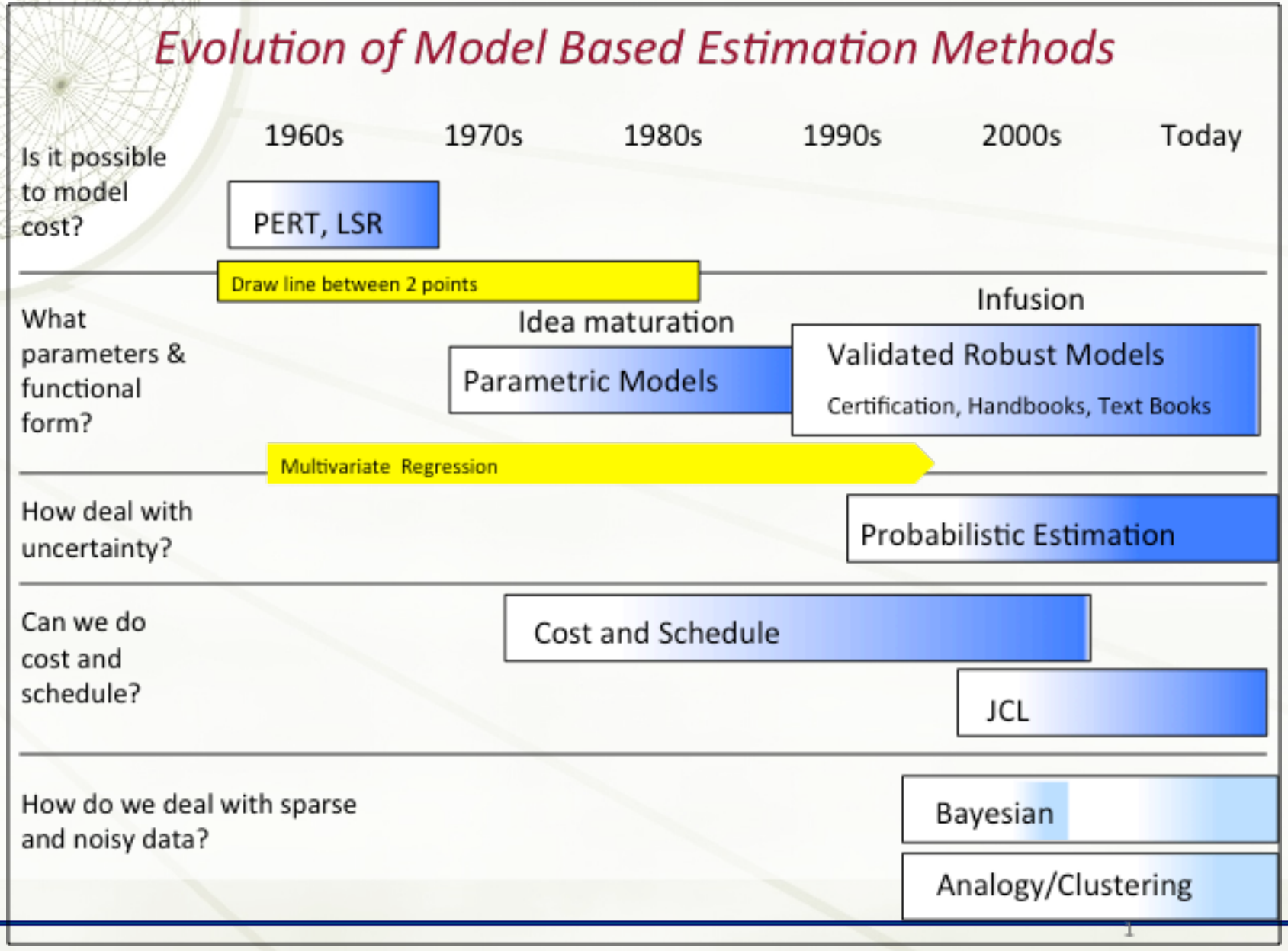
A Script Writers Version of the Conversation Between a SME and DME

- DME: (All excited like he just solved the mysteries of the universe) Look unsupervised learning with xxx outperforms all the other methods based on the median MRE.
- SME: But is there a formal test that we can use like some kind of F-test that tells us one result is actuall siginificantly better than the other?
- DME: That does not matter and is irrelevant because xxx.
- SME: Is there a "model" I can show my stakeholders?
- DME: Sort of but it changes everytime as it depends on what you are comparing to.
- SME: The clusters make no sense.
- DME: But the MREs are great.
- DME: By the way I have a brand new method that works even better.
- SME: (Walks away feeling as if they are walking on quicksand wondering why he ever got involved with this DME)

SME: (Frustrated) What are the actual parameter values in the locally calibrated COCOMO so I can see if they make any sense.

DME: We do not know we just know how the different types of models perform. That's a lower order question. We can do all sorts of things.

History of Industry Cost Estimation Methods



One of the Keys to Success was this time we have a hook

- The purpose of the model is to
  - Supplement current estimation capabilities
  - *Be effective in the very early lifecycle when our knowledge is fuzzy*
    - uses high level systems information (Symbolic Data)
  - *Be usable by Cost Estimators, Software Engineers and Systems Engineers*
- *Methodology handles*
  - small sample sizes
  - noisy and sparse data

Keys to Success - 1

1. Engage, engage, and engage again.

- Plan for infusion from day one. New methods require new processes and new ways of thinking, these do not change easily.
- Be patient. It takes time to introduce something fundamentally new. Especially into a large organization or community (e.g. NASA).

2. Users before algorithms.

- Clearly understand the industry users problem they are trying to solve. You must meet a recognized need even if you have to help the user articulate what that need is.

3. Learn to Communicate.

- Unless the target community accepts the idea, they will not use the idea.
- Learn the users vocabulary and patently teach them the key aspects of your vocabulary
- Constantly communicate to develop a shared language and at least consistent mental models

4. Validation is important

- Validate continuously both at the technical level and with the SMEs.
- Make things visible to the user but do not burry them in overly technical details of the algorithm.
- But be aware that merely having excellent validation results is not enough to guarantee success

Keys to Success - 2

5. Broad skill set, big toolkit

- Initially prototype and experiment with a variety of techniques and testing them with the users to determine what works best for both the available data, the usage scenarios, and the existing mental models of the stakeholders

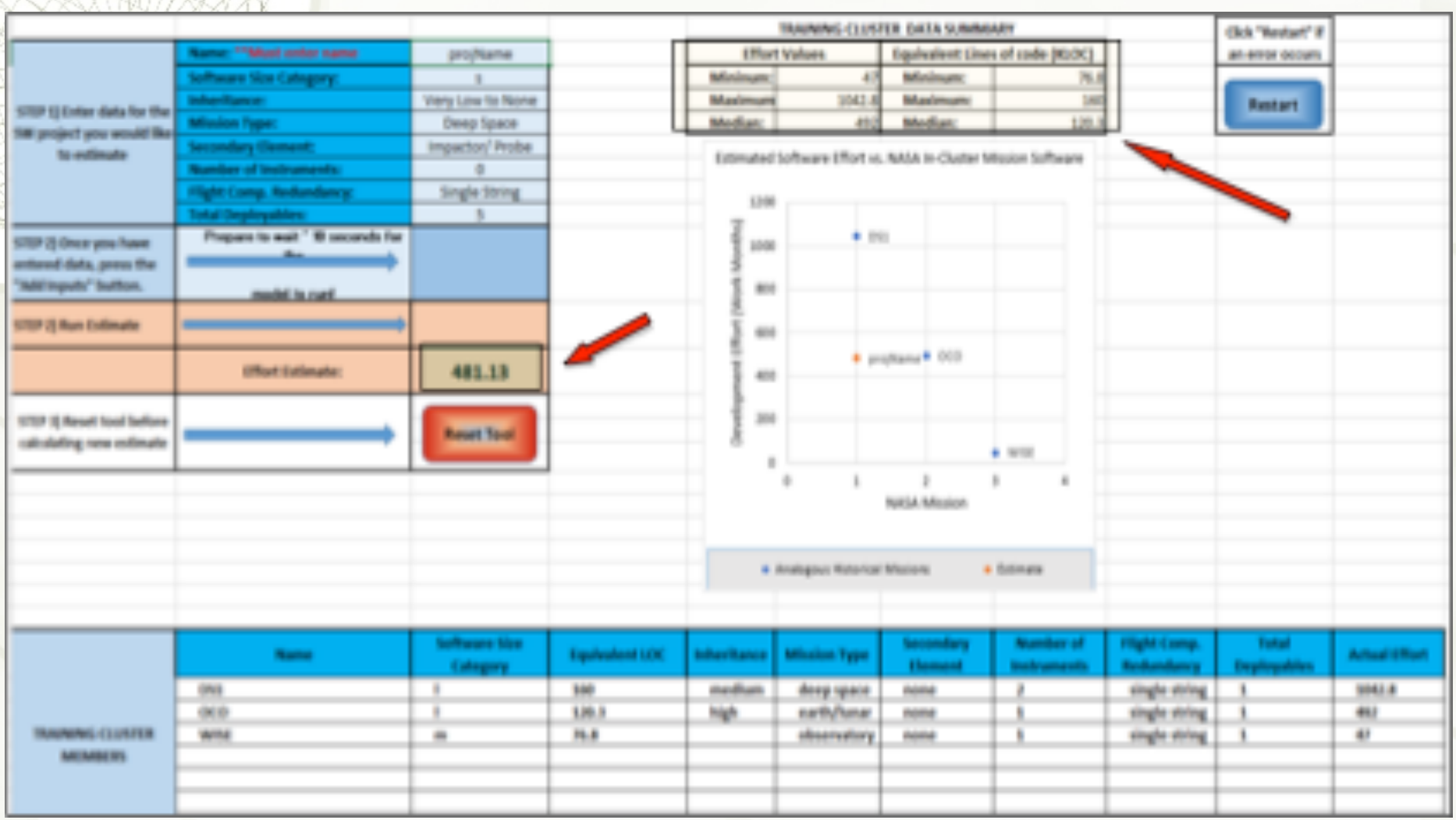
6. Visualization

- Humans by nature but especially external users need to be able to visualize results and the key internal data structures. Summary statistics provide no insight and create no belief or trust in the results. This enables aspects of validation and also supports the evolution of new insights and new ways to communicate the results.

7. Be humble and open to unexpected ideas.

- The truth is a blend of what the data says and the experts many years of experience. The perfect answer is rarely the best answer. Sometimes the data is partly wrong.

First Stab Prototype Interface



Next Step

This capability will be released through the NASA ONCE Portal on December 4th for use by NASA and its contractors so we will launch

But …