Adopting a Data Science Paradigm
Merging Traditional Cost Methodologies with Advanced Computational Analysis

Kyle Ferris, Eric Hagee, Zoe Keita, John Maddrey
Tecolote Research, Inc.

2021 Joint IT and Software Cost Forum
Data Science as a Paradigm

- Fundamentally, data science entails the development of structured datasets towards addressing research questions or mission requirements.

- The field of Data Science emerged in response to recent advances in computational data processing.
  - The significant volume, velocity, and variety of data made available through online platforms, applications, databases, and Internet-of-Things (IoT) devices makes automated data collection, modeling, and analysis a necessity.

- Oftentimes, organizations find themselves having access to more data than they are able to process.
  - There is a critical need for specialists that are able to sift through the “noise” in order to methodically collect, normalize, structure, and present useful data for stakeholders.
Data Science Process

Identify a Research Requirement

Establish Data Collection Criteria & Parameters

Develop an Automated Data Collection Process

Develop Data Normalization & Modeling Workflows

Run Model to Simulate and Identify Data Trends

Utilize Backend Analytics to Communicate Findings
The Data Science Paradigm

- The data science process can be defined in several ways, but all definitions describe the same fundamental goals and desired outcomes.

- The data science process is similar in structure to the cost analysis process, but details surrounding data collection/normalization, modeling, and analytical environment are important differentiators.

Data Science Lifecycle

Collection > Cleaning > Exploratory Analysis > Model Building > Model Deployment

Data Engineering

Data Analysis

ML Engineering

Data Science

Data Science in the Cost Community

**Advanced Analytics**
- Automated mining of cost data based on pre-selected criteria
- Statistical analysis on structured datasets (e.g., regression analysis, learning curve, analysis of variance, pairwise correlation, etc.)
- Automated data visualization using custom programs and/or applications, complete with interactive graphical user interfaces (GUIs)

**Software Development**
- Development of custom programs and/or applications for advanced analytics, modeling/simulation, and data visualization
- Database architecture, engineering, and management for unstructured datasets and/or data repositories
- Development Operations (DevOps) for data pipeline optimization as well as model training, testing, predictions, and deployment

**ML / AI**
- Machine Learning (ML) / Artificial Intelligence (AI) is emerging in the cost community as automated tools for selective data collection and predictive analysis.
- Natural Language Processing (NLP) for analysis of analogous program requirements
- Data imputation based on automated correlation and weighted regression analysis
Why Should We Care?

- Modern data science methodologies and tools are vital to evolving data collection and management requirements.
  - Traditional cost analysis will need to incorporate elements of software development, Machine Learning (ML), and Artificial Intelligence (AI) for improved analytics.

- It is beneficial to think of data science as a *complement* to cost analysis, rather than something that *competes* with it.
  - Despite continued advances in computer automation and artificial intelligence, there will always be a need for analysts to assign value and interpret meaning for data outputs.

- Because cost analysts are typically skilled with statistical modeling and analysis, they are well postured to branch into the wider field of data science with complementary skillsets in computer programming and data visualization.
Traditional Cost Analysis Paradigm

- As a systematic process, cost analysis is proven to help acquisition stakeholders understand the financial scope involved in research, investment, maintenance, and disposal for a long-term program.
  - The statistical methodologies and rigor involved with defensibly projecting future costs – to include time phasing, regression analysis, weighted factors, and extrapolation of actuals – should not be discounted regarding the development of accurate estimates.

- Rather than viewing the traditional cost analysis paradigm as “dated” or “inadequate”, it is appropriate to recognize that the traditional paradigm remains effective – but simply does not include the advanced computational data processing and analysis workflows that newer technology can offer.
Traditional Cost Analysis Process

**Figure 1: The Cost Estimating Process**

<table>
<thead>
<tr>
<th>Initiation and research</th>
<th>Assessment</th>
<th>Analysis</th>
<th>Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your audience, what you are estimating, and why you are estimating it are of the utmost importance.</td>
<td>Cost assessment steps are iterative and can be accomplished in varying order or concurrently.</td>
<td>The confidence in the point or range of the estimate is crucial to the decision maker.</td>
<td>Documentation and presentation make or break a cost estimating decision outcome.</td>
</tr>
</tbody>
</table>

**Analysis, presentation, and updating the estimate steps can lead to repeating previous assessment steps.**

- Define the estimate's purpose
- Develop the estimating plan
- Define the program structure
- Identify ground rules and assumptions
- Conduct sensitivity analysis
- Conduct a risk and uncertainty analysis
- Develop the point estimate and compare it to an independent cost estimate
- Document the estimate
- Present estimate to management for approval
- Update the estimate to reflect actual costs/changes

Source: GAO.


URL: [https://www.gao.gov/assets/gao-09-3sp.pdf](https://www.gao.gov/assets/gao-09-3sp.pdf)
Limitations of Traditional Cost Analysis

Manual Data Collection & Normalization
- Slow workflow process with little to no automation
- Tedious and repetitive tasks with high probability for human error
- Limited scope of available data due to slow collection cycles and time constraints

Static Data Management
- Convoluted and/or overwhelming data fields to populate or review
- Inability to house very large datasets (Excel/Access)
- Limited workflow customization options using canned macro functions (Excel/Access)
- Limited GUI customization options using VBA code (Excel/Access)

Cost Estimating Methodologies
- Inaccurate understanding or reporting of project scope and requirements
- Indefensible and/or unsubstantiated inputs
- Heavily biased inputs accounting for human optimism/pessimism
- Inaccuracies caused by manual normalization errors
Enterprise Data Management

**Cost Analysis Paradigm**
- Aligned with enterprise acquisition process
- Data calls to define and understand scope of cost estimate
- Deliverables satisfy specific tasking (e.g., produce a Life Cycle Cost Estimate, Independent Cost Estimate, Business Case Analysis, etc.)
- In Federal Government, data collection is often limited to internal sources and Subject Matter Expert elicitation

**Data Science Paradigm**
- Enterprise mission drives data collection/analysis
- Data team works with product owners to translate enterprise requirements to data analysis requirements
- Iterative work with product owners
- Define focus of research based on enterprise requirements and availability of relevant data to address requirements
- In Federal Government, open-source data collection may be a requirement, though in practice data is often collected internally
## Data Preparation/Modeling

<table>
<thead>
<tr>
<th>Cost Analysis Paradigm</th>
<th>Data Science Paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typically uses small amounts of analogous and/or historical data</td>
<td>Traditionally uses larger amounts of unstructured datasets</td>
</tr>
<tr>
<td>Normalization of data for cost, quantity, and duration</td>
<td>Data size requires more intensive data normalization</td>
</tr>
<tr>
<td>Usually uses linear or non-linear regression</td>
<td>Missing data may be imputed</td>
</tr>
<tr>
<td>Utilization of tools like Microsoft Excel and ACEIT</td>
<td>Machine learning methods such as neural networks, decision trees, etc. may be employed</td>
</tr>
<tr>
<td>Cost estimators may be considered to be main drivers</td>
<td>Data exploration using programming languages (Python, R, etc.)</td>
</tr>
<tr>
<td>Data analyst/Data scientist line usually crossed with the introduction of machine learning</td>
<td></td>
</tr>
</tbody>
</table>
# Evaluation and Deployment

## Cost Analysis Paradigm
- Cost estimates may be iteratively documented, but documentation is largely added and finalized prior to stakeholder review.
- Finalized estimates are presented to stakeholders for approval.
- Deliverables are typically limited to a cost model and associated Microsoft Office files.

## Data Science Paradigm
- Documentation accompanies programming efforts.
- More iterative presentations and adjustment before delivering a final product.
- Product likely to include various programming files piped into a Independent Development Environment (IDE) or application that can integrate files into a curated output.
- Final product may include customized programs and/or web-based tools for end-user analytics.
Cost analysis and data science methodologies are based off the same theoretical basis.

The cost analysis paradigm usually involves direct tasking with smaller enterprise-owned datasets.

The data science paradigm requires more collaboration with enterprise stakeholders to determine how available data can continuously address mission requirements.

This likewise requires a wide-range of technical skillsets (programming, statistical modeling, analysis, visualization) to assign value and predictive trends to data.
Data Science Mission

1. Integration of advanced analytical techniques and programming expertise to provide data driven forecasting and modeling into cost estimates
2. Advance industry best practices in handling, modeling, and communicating cost data
3. Evolve past Subject Matter Expert (SME) input to focus on historical actuals for cost estimation
4. Transition cost estimators from data consumers to data builders
Evolving the Current Paradigm

- Current cost estimating methodologies are slow, and focused heavily on process as opposed to decision making support.
- The benefit of modeling from data allows cost estimation to provide quicker results without bias from SME input.
- Processing data for analytics can be an automatic process where results are refined as a program evolves.
- There exist far more defensible methods for forecasting than basic linear regressions via Monte Carlo simulations, which are not commonly used within the cost community.
Data Science Vision = Process Evolution

- Transition from ad hoc reporting to continuous analytic processes that adjust to changes without compromising the validity of previous estimates
- Leverage data across any relevant source, no matter the format
- Prioritize *communication* of analysis over the analysis itself
Data Science as a Service

- Construct a fully integrated data science stack for cost estimation efforts
- Enable cost estimators to streamline estimating processes and gain efficiency in targeted deliverables
- Shift focus to decision support from process requirements
Developing a Data Science Curriculum

Why train instead of hire?

- Instead of competing for a limited pool of job seekers, look to the current employee talent pool
- “Employers are already struggling to fill Data Science and Analytics jobs, as evidenced by the length of time unfilled roles remain open. On average, DSA jobs remain open for 45 days (Markow et al., 2017)”
- Upskilling can be a much smaller investment than hiring and training a new worker.

To effectively create a comprehensive data science training plan and maximize your outcomes, the curriculum should serve 3 functions:

- Train the workforce
- Institutionalize a knowledge management repository
- Serve as a key driver for scaling analytics
Preliminary Curriculum Planning

**Data Strategy & Roadmap**
- 62% of Insights Leaders have a data science development plan and road map in place, compared with only 28% of Insights Laggards and 29% of The Pack (Forrester, 2016)
- Identify current and future needs
- Brainstorm current and projected use cases

**Identify Required Skillsets**
- Determine which skills matter most for the organization’s aspirations (as describe in Step 1)
- Differentiate between broad skills and deep skills

**Gap Analysis on Employee Skills**
- Establish a baseline
- Identify the gaps in skillsets between the baseline and requirements
Data Science Skillsets

- Data Literacy
- Knowledge Literacy (fundamentals in mathematics, probability, statistics)
- Exploratory Data Analysis, Business Acumen and Storytelling (e.g. data visualizations, soft skills)
- Programming
- Data Extraction & Wrangling
- Applied Statistics/Advanced Mathematics & Experimental Design
- Deployment & DS Platforming
- Machine Learning, Data Engineering, NLP
- Deep Learning

<table>
<thead>
<tr>
<th>phase</th>
<th>PROFILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientist</td>
<td>ML Engineer, Data Scientist, DevOps Engineer, Data Architect</td>
</tr>
<tr>
<td>Data Analyst</td>
<td>Data Scientist</td>
</tr>
<tr>
<td></td>
<td>Data Analyst, Business Analyst, data consumers, end users</td>
</tr>
</tbody>
</table>
The most effective strategies incorporate techniques that make the most of the existing internal personnel as well as external resources.

**Learning Environments**
- L&D program
  - Traditional approach to upskilling a workforce
- Capability academies
- Data labs and workshops
  - Key for continuous development and re-learning
  - Key for raising awareness

**Specialization versus generalization**
Training Considerations

- Training will take time
- Define success flexibly. Not every employee needs to be a master coder
- Practicality of solutions should be constantly analyzed
- Accompanying training should be a promotion of a data-driven culture
References


LaPrade Annette et al. The enterprise guide to closing the skills gap. IBM Institute for Business Value. September 2019. 91026091USEN-01


References


References


