

Adopting a Data Science Paradigm

Merging Traditional Cost Methodologies with Advanced Computational Analysis

Tecolote Research, Inc.

ICEAA 2022 Professional Development & Training Workshop

Kyle Ferris, Eric Hagee, Zoe Keita, John Maddrey



Overview

- ❖ **Cost Analysis Paradigm**
- ❖ **Data Science Paradigm**
- ❖ **Cost Analysis vs. Data Science Paradigm**
- ❖ **Data Science Mission**
- ❖ **Data Science Vision & Evolution**
- ❖ **Data Science Curriculum & Training Development**
- ❖ **Conclusion**

The Cost Analysis Paradigm

- ❖ As a systematic process, cost analysis is proven to help stakeholders understand the financial scope involved with the research & development, investment, maintenance, and disposal of federal government programs
 - ❖ Traditional cost estimating methodologies involved with defensibly projecting future costs – such as regression analysis, analogous factoring, and economic extrapolation – should not be discounted towards the development of accurate estimates
- ❖ It is important to recognize that the traditional cost analysis paradigm remains effective – but lacks the advanced analytics, modeling, and machine learning capabilities that newer technologies can offer

Traditional Cost Estimating Process

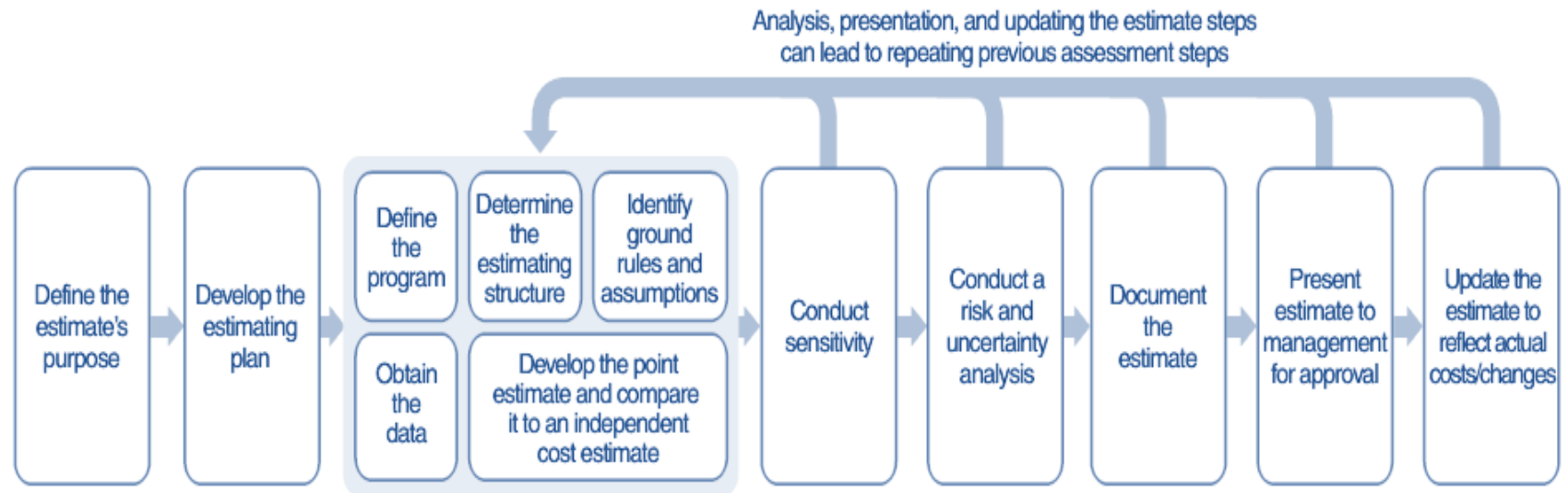
Figure 1: The Cost Estimating Process

Initiation and research
Your audience, what you are estimating, and why you are estimating it are of the utmost importance

Assessment
Cost assessment steps are iterative and can be accomplished in varying order or concurrently

Analysis
The confidence in the point or range of the estimate is crucial to the decision maker

Presentation
Documentation and presentation make or break a cost estimating decision outcome



Source: GAO.

GAO: *Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Capital Program costs.*

URL: <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)

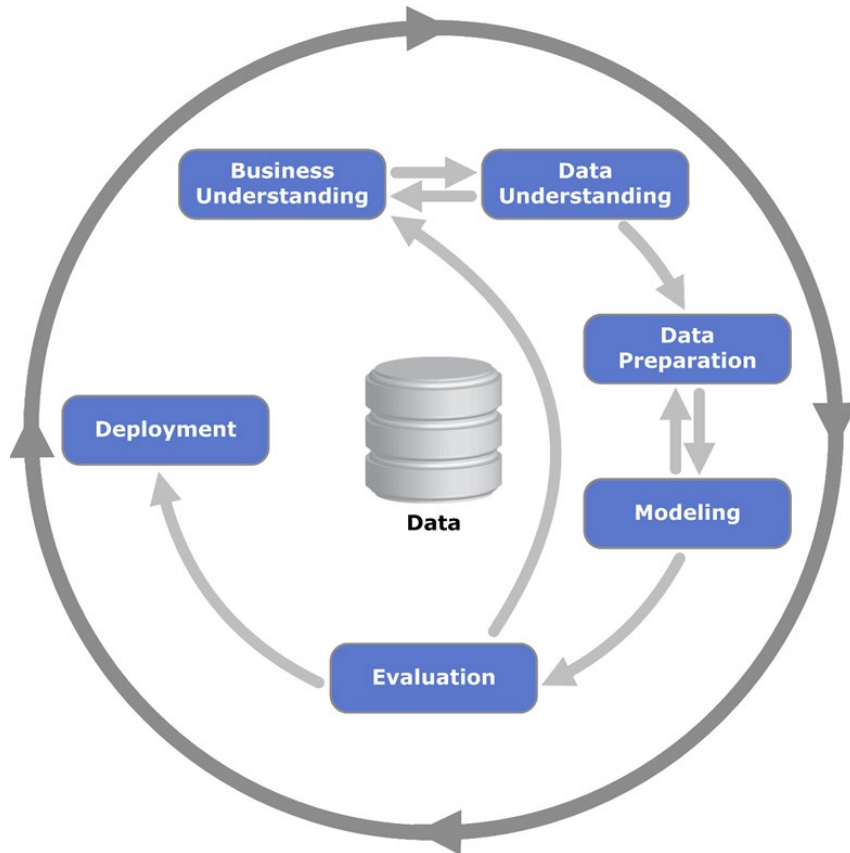
Cost Estimating Methodologies

- 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

Data Science as a Paradigm

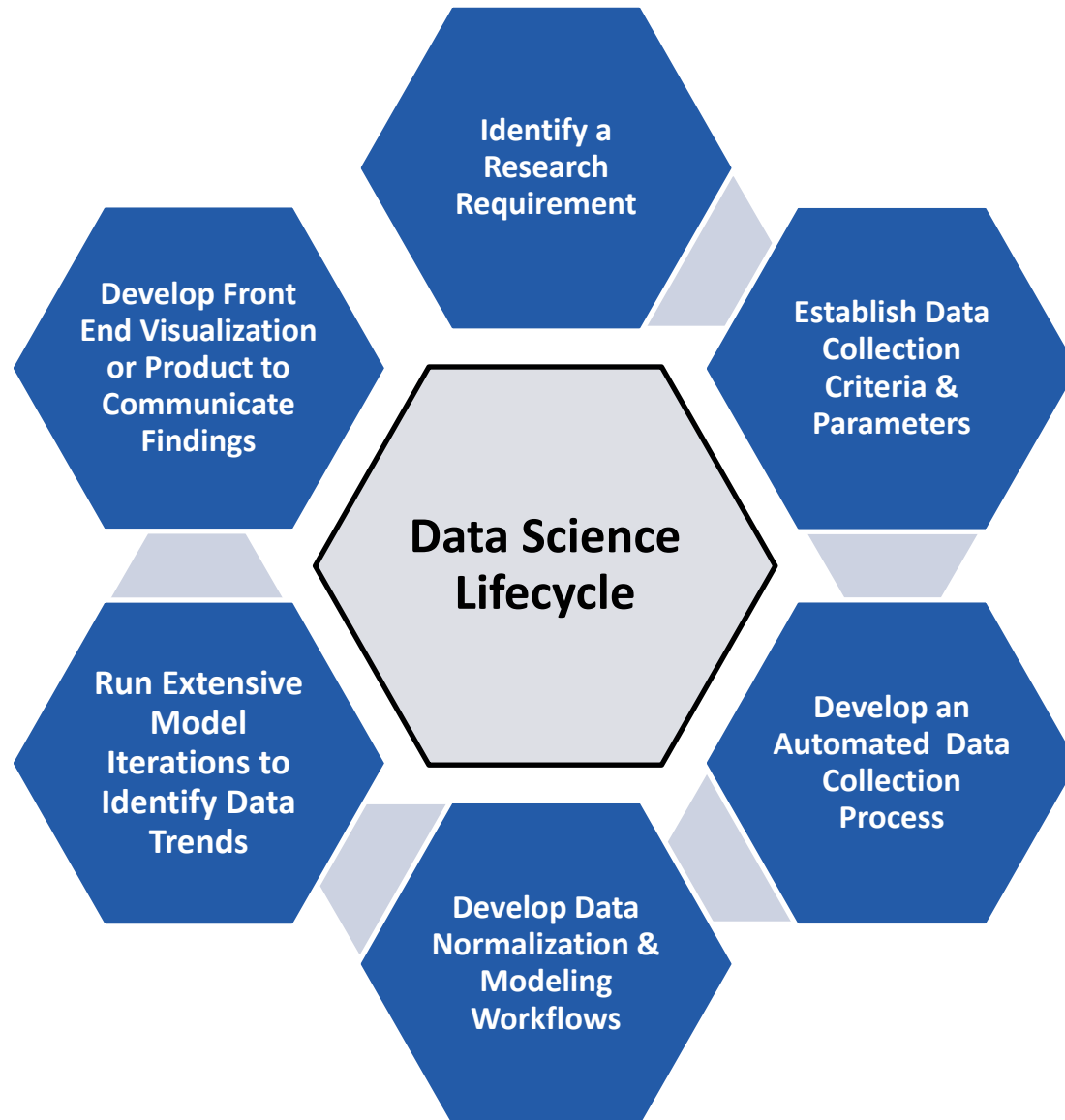
- ❖ Fundamentally, data science entails the development of structured datasets towards addressing research questions or mission requirements
- ❖ The field of Data Science emerged in response to recent advances in computational data processing
 - ❖ The significant ***volume***, ***velocity***, and ***variety*** of data made available through online platforms, applications, databases, and Internet-of-Things (IoT) devices makes automated data collection, normalization, and modeling 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 large unstructured datasets in order to methodically collect, validate, and normalize data using customized programs

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

The Data Science Lifecycle



Presented at the 2022 ICEAA Professional Development & Training Workshop: www.iceaaonline.com/pit2022

The Data Science Process



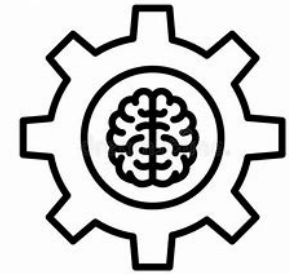
Collection & Normalization

Data Analytics

ML / AI

Data Science

Data Science in the Cost Community



Advanced Analytics

- Automated mining of cost data based on pre-selected criteria
- Regression analysis, learning curve, analysis of variance, pairwise correlation
- Automated data visualization via custom programs and/or applications

Programming

- Custom programs and/or applications for advanced analytics, modeling/simulation, and data visualization
- Database architecture, engineering, and management
- DevOps for data pipeline optimization, testing and deployment

ML / AI

- 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

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 deliverables address shifting enterprise requirements
- 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

Cost Analysis Paradigm

- Typically uses small amounts of analogous and/or historical data
- Normalization of data for cost, quantity, and duration
- Usually uses linear or non-linear regression
- Utilization of tools like Microsoft Excel and ACEIT
- Cost estimators may be considered to be main drivers

Data Science Paradigm

- Traditionally uses larger amounts of unstructured datasets
- Data size requires more intensive data normalization
- Data exploration using programming languages (Python, R, etc.)
- Missing data may be imputed
- Machine learning methods such as neural networks, decision trees, etc. may be employed
- Data analyst/Data scientist line usually crossed with the introduction of machine learning

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

- Iterative documentation accompanies programming efforts
- Iterative presentation of analysis via sprints.
- 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 front-end applications and/or web-based dashboards

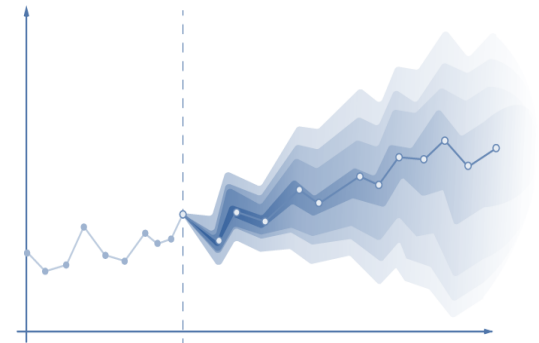
Key Takeaways



- ❖ 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

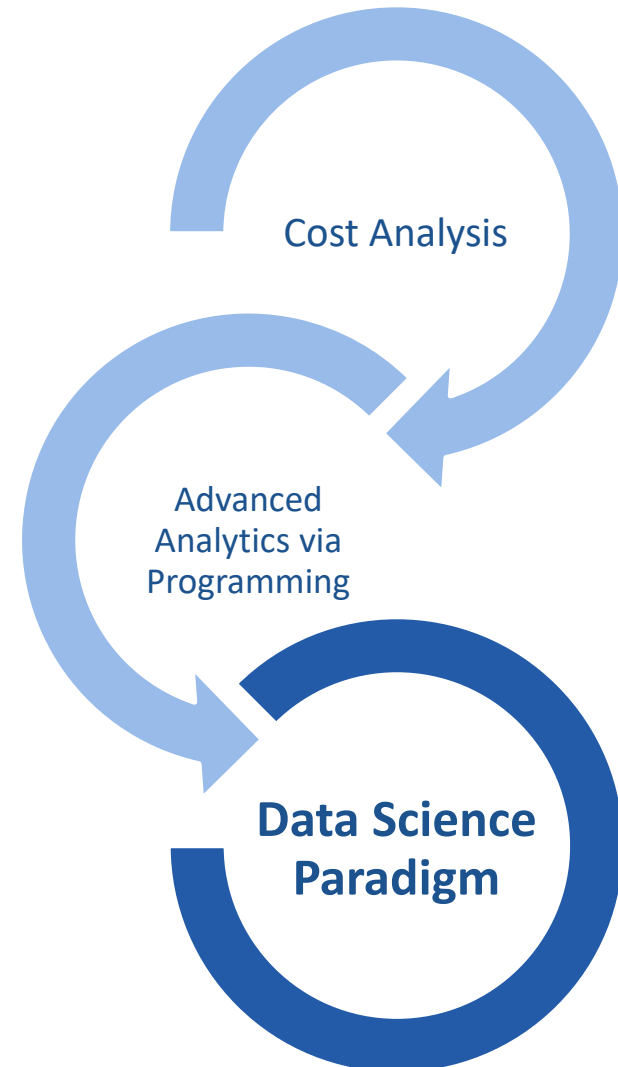
Data Science Mission

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*



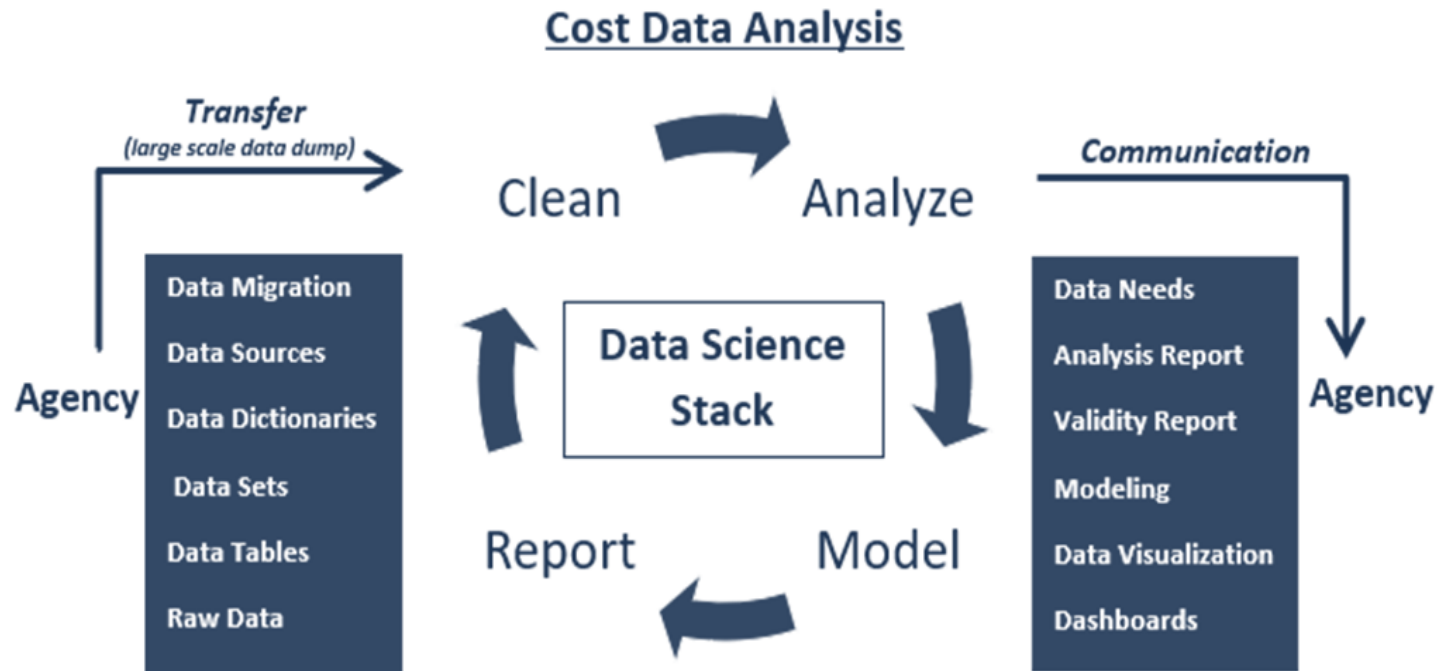
Evolving the Current Paradigm

- ❖ 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
- ❖ Far more defensible methods for forecasting exist than just basic linear regressions via Monte Carlo simulations



Data Science Vision = Process Evolution

- ❖ Transition from ad hoc reporting to continuous analytic sprints 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 (see “Managing Data Science” presentation)
- ❖ Enable analysts to standardize data science operations and gain efficiency towards addressing iterative 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

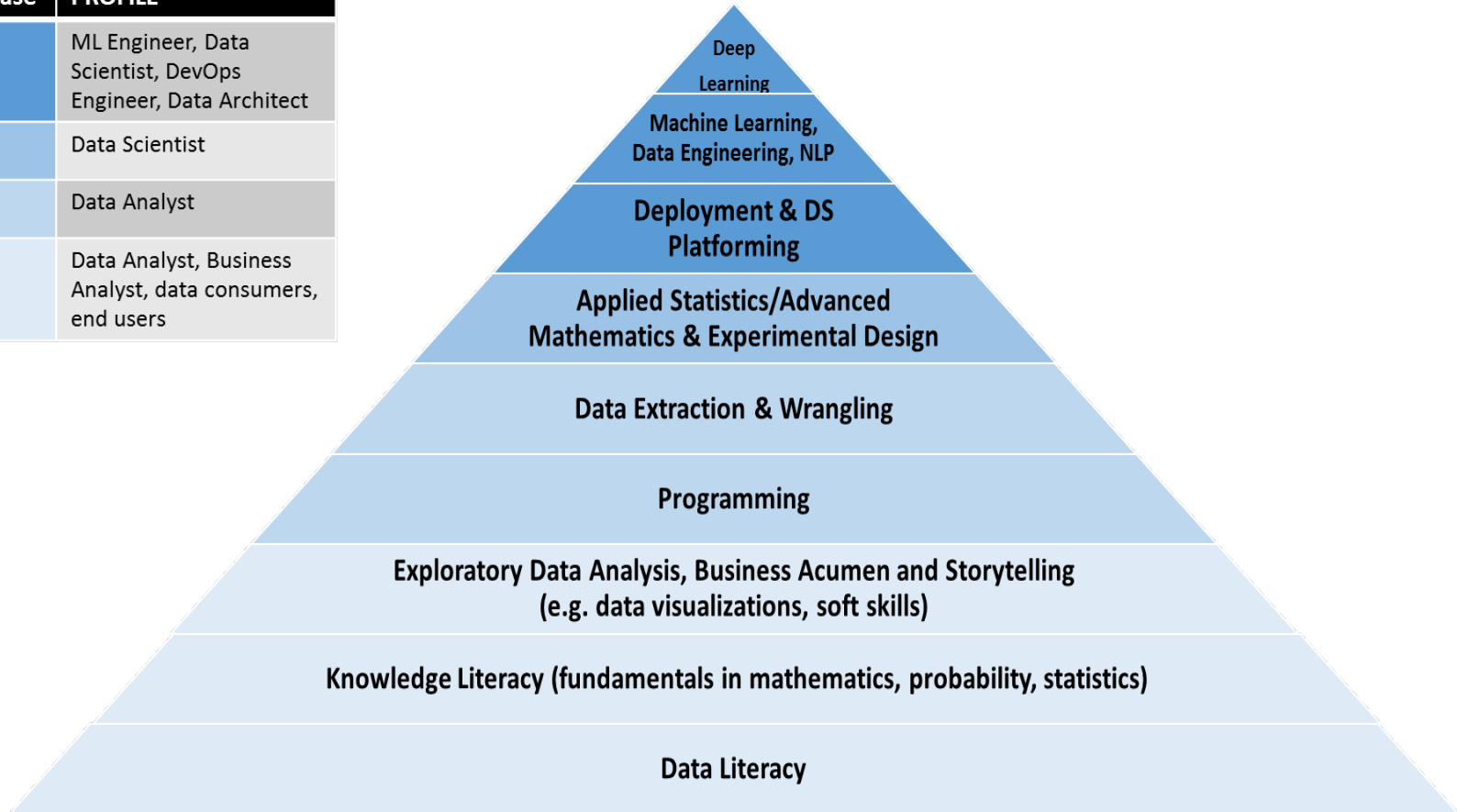
- Determine which skills matter most for the organization's aspirations (as describe in Step 1)
- Differentiate between broad skills and deep skills

Gap Analysis on Employee Skills

- Establish a baseline
- Identify the gaps in skillsets between the baseline and target

Data Science Skillsets

phase	PROFILE
	ML Engineer, Data Scientist, DevOps Engineer, Data Architect
	Data Scientist
	Data Analyst
	Data Analyst, Business Analyst, data consumers, end users



Program Development

- ❖ 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



Conclusion

- ❖ 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.

References

Oracle, Datascience.com. “Scaling Data Science,” 2018.

Forrester Consulting. “Data Science Platforms Help Companies Turn Data Into Business Value,” December 2016

Bisson, P., Hall, B., McCarthy, B., & Rifai, K. (2018, May 22). *Breaking away: The secrets to scaling analytics*. McKinsey & Company. <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/breaking-away-the-secrets-to-scaling-analytics>.

Keller, Scott & Schaninger, Bill. (2019, September 19). *The Forgotten Step in Leading Large-Scale Change*. McKinsey & Company. <https://www.mckinsey.com/business-functions/organization/our-insights/the-forgotten-step-in-leading-large-scale-change>.

LaPrade Annette et al. The enterprise guide to closing the skills gap. IBM Institute for Business Value. September 2019. 91026091USEN-01
Bersin, Josh. (2019, October 5). *The Capability Academy: Where Corporate Training Is Going*.

Josh Bersin. <https://joshbersin.com/2019/10/the-capability-academy-where-corporate-training-is-going/>. Tyagi, Harshit. (2021, January 16). *Data Science Learning Roadmap for 2021*. towards data science. <https://towardsdatascience.com/data-science-learning-roadmap-for-2021-84f2ba09a44f>. *Teachers College Record* Volume 117 Number 4, 2015, p. 1-22 <https://www.tcrecord.org/library> ID Number: 17856, Date Accessed: 7/29/2021 7:39:08 PM

Panetta, Kasey. (2019, February 6). *A Data and Analytics Leader’s Guide to Data Literacy*. Gartner Inc. <https://www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy/>.

Bersin, Josh & Zao-Sanders, Marc. (2020, February 12). *Boost Your Team’s Data Literacy*. Harvard Business Review. <https://hbr.org/2020/02/boost-your-teams-data-literacy>

Rout, Amiya Ranjan. “How to Become Data Scientist – a Complete Roadmap.” GeeksforGeeks, GeeksforGeeks, 9 Mar. 2021, www.geeksforgeeks.org/how-to-become-data-scientist-a-complete-roadmap/.

Waller, David. (2020, February 06). *10 Steps to Creating a Data-Driven Culture*. Harvard Business Review. <https://hbr.org/2020/02/10-steps-to-creating-a-data-driven-culture>.

Markow, Will et al. *The Quant Crunch: How the Demand for Data Science Skills is Disrupting the Job Market*. Burning Glass Technologies, 2017.

Ermakova, Tatiana et al. (2021) Beyond the Hype: Why do Data-Driven Projects Fail? *Proceedings of the 54th Hawaii International Conference on System Sciences*. Scholar Space. <https://scholarspace.manoa.hawaii.edu/bitstream/10125/71237/0498.pdf>.

References

Springboard. (2019, March). The Data Science Process. Kdnuggets. <https://www.kdnuggets.com/2016/03/data-science-process.html>

Wirth, Rudiger and Hipp, Jochen. (2000). “CRISP-DM: Towards a Standard Process for Data Mining”. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. 29-39. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.198.5133>

Mason, Hilary and Wiggins, Chris. (2010, September 25). A Taxonomy of Data Science. Dataists. <http://www.dataists.com/2010/09/a-taxonomy-of-data-science/>

Davenport, Thomas H. and Patil, D. J. (2012, October). “Data Scientist: The Sexiest Job of the 21st Century”. Harvard Business Review. <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

Bandyopadhyay, Raj. (2017, January 9). The Data Science Process: What a data scientist actually does day to day. Medium. <https://medium.springboard.com/the-data-science-process-the-complete-laymans-guide-to-what-a-data-scientist-actually-does-ca3e166b7c67>

Nantasenamat, Chanin. (2020, July 27). The Data Science Process: A Visual Guide to Standard Procedures in Data Science. Towardsdatascience. <https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b>

Wiegand, Greg, Saood, Shavaiz, and Shea, Richard. (2018, June). The Art of Employing Data Science to Improve Cost Data Analysis. International Cost Estimating and Analysis Association 2018 Professional Development and Training Workshop. <https://www.iceaaonline.com/ready/wp-content/uploads/2018/07/EA10-Paper-The-Art-of-Employing-Data-Science-Wiegand.pdf>.

Wilson, Josh and Baker, Laura. (2016, June). Integrating Cost Estimating and Data Science Methods in R. 2016 International Cost Estimating and Analysis Association Professional Development & Training Workshop. <http://www.iceaaonline.com/ready/wp-content/uploads/2016/06/PA11-ppt-Integrating-Methods-in-R.pdf>.

References

Svelhak, Chris. (2019, May). Clearly Communicating Your IGCE To Decision Makers: The Art of the Outbrief. 2019 International Cost Estimating and Analysis Association Professional Development & Training Workshop.

<https://www.iceaaonline.com/ready/wp-content/uploads/2019/06/CV05-Clearly-Communication-Your-IGCE-to-Decision-Makers-Svehlak.pdf>.

Roye, Kimberly and Smart, Christian. (2019, May). Beyond Regression: Applying Machine Learning to Parametrics. 2019 International Cost Estimating and Analysis Association Professional Development & Training Workshop.

<https://www.iceaaonline.com/ready/wp-content/uploads/2019/06/ML06-Paper-Beyond-Regression-Applying-Machine-Learning-Roye.pdf>.

McDowell, Jeff and Clark, Courtney. (2021, May 20). Data With a Purpose: Technical Data Initiative. International Cost Estimating and Analysis Association 2021 Professional Development & Training Workshop.

<https://www.iceaaonline.com/ready/wp-content/uploads/2021/06/MLD06-ppt-McDowell-Data-With-A-Purpose.pdf>.

Roye, Kimberly, Hilton, Dustin, and Smart, Christian. (2021, May). Dealing With Missing Data – The Art and Science of Imputation. International Cost Estimating and Analysis Association 2021 Professional Development & Training Workshop.

<https://www.iceaaonline.com/ready/wp-content/uploads/2021/06/MLD08-Paper-Roye-Dealing-with-Missing-Data.pdf>.

Eden, Jeremy. (2019, May). How to Build a Data Science Cost Estimate with R Studio. 2019 International Cost Estimating and Analysis Association 2021 Professional Development & Training Workshop.

<https://www.iceaaonline.com/ready/wp-content/uploads/2019/06/DM03-Paper-How-to-Create-a-Cost-Estimate-Using-Data-Eden.pdf>.

“AI Stack”. Carnegie Mellon University: Artificial Intelligence. <https://ai.cs.cmu.edu/about>