Managing Data Science:

A Stacked Approach to Integrating Advanced Data Analytics

Tecolote Research, Inc.

2022 ICEAA Professional Development & Training Workshop

John W. Maddrey, Eric Hagee, Kyle Ferris

Observation

"As federal agencies increasingly look to adopt artificial intelligence and machine learning for their missions and back-of-house business processes, they often hit one major early stumbling block: preparing their data, which can include data fusion from multiple sources, cleansing, transformation, validation and publishing. Agencies often have data stored in multiple silos and data lakes, making discovery difficult. In addition, that data is rarely in standardized formats, especially with regard to formats usable by AI and ML. But by applying DevOps principles to their data strategies ***[DataOps]****, they can overcome this stumbling block much quicker, facilitating implementation of AI and ML tools."*

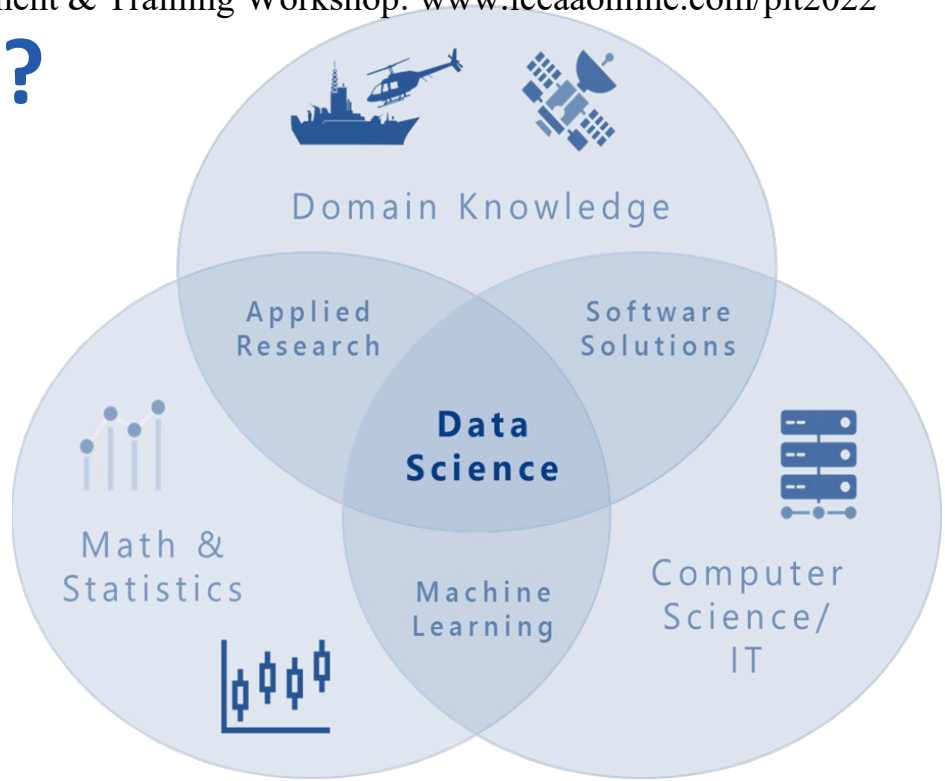
"Agencies should apply DataOps to their data for AI, machine learning", Federal News Network, August 20, 2021. Insight by Geocent.

BIG IDEA

- Integration of data science into cost estimation is here…and it got here quick
- We need to standardize a structure by which we can understand what data science is in cost, and how it operates
- The Data Science Stack is our proposed structure to provide clarity

What is Data Science?

- A multi-faceted field of study involving the manipulation, transformation, visualization, analysis, and modeling of structured and unstructured data.
- Typically involves three main areas of study:



Mathematics	descriptive statistics, hypothesis testing, multivariate analysis
Computer Science	programming, algorithms, data structures, software solutions
Domain Knowledge	federal acquisitions process, program management, systems engineering

Shifting Paradigms

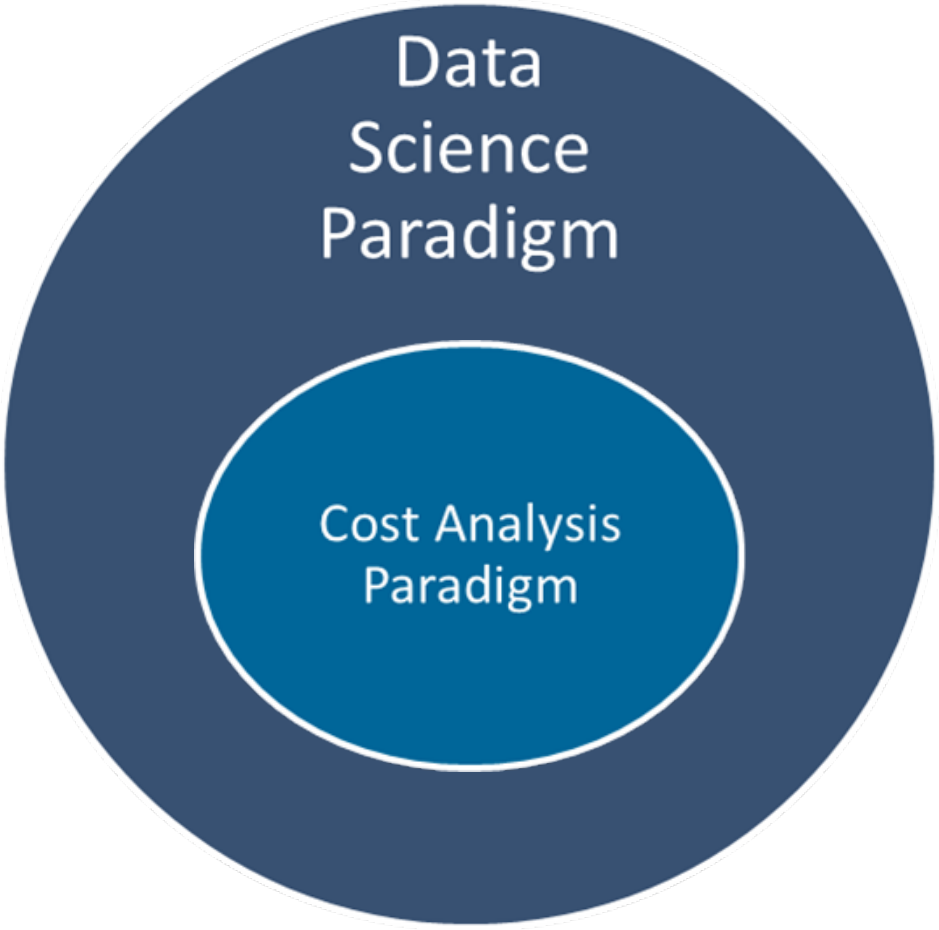
Cost Analysis Paradigm

Small Enterprise Data Sets	Simple Regression and Expert Opinion	Direct Tasking for Specific Product

Data Science Paradigm

Large/Multiple Data Sets	**Complex Data-derived models**	**Collaborative, Iterative Development of Product**

Cost Analysis and Data Science



Functionally, the cost community is engaged in data science.

In many respects, cost is a component of data science.

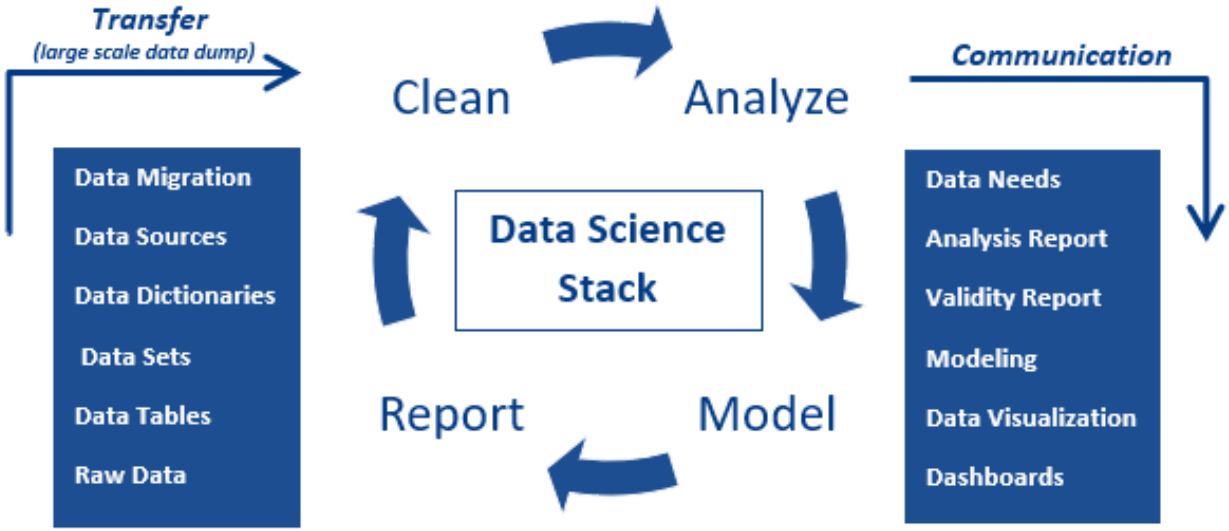
However, the accessibility of computational tools, and a paradigmatic shift in thinking has lead to a focus on data driven solutions.

Challenges and Solutions

Data Challenges	Data Solutions
Where is the data?	Development of data repositories
Is the data good?	Cleaning and maintenance of data
What is the data?	Exploratory analysis
Is the data accessible?	Development of databases

What's Happening?

- In light of the new "data driven" solutions mindset, cost organizations are tasked to go beyond estimating.
- Several cost communities have been setting up data teams to go out, find, collect, and clean data for a whole host of issues related to program management. ***Data is not so much a commodity as it is an expertise.***
- Opportunity to use our skillset and offer data science as a service (DSaaS)



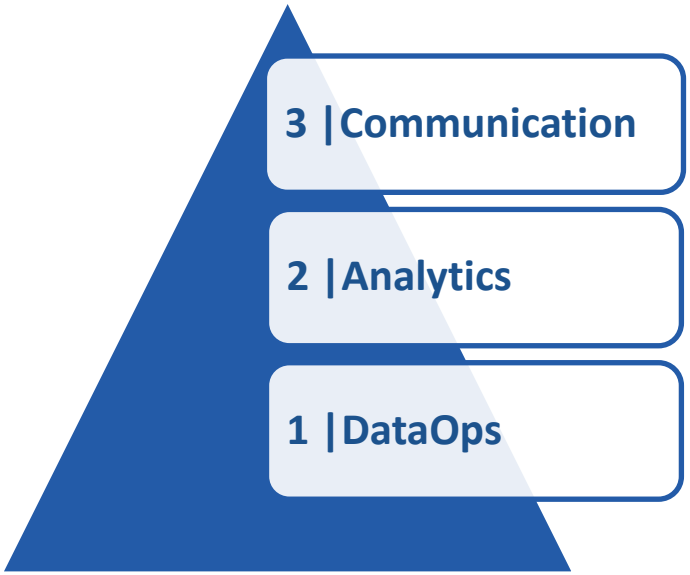
Expectations vs. Reality

- **Setting up a data operations program is new, and, ironically, there is not a lot of data or information on how to do it.**
- **Expect to be a "data investigator", rather than a data scientist**
- *"If we get a group of smart people together, then they can deliver products in a timely manner"*
  - **No level of expertise can compensate for poor resources and support**
- *"We have the data"*
  - **You might not. And if you do, it may not be relevant and/or actionable**
- *"Our enterprise IT infrastructure is well maintained"*
  - **Are you sure about that? In many cases it isn't**
- *"In our industry, it can take years just to get a functioning data set to work off of…"*
  - **That doesn't mean you shouldn't invest in long term improvements**

The GOAT Principle | Align Your Expectations

G	**Grind:** Early on data mining is a manual process, finding data sources is difficult, and validation is time consuming. Getting the tools to make things efficient can take time
O	**Optimize:** Find patterns in the data that make cleaning and mapping data a routine process, setting up timed intervals to retrieve data. Identify individuals familiar with certain data sets
A	**Automate:** Develop scripts that can pull in, manipulate, clean, and validate data to set up for storage or analysis
T	**Thrive:** Enjoy the fruits of your labor…but realistically you will have done a good enough job that people will want you to do the same thing with other data.

Functional Areas



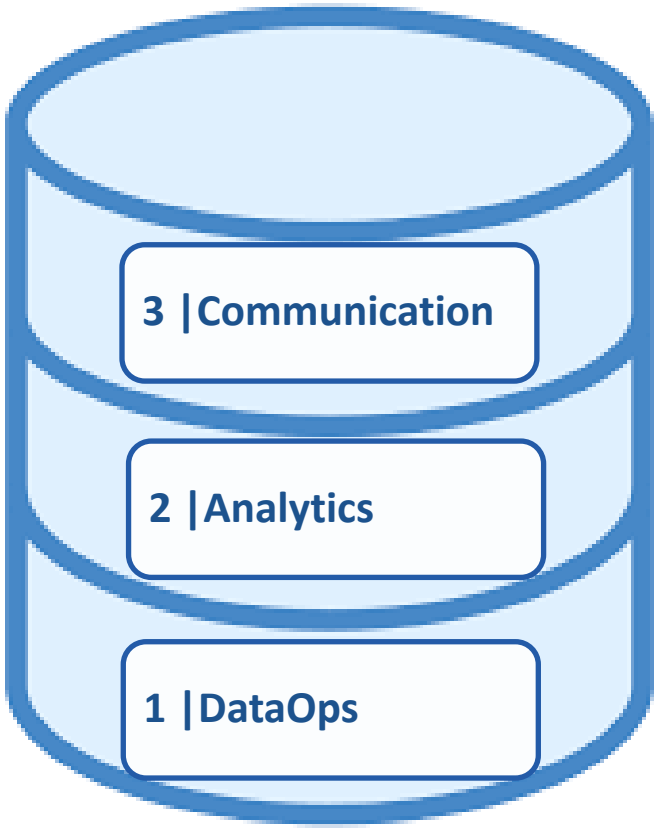
DataOps	• Foundational policies and procedures in collecting, cleaning, storing, and protecting data. • Physical and digital solutions needed to deal with the data. • Data sources
Analytics	• Statistical/Mathematical analysis of the data. • Includes model building
Communication	• The means to articulate the data and the analysis.

The Data Science Stack

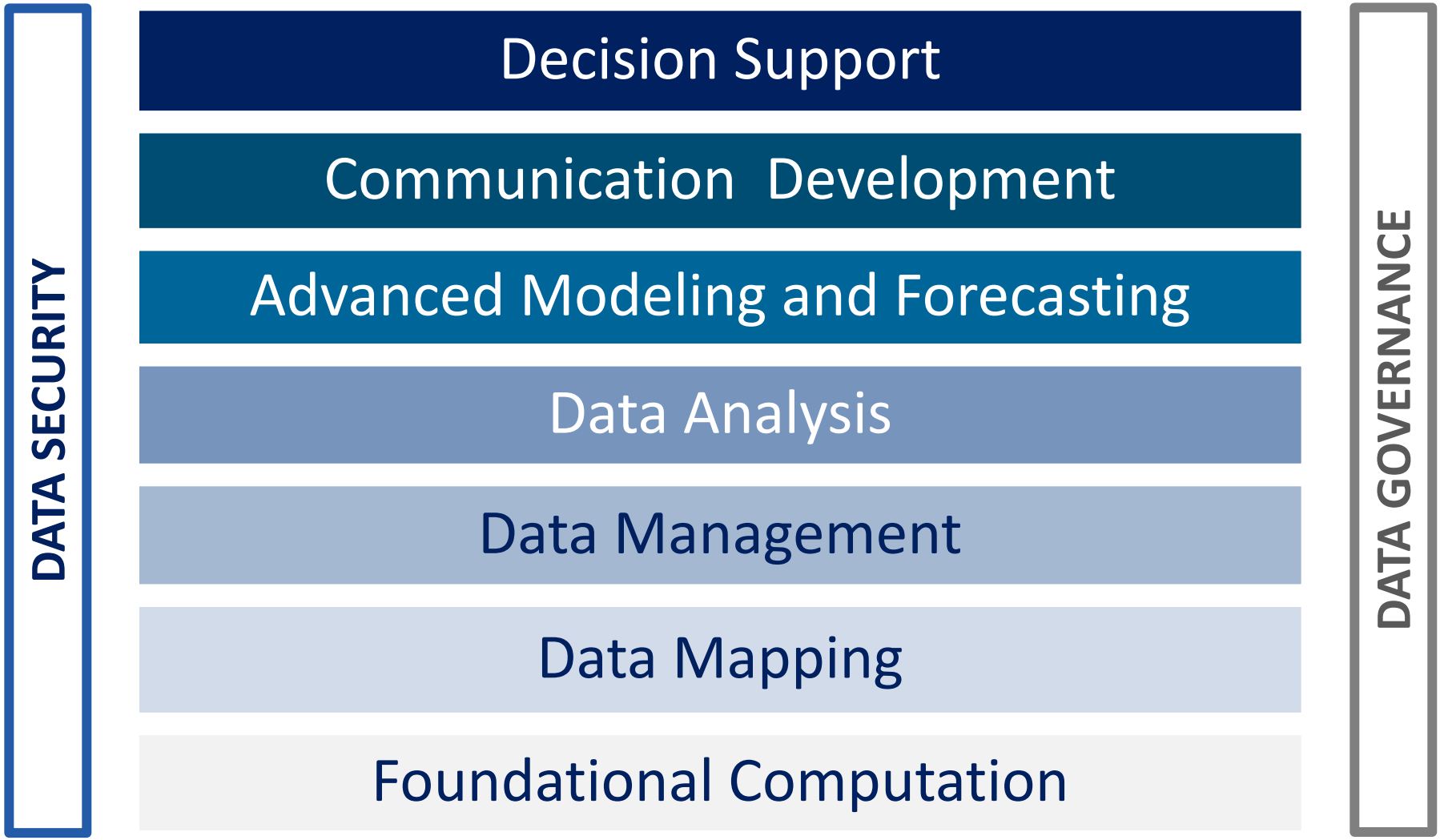
The Data Science Stack (DSS) is a hierarchical system comprised of levels organized to structure, manage, and grow a data science program

Levels of the DSS are organized to be a road map of sorts where each level of the DSS is predicated on development of preceeding levels

The DSS follows a basic structure – communication is built on analytics, analytics is built on DataOps



The Data Science Stack



The Stack | DataOps

- **Foundational Requirements:**
  - Consists of the IT infrastructure, resources and policies necessary to successfully implement data science operations
  - Includes appropriate cybersecurity measures, hardware/software acquisition, and talent acquisition
- **Entry-level of the DSS: establishes an operational environment**
- **Requires management expertise to establish data operations**

The Stack | DataOps

- **Data Mapping:**
  - Process of describing the flow of data in a system. Understanding how data enters a system, is transferred, stored, and validated. This step also identifies users of the data and who has configuration management of the system.
  - Sometimes referred to as data pipelining, spells out how data moves, who moves it, what data sources are being used, and how data (being prepped for analysis) is handled.
- **Anecdotally, many data teams get stuck here. They cannot identify reliable data sources or do not know how to gain access to required systems and/or networks.**

The Stack | DataOps

- **Data Management:**
  - Activities that lay out data acquisition and validation policies. These activities describe the relationship between various data sources, how data is acquired from them, how data is validated, and reporting standards.
  - Here we put pen to paper and set up a Data Governance Package (DGP)
  - The DGP is a standard work package that details how an organization deals with data
- **Note of caution…before you go talking about setting up databases, you should have a working DGP**
- **At this point in the DSS we can veer off and start talking about databases, repositories, data lakes, etc.**

The Stack | Analytics

- **Data Analysis:**
  - The mathematical processes by which new information is deduced from a data set to prepare customizable or operational data reporting
  - Statistical normalization, pattern identification, identification of outliers, etc.
- **This is the first step as to where we deal with data and in any process things can get side tracked.**
  - In addition to doing all of the fun math, we also need to begin the best practice of error reporting. This is a formal process that addresses issues in the data that need to be fixed within our data management.
- **At this point we can move to develop a data research program**

The Stack | Analytics

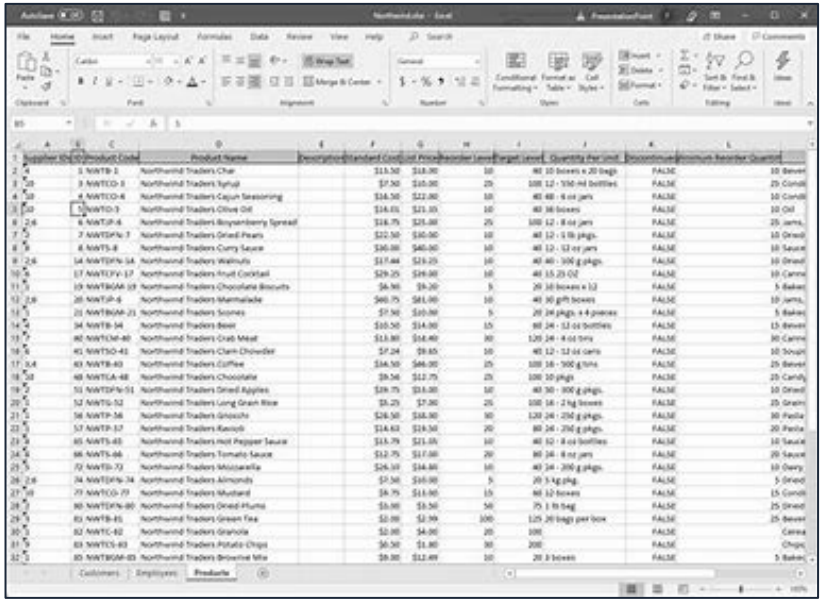
- **Advanced Modeling & Forecasting:**
  - Expansion of classic analytical best practices with the utilization of new capabilities in data science: namely machine learning and artificial intelligence that improve the accuracy, reliability, traceability, and defensibility of our analytical products.
  - By no means are we suggesting that you throw out tried and true methodologies!
  - However, with well maintained data processes and advanced tools… we can do a lot more.
- **This step can be difficult for the workforce to adapt to…please see "The Data Science Paradigm" to contextualize**

The Stack | Communication

- **Communication Development:**
  - The medium by which analytical products are presented. Results should be clearly representative of the data, methodologies should be explained, and data sources/dictionaries referenced
- **Standard or customized data visualization. Each organization may have a different set of requirements for reporting results**
- **Development should be simple and customer focused**
- **Be prepared to develop dashboards or applications…..the future is now**

The Stack | Communication

- **Decision Support:**
  - The inferences and conclusion that can be made based on a well developed data-driven approach.
- **Here we address a myriad of problems that the cost community faces**
- **Better agility and usefulness in a program office**
  - Move beyond a spreadsheet



Why the Stack?

- **The stack accomplishes two things:**
  - First, it provides a framework of activities that build off of each other to structure and maintain a data science program
  - Second, provides something of a work break down structure that allows us to estimate and gauge cost and schedule
- **The goal of the DSS is to engage the community and propose a framework by which a data science process is integrated into existing workflows**
- **The community is in its infancy, and a lot of people are making their own policies and procedures on how to deal with data.**
  - The DSS seeks to take what the community is doing, and structure it in a way that is reproducible.
- **Take these ideas, and start your own data program!**

How to use the DSS

- **The DSS is not just a check list**
  - Development is a continuous endeavor when it comes to data; think of life in terms of sprints, and not as a series of events that leads to a conclusion.
- **In each level of the stack you have different areas of expertise. Use this to find the right people for the right task**
  - You do not always need a team of "data scientists" with hefty salaries to carry out data science activities
- **Each level of the stack is built on what comes before it. But, that does not mean that each level needs to be developed sequentially**
  - Think holistically of how the stack is being developed

QUESTIONS

Thank you all for participating!

References

Ferris, Kyle; Hagee, Eric; Keita, Zoe; Maddrey, John. "Adopting a Data Science Paradigm: Merging Traditional Cost Estimating Methodologies with Computational Analysis". Tecolote Research, Inc., September 2021.

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/.

References

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.

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

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

Bandyopadhay, 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 Devlopment 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.