Department of the Air Force

Integrity - Service - Excellence

Bridging the Gap between Data Science and Cost Analysis

Sarah Green

Sept 2021

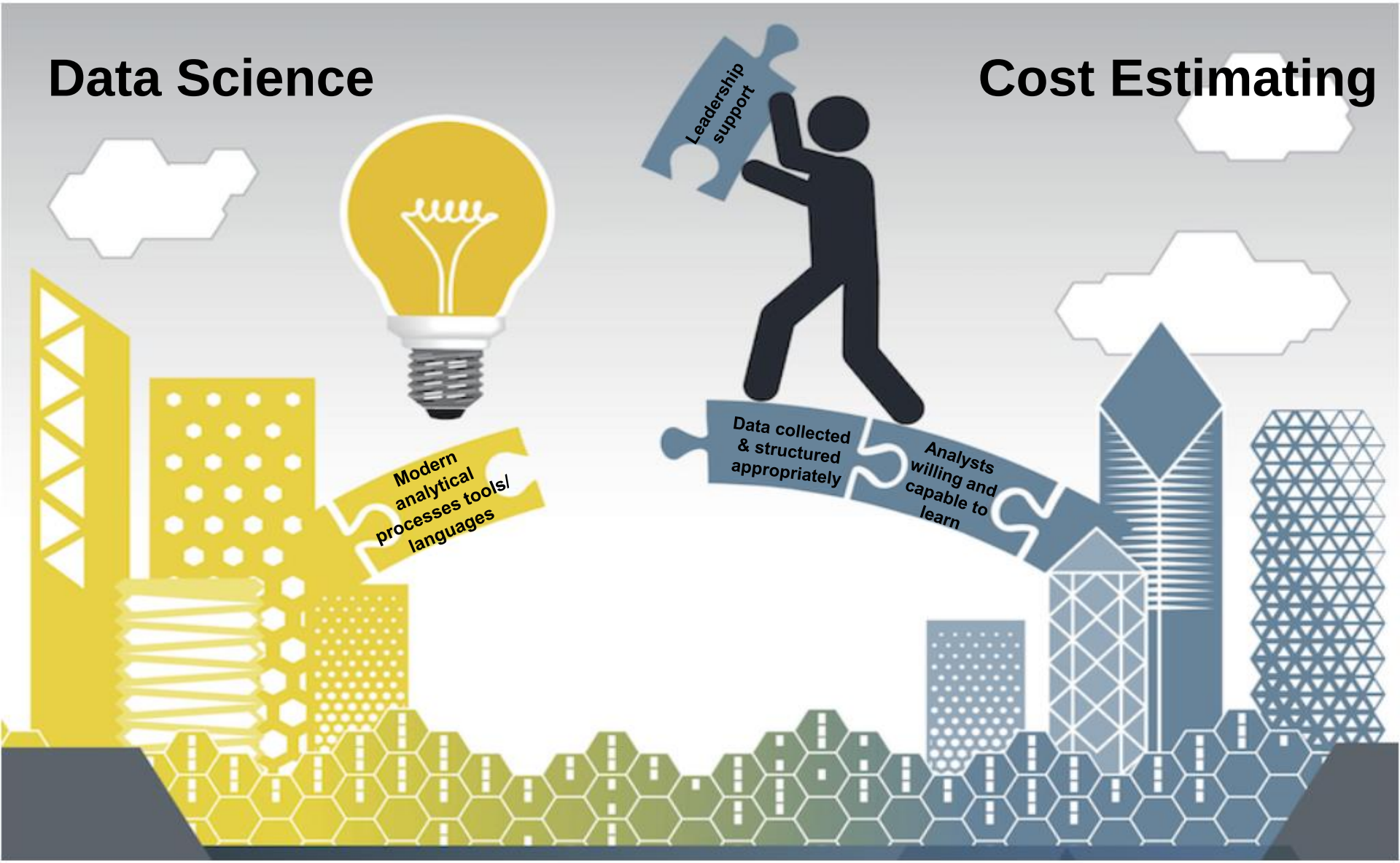
Bad News

- ***Generally speaking*, the cost community is way behind the curve from an data science perspective vs. current state-of-the-art**
  - Our **PROCESSES** are outdated
    - Often, each analyst/team of analysts separately pulls down data from various sources which is repetitive & monotonous
    - Lack of "flattened" structure/format for data, lack of consistent methods/ analysis
    - Result is inconsistent, compartmentalized and unstructured datasets and "ad-hoc"/ stove-pipped analysis
  - Depending on the organization, modern **TOOLSETS** are either non-existent or under-utilized
    - Even with forward-thinking organizations that have adopted some of the industry's best tools and nearly unlimited/free online training, we are years behind the curve
    - CAPE-lead Data Tools Tiger Team Survey to be released soon
      - Hypothesis is that IT/security concerns and adoption are likely the top issues facing the community
  - **PRESENTATIONS** to leadership are often static vs. dynamic

Good News

- **We have more data than ever before**
  - For DoD- Complete overhaul of CSDRs with “FlexFiles”
    - Cost data
    - Software data
    - Technical data
    - Programmatic data
  - Advana +750 data sources
  - Sites like DACIMS, PMRT, EVM-CR etc.
- **As a community (generally speaking) we have the necessary building blocks to succeed in the field of data science**
  - Mathematicians, engineers, as well as a variety of other technical backgrounds
  - Many of our job responsibilities already overlap with those in the data science field

Bridging the Gap



Bridging the Gap – Leadership

- **Specifically what Leadership can do:**
  - Invest in the resources to shift towards data science
    - Free up your current staff and dedicate/train them in data science OR hire data scientists and train them in cost estimating (pros/cons)
    - Surge support is an option but not necessarily recommended since data science is absolutely here to stay
    - The resources you invest in advancing data science will pay off many times over with both efficiencies and improved quality of estimates
    - In many if not all cases of data-driven organizations, data science divisions are being stood up as stand-along entities
  - Ensure access to modern tools & investigate cloud based solutions
  - Incentivize analysts to learn data science principles, methods, languages, tools etc.
  - Plan ahead- what is the 1-5 year plan from a personnel/tools perspective?

Bridging the Gap – Analysts

- **Specifically what Analysts can do:**
  - Be proactive and get trained
    - No matter what your career trajectory learning data science principles and tools WILL benefit you longer term!!
    - If you can meet the technical expectations for a cost analyst then you can master data science principles and tools
    - ALL the training you will ever need on data science is available via open source/ virtual although more traditional classes are an option as well
  - "Flatten" your data tables- this is essential
    - See Adam James' (award winning) ICEAA presentation here for details on how to execute
  - Visualization tools (Tableau/Power BI) have a high payoff and are very intuitive/ easy to learn – do the work & take the training!
  - Liaise and Collaborate with larger cost community to share knowledge and leverage methods/tools currently existing
    - Data Tools & Analytics User group (more on this later)

Benefits

Current

- Traceability is dependent on documentation & process used by analyst (rare to have transparency)
- Only saved versions are kept- can lose trace to data in certain versions of models if not properly handled
- Often have issues with compatibility of desktop versions
- Mostly manual steps – not easily repeatable and often not well documented
- Extremely difficult to get desktop tools approved on high side

Vault

- Complete step by step traceability to original, raw data
- “Time Capsule” - can tell exactly who made change & when – and can revert back to a previous version at any time.
- No compatibility issues with different versions of desktop software once in the cloud
- Automated steps from raw data to final product so that it’s repeatable on new data that’s received
- Can replicate environment on the high side (SIPR now, JWICS 2022)

Benefits, continued

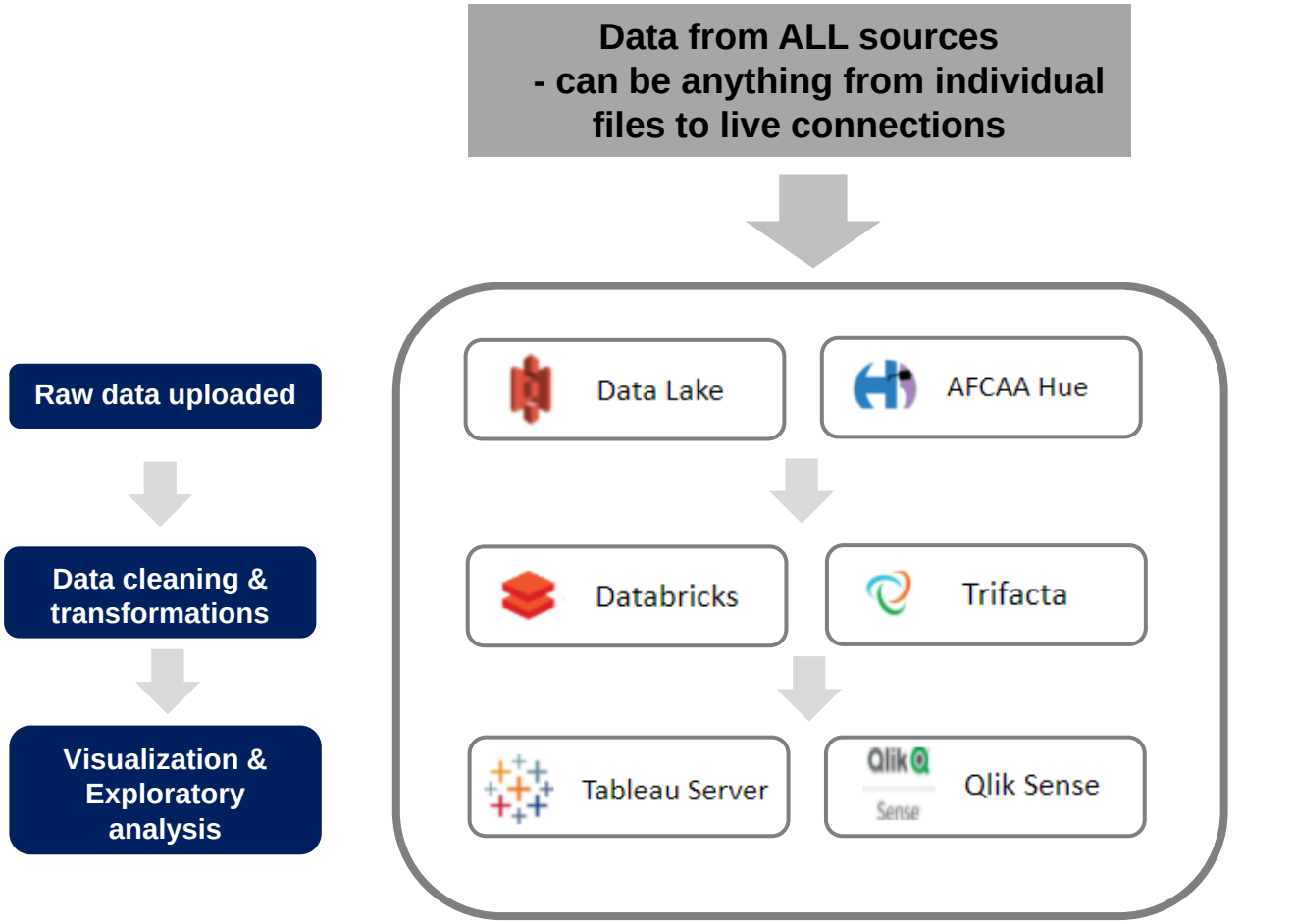
Current

- Performance limited to desktop compute
- Analysts download CSDR data and build relational models on their own
- Mostly access to files is controlled by access to folders in file structures and passwords
- Power of languages like R/ Python extremely limited bc of desktop versioning issues and most importantly the lack of ability of most analysts to learn
- **Analysts need to apply processes/methods independently**

Vault

- Performance will scale based on compute available in “cluster”- equivalent of groups of machines
- Capability to build centralized relational models that will be able to ingest data form
- Access can be controlled and managed easily for each
- For languages like R/Python almost any tools developed can be leveraged in the Vault with one-time set up/debugging effort
- **One analyst, one time= entire organization benefits**

Vault: Proposed Process



AFCAA only Vault "Tenant" Space

Vault Demo

- **Live Demo of Vault capabilities (randomly generated data)**
  - Demo will demonstrate results currently prototyped with multiple tools and our proposed process going forward

User Group Overview

- **What this User Group IS**
  - Analysts from a wide spectrum of different government organizations that are **CURRENT users of advanced data analytic tools** and can represent their organization by talking specifically about them and ideally be able to demonstrate exactly how they are using those tools
  - Liaising with Data Tools Tiger Team (whose mission is to better inform leadership on what is needed to identify, procure and adopt the right tools for the cost community)
  - Government civilians
  - Contractors directly supporting a government organization
  - Meets regularly every 3 weeks
- **What this User Group IS NOT**
  - Providing training for novice users
  - Making authoritative decisions about which tools should be used
  - Industry contractors

Goals of User Group

- **Short term Goals :**
  - Create a community of analysts using data analytics tools to collaborate
  - Discuss each tool in detail to include the different ways that each organization is using the tools to their advantage
  - Demonstrate and share results with the group so we can consolidate best practices and lessons learned
  - Collaborate in order to avoid duplicative data analytic efforts and leverage work that has already been done to the greatest extent possible
  - More widespread outreach to the cost community to help with adoption of tools
- **Potential longer-term goals:**
  - Collaboration on products to be eventually hosted in the DTM Hub

Email me for more information or to request to join: sarah.green.10@us.af.mil

Disclaimers

The views expressed are those of the author and do not reflect the official guidance or position of the United States Government, the Department of Defense or of the United States Air Force.

The appearance of hyperlinks does not constitute endorsement by the Department of Defense or the Department of the Air Force of the information, products or services contained therein