CMS AI Playbook

Centers for Medicare & Medicaid Service

Version 2.0

Table of Contents

1. Objectives & Audience

HHS AI Strategy - 3.1 AI Council and AI Community of Practice (CoP) – "foster enterprise-wide AI adoption by sharing lessons learned, identifying AI opportunities, providing peer recommendations for scaling AI use cases, and supporting shared access to AI tools, resources, innovation labs, and best practices." [1]

This document begins with the principles (Section 2) that enable scalable artificial intelligence (AI), focused on research and development (R&D) and innovation within an organization. Once we discuss those principles and how to apply them within the agency, we discuss the operating model (Sections 3, 4, 5) that allows for rapid iteration, application of learnings and measurement of the impact of change to the organization. We close with the phases (Section 6) of adoption and the key organizational constructs to stand up for effective rollout agency-wide. The AI Playbook Appendix A attachment contains specific projects within CMS where the application of these principles and techniques was used on data sources from the agency's ServiceNow instance.

2. Design Principles

AI programs differ from standard software development projects. In general, more time is spent "cleaning" and investigating the data to be used in AI modeling than in developing the models themselves. The principles in this section define best practices to keep in mind. Section 3 discusses the design phase, in which these principles inform the research processes needed to kick off a successful AI program.

2.1 Think Big, Start Small

Successful programs often require holding two ideas in tension with each other — *think big* and *startsmall*.

To *think big* means to establish an ambitious program vision affecting a specific change over multiple years, and to articulate and re-articulate it to key stakeholders with each progress update. To *start small* means to identify the smallest unit of the program that can be de-risked and demonstrate proven value to begin a virtuous cycle of outcomes and investment with minimal wasted effort or rework.

In contrast, we should avoid *thinking small* and *starting big*. To *think small* is to propose a hard-to-justify allocation of management attention and financial resources. To *start big* is to engage in a program so large, it fails to identify and resolve major risks before a vicious cycle of no outcome, flagging investment and substantial wasted effort and rework takes hold.

Principle in brief: Begin with an ambitious goal that impacts a significant portion of your organization or end users, but make sure that there are smaller pieces within that goal that can be tested quickly and easily. (See Section 3.1, 3.6)

2.2 Work Backwards from Customer Value

Successful programs often require an approach that's less intuitive to AI R&D teams — WorkingBackwards from Customer Value.

Working Backwards means reasoning through all possible valuable outcomes and all the paths that could plausibly lead to them, navigating through what is often a "value maze" of missing data, challenging system dependencies, unexpected stakeholder hesitations, and more. This process typically surfaces one or a few leading candidates to focus on developing further.

In contrast, a more common pattern is for AI R&D teams to be handed a program vision and asked to make progress against it, presumably by working forwards from data and algorithms. However, this way of working forwards from technology often leads to dead ends in the value maze after much effort, resulting in multiple cycles of wasted effort and starting over.

Principle in brief: It's important to start from the perspective of the end user you wish to impact. Begin with the user's overall experience rather than the traditional approach of improving a single application or business process. (See Sections 3.3, 3.4)

2.3 Responsible AI as a Principle

AI is a growing industry with high impact, high visibility, and varying levels of trust, especially in the federal space. Practicing AI responsibly calls for establishing elements of transparency, explainability, accountability, and security, among others. High quality AI programs gain trust and raise effectiveness by valuing responsible AI throughout design, build, and testing phases.

As a design principle, this indicates following a human-centered approach to AI development so that the needs of the user as well as the impact on demographics and society are accounted for. Existing and potential biases must be kept in check, human-AI interactions must be designed for explainability, and consistent maintenance should be made on data and algorithms so that the AI evolves in parallel to the real dynamic environment.

Principle in brief: Take a human-centered approach in your design to ensure your AI endeavors are carried out responsibly to produce the intended impact through the most effective and trusted means. (See Sections 4.2, 7)

2.4 Put Documentation to Use/Make Documentation Useful

Many teams and programs expend excess resources in the earlier stages of an AI project when they start from scratch rather than utilizing existing knowledge and processes, as well as in the later stages when they do not have adequate documentation to present to users and other stakeholders.

Documentation is any record kept providing evidence of work (i.e., research, templates, lessons learned, instruction, explanations) that can be used to support future efforts or decisions. Professional and proficient AI programs should maintain a system of organized, accurate, and consistent documentation that is made appropriately accessible to colleagues, partners, and users.

Principle in brief: Don't recreate the wheel. Use what is available and, likewise, document your work so that the artifacts can be leveraged by yourself and your colleagues to bypass future duplicate efforts and avoidable frustrations. (See Sections 6.2.2)

2.5 Tackle the Big Risks Early

Many projects suffer due to picking off simple problems to start with. While this approach allows for early wins, it sets the project up for the perception of delay later as bigger challenges and larger risks materialize.

AI projects should start with the most complex data sets, the greatest regulatory challenges, and the largest legacy of technical debt. From the earliest releases of the new model, new application, or new data products, users receive maximum value. By starting with the most complex project risks, organizations also ensure that technology platforms, regulatory frameworks, staff skills and project management discipline are mature and will scale as the organization adopts AI more widely.

Principle in brief: An organization's areas of greatest challenge are often the areas that can gain the most from AI programs. Design your AI program to support those solutions from the beginning to ensure that you don't have to re-design later. (See Sections 3.2, 3.5)

2.6 Human Learning Before Machine Learning

Data Literacy is foundational to every AI program and its success. It is the organization-wide programs that facilitate the rollout of methods, processes, technology awareness, and education to move cultural change towards a data-centric organization. The early focus on mindset, culture, skills, and approach, enables the later adoption and scaling of technology and engineering.

While machine learning is valuable work, it should be done after the organization has decided on what skill sets are expected of team members, how data will be consumed, the operating model, and what impact is expected for AI.

This human learning up front accelerates the later decisions on technology standards and creates a culture of data first, solving hard problems early and rapid iteration.

Principle in brief: Lay a strong foundation for AI programs by Investing in the human and organization understanding of AI programs **before** they begin. (See Sections 3.2, 3.4, 3.5)

2.7 Data Rarely Speaks for Itself

Many organizations sit on piles of data collected over many years. This data may yield insights, but should be viewed with caution. Bias frequently impacts this data due to poor processes, manual input methods, poorly-documented changes to applications and workflows, and lack of annotation of the associated data sets.

Successful AI programs pair data scientists with business process experts to create AI R&D teams that work to understand the data relationships, data meaning, and impact of business process changes before presenting results. This Exploratory Data Analysis (EDA) is fundamental to finding meaning in data, communicating it with appropriate business context and collectively working to define business processes to measure future improvements.

Principle in brief: Pair data scientists with business process experts (i.e., the data owner) so that the context is well understood. This partnership also ensures that the program solves the right problems. (See Section 3.1, 3.2, 3.4)

2.8 AI Mindset vs BI Mindset

Data Literacy should focus on building an AI mindset, one that focuses on augmenting the human decision makers' expertise with recommendations, automated decisions and work prioritization. This enables faster, more effective decision making for the organization.

The business intelligence (BI) mindset is familiar, focusing on dashboards, reporting, metrics, and relatively static representations of how an organization executes. In contrast, by focusing on how and where they apply AI to the business process, organizations can realize significantly more powerful decision making, augmenting their teams' work while improving in real time in conjunction with human-AI feedback loops.

Principle in brief: AI programs should go beyond static data displays to real-time, adaptive assistance to human decision-making. (See Sections 3.5, 3.7)

2.9 Augmented Intelligence Before Autonomous Intelligence

AI projects should begin with the goal of augmenting human decision makers with better information through AI. This can come in the form of recommendations on next actions, data relationships and the reasoning behind them, or risk areas for evaluation and resolution.

By beginning with augmented intelligence, we build data-rich applications enabling humans to make final decisions, execute on actions, and identify errors in recommendations and paths from the AI models. This identification is key to ensuring that error conditions are captured. The “right” answer is collected and annotated with the reasons to enable data scientists to improve future models.

Model improvement allows organizations to establish measures of when autonomous intelligence is appropriate and relevant for the organization.

Principle in brief: The goal of an AI program should not be to automate decisions or to replace the human element. Instead, they should augment a person’s ability to make informed decisions. (See Section 3.7)

2.10 Better Data Beats Fancier Algorithms

AI programs should prioritize using EDA and experimentation to improve existing data sets before focusing on measures of model accuracy and speed.

Many organizations are saddled with data sets that are messy, lack uniformity in feature availability and have variances in quality that affect the ability to build effective models. Early EDA identifies these areas of data concern, allowing teams to determine the best way to resolve them. There are several options: improve application and business processes with a focus on the quality of data collection and build the necessary frameworks for experimentation including the associated centralized services (like data catalogs and feature stores).

Principle in brief: Garbage in, garbage out; data sets should be cleaned before changing algorithms to improve model performance. The EDA process can point to the most critical areas for cleanup. (See Section 3.4, 3.7)

2.11 Design for Scaled Experimentation on Day One

The objective of AI at scale is to quickly adapt to large and dynamic data ecosystems. It allows teams to iterate through new data sets, understand the impact of business process changes, and assess the potential of new types of data being collected, purchased or shared across agencies.

To support AI at scale, just as with smaller AI projects, organizations should first agree on the skills, approach, and mindset before setting technology standards. Staff should be trained on collaboration for AI projects, exposed to new ways to divide work amongst team members and how to work with outside and third parties to enable higher levels of scale as organizational needs grow and evolve. Soon after, technology standards should be defined to enable simplified access to data, collaboration amongst team members and alignment with data literacy programs.

Principle in brief: Adaptability requires a scalable foundation; when an AI finding prompts a change, staff skills, technology standards, and infrastructure should already be prepared to handle the change. (See Section 3.2, 3.5)

2.12 Ethics by Design

Ethical considerations evolve as the use of analytical techniques continues to expand and data volumes increase. These ethical considerations span equality, unemployment, security, and legal accountability. All new data science projects should assess their ethical impact and work to define early metrics to quickly determine if the outcomes increase or decrease inequality. Clear, concise reporting on the decisions influenced by analytical tools maximize transparency and increase institutional knowledge of recognizing and mitigating or removing bias.

Principle in brief: Include metrics to determine impacts to equality from the beginning and be prepared for those measures to change as the AI program evolves. Subject matter experts (SMEs) and data owners are good sources for determining these metrics. (See Section 3.5, 3.9)

2.13 Security & Privacy by Design

AI R&D teams are confronted with the growing need for ethical data use and the concurrent expansion of new laws and regulations requiring the disclosure of how data is used. Teams should take compliance measures into consideration at the very earliest stages, while also considering the necessary frameworks, data literacy objectives and technical components for reuse to ensure compliance with legal and agency requirements.

As AI projects begin, organizations should ensure that early standards are set for the tagging of data, the inventory of data sets and the tracking of regulatory and legal requirements against those data sets. While automation of enforcement is a lofty goal, manual methods may be considered at the beginning. Enforcement methods can improve incrementally as the project progresses. The important requirement is to begin enforcement from the beginning.

Close partnerships between data consumers and privacy teams build trust and a shared understanding of data owner expectations and consumer obligations.

This partnership can be used to develop formal training and education plans for deployment through data literacy programs across the agency.

Principle in brief: You can't ensure the security and privacy of data if you don't know what you have or what's expected. Institute a policy for tracking data and compliance needs from the beginning. (See Section 3.9)

2.14 The Empowered AI R&D Team

The most impactful AI teams are the ones that can execute quickly, ensure they are meeting the needs of agency stakeholders and build growing relationships that reinforce the earlier dynamics leading to successful AI adoption at scale.

Agency leaders should ensure that AI R&D teams have the full context of agency needs, challenges, constraints, and unknowns. Teams in possession of this context make decisions based on ground-truth across the agency, shifting focus to maximize impact and align with agency needs as they iterate on experiments.

As development moves forward, stakeholders should be available to review the AI R&D team's results and to ensure findings are shared widely for adoption, feedback, and iteration.

Principle in brief: Empowerment is an ongoing process; AI R&D teams need both the autonomy to experiment and organizational context throughout the program's existence. (See Section 3.1)

3. Running a Lean Research & Design Phase

With the principles from Section 2 in mind, we enter into Research and Design. For AI programs, it is essential to spend time learning about the people, the data, and the infrastructure involved before building anything for production. We'll cover the Engineering phase of an AI program in Section 4.

3.1 Stakeholder Expectation Setting

Successful AI adoption comes from a partnership in exploration, understanding and actions between AI R&D teams and the functional stakeholders that can take findings in data and turn them into operational changes across the agency.

The stakeholders must dedicate significant time in the early stages of any program to become partners in the exploration process. They must understand the outcomes of experiments and help theteam calibrate after failures and learnings. The stakeholders are there as partners to help locate new data sets for exploration and to provide guidance to AI R&D teams on what findings mean in a larger business context.

Oftentimes the stakeholders are key to understanding organizational changes over time that occur inthe data sets and large one-time events located by the AI R&D teams.

Key Action Items for This Phase

- Define AI program vision in collaboration with stakeholders - remember to *think big,start small.*
  - Select a small program unit for early experiments.
- Identify all stakeholders
  - Be sure to include the SMEs who can provide context for the data.
  - Have additional stakeholders been identified during the EDA process?
  - Has User Research (See Section 3.3) identified additional stakeholders?
- Help the stakeholders understand the differences between an AI program and traditional application development.
  - Empowered teams need full agency context both at the start and as changes occur, which may be best provided by the stakeholder.
- Address the potential benefits and risks associated with this AI program.
- Determine the methods for sharing findings with stakeholders and respond to feedback.

3.2 Operations Research

A strong AI culture has its foundation in research-based approaches for all aspects of understanding, hypothesis development, testing and validation. Operations research brings analytical methods to the forefront for breaking down business processes and outcomes into their respective components to allow for measurement, experimentation, and improvement.

A strong operations research team will identify key personas across the organization and the journey between them that enables the desired organizational outcomes. These journeys can be broken down to allow for localized testing of new strategies, focus on data improvement and measurement of impact.

Operations research is a partnership between business analysis teams, AI R&D teams, and SMEs, but always with an eye towards what is possible through experimentation and process changes.

Key Action Items for This Phase

- Identify the owners of all impacted systems and included them as stakeholders.

- Identify SMEs to interview.
- Identify existing dependencies between applications, tools, and business processes.
- Determine what skills are present in your teams and where training could be added to your Data Literacy program (See Section 6.2.1).
- Understand stakeholders' and/or SMEs' "wish lists" for how the AI Ops pipeline would function.
- Understand how the applications and systems that capture data may affect the data: incomplete, incorrect, or corrupted data.
- Understand where business process stakeholders could most benefit from augmented decision-making.
- Determine where the *big risks* lie - what challenges exist in current business processes, what gaps prevent integrating data sources, etc.
- Identify any specialized needs inherent in the existing tools that access the data sets to be used in the AI program.

3.3 User Research

As a subset of the operations research focus, agencies must look at their user base and ensure that they have a full understanding of users' needs, outcomes and the potential obstacles to those outcomes. The focus of user research is to carefully document the users' journey through the application or process from which data is sourced. The best practice method is to develop a persona, which is a fictional representation of a user type that serves as a point-of-view for navigating the application. These personas enable researchers to create hypotheses about what interactions will improve the experience and outcome for the user base. This process generates additional data about the experiences and outcomes that can be used to prove or disprove the hypothesis, generate additional experiments to execute and allow deeper visibility into the outcomes of our business processes. Many of the principles from Human Centered Design (HCD) form the foundational approach into this research. The goal of user research is to understand the needs of the user community and feed the HCD structures within the organization for ongoing feedback and measurement of impact.

Key Action Items for This Phase

- Conduct interviews to understand the experience of end users, business process users, and others with the existing business processes.
- Pinpoint the pain points experienced by each type of user in the current business process and/or application.

- Design user personas to be emulated from a combination of end user and SME input.
  - Consider the full life cycle of that persona's experience with the agency.
  - Decide if additional data sets should be incorporated in the AI program.
- Learn where the greatest value can be found for each user type.
- Identify areas of potential bias to the data through discussions with end users and SMEs.
- Analyze potential bias in the experiments or questions to be answered via the AI model and determine mitigating factors.

3.4 Data Research

Every person in the organization consumes data, enriches it with business context, and produces additional data throughout the journeys they participate in across the business. This growing volume of data provides insight into not just activities and measures of success, but also points of inefficiency, points of friction and other areas for possible intervention for improving outcomes and execution.

Data Research, the EDA portion of all projects, is critical to understanding the relationships across our data sets and the implications of those relationships. These implications can be unexpected outcomes or engagement points to be leveraged to improve outcomes. Bias is one of the components to be considered early to identify ways to test for and prevent bias in data and models.

Data Research expands beyond first-party data and includes third party data the organization can acquire for later enrichment or detailed analysis. Third party data adds additional complexities for purchase, right to use, timing of destruction and downstream decisions that should be carefully understood in conjunction with measuring the value of the data to determine if the overhead of third-party data sourcing is worth the added value during analysis and decision making.

Key Action Items for This Phase

- Understand the links between the data sets under consideration for the AI program.
- Reach out to any other organizations whose data would complement the AI program.
- Learn how each person in the organization interacts with the data sets.
  - How do they change the data?
  - Do they have to manipulate the data for consumption?
- Determine if data quality can be improved through training programs for those who interact with the data.

- Document how the business context impacts the data
  - How, when, why is the data collected?
  - How do data sets relate to one another?
  - Is there any duplicate data?
  - What is the age of the data?
  - What is the size of the data sets?
  - Limitations on data usage or sharing?
- Understand the source of the data
- Understand how the data has changed over time
- Understand the type of data to be considered
  - Is this personally identifiable information?
  - Is this purely technical data?
- How can data be cleaned to improve model performance?
- What data would be appropriate for the AI program to ensure that sufficient data is included.
- Understand the areas in which augmented decision-making will most benefit the agency.
- Interview data consumers to understand their reporting needs, and how the AI program's results could improve or replace them.
- Calculate the size of the data sets for your AI program to determine technology needs. Be sure to consider future growth and the addition of new data sets.

3.5 Systems Research

Today's systems were not built for the modern era of AI-powered recommendations, decisions, and rapid experimentation. They have many limitations including inflexible data models, lack of exposed integration points, scalability and performance limitations, and end-of-life commercial software which leads to operational challenges.

Once an organization has an understanding how to appropriately experiment with a growing number of data sets, it will be critical to build the necessary systems and platforms to enable data collection, analysis and modeling, in addition to enabling strong collaboration between team members. By defining standard technology stacks for team members to consume, collaboration can be maximized and the focus of data literacy programs can ensure the necessary technical skills across the organization.

Modern systems should be built with small, modular components to enable scalability, rapid changes brought on by agency needs and enable rapid iteration of technology as the market matures and new capabilities become available.

Key Action Items for This Phase

- Determine on what data structures are the existing data sets are based.
- Identify the hardware on which the data currently reside.
- Determine the final disposition of the data sets.
  - On premise servers
  - Cloud
- Select a training data collection strategy.
- Focus on the system components that present the greatest risk.
  - Should AI data engineering be done with existing technologies or move to the cloud?
- Develop any training needed to accommodate any changes recommended by the AI program.
  - Should any training stakeholders be identified and included in the program.
- Understand the computing power needed to run your team's experiments. Be sure to include future growth and scalability in your calculations.
- Calculate how quickly new data is added to the data sets.
- Decide the tolerance for latency between data acquisition and model updates.
- Choose a method for tracking experiment results.
- Select technologies for the Data Science Toolkit (See Section 6.1.2) for your organization.
  - If one already exists, deploy it.

3.6 Generating Testable Value Hypotheses

Generating multiple testable theories early in the process allows AI R&D teams to experiment and identify the most promising outcomes for future focus.

The first step is developing hypotheses about possible interventions with the user population. Each intervention identifies a targeted improvement, the timeframe for the experiment, the segment of the population to participate, and the data necessary to use and collect during the experimentation phases.

Each experiment and its output data should be leveraged to generate the next round of hypothesis, segmentation, and targeted outcome improvements.

Key Action Items for This Phase

- Be sure to *start small*. Identify a small experiment that can be explored as part of EDA.

- Use the results from Operations, User, and Systems research to determine where the most value can be found. Design experiments to achieve that value.

3.7 Rapid Prototyping & Hypothesis Testing

The intersection of reusable platforms and technology components, coupled with the ongoing stream of testable hypotheses enables organizations to quickly mock-up experiments, measure the outcome and determine if the experiment warrants either further research, a change in approach, or yields no value to the organization.

The rapid prototyping processes must be fed with continuously increasing data volumes from across the organization, augmented by third party data where appropriate, for complete visibility into experiment outcomes in a rapid period of time.

One key lesson for many organizations is that the first few experiments will often be throw-away work. These experiments are to build team cultural dynamics and depth of understanding of both the data and how it represents our personas and journeys. It will often be run on lower quality data. These experiments are still valuable as a learning tool for the organization and only strengthen the approach, methods, communication patterns and expectations for the team.

Key Action Items for This Phase

- Focus on areas in which augmented decision-making and real-time models can most benefit the human element.
- Identify data sets that require cleaning and data sets that can be integrated into the AI program to improve results before relying on algorithms to do the same.
- Make note of the reasoning and results as models are tested and trained.

3.8 Designing for Production

Production operations take on new meanings in data-centric organizations. The need to ensure availability of data with scalable compute resources is critical to ensuring AI R&D teams can operate at high velocity with ever growing data volumes and complexity.

Platforms should be built with privacy by design approaches and zero-trust architectures to ensure data protection at each stage of processing. Zero trust architecture ensures that small, discrete services across the environment can become collectives for application deployment, while maintaining flexibility and security boundaries.

Key Action Items for This Phase

- Design and implement systems for tracking the sources of data to be used in the program.
- Design and implement systems for tracking the relevant privacy and security risks.
- Determine the performance requirements for AI system components.
- Determine the performance impacts of any changes recommended by the AI program and include them in reporting to stakeholders.
- Include any system impacts or requirements of changes recommended by the AI program in reports to stakeholders.

3.9 Planning Engineering

Planning engineering is the focus on developing the most economical engineering solutions for a defined problem set.

Oftentimes the speed of data analysis and minimizing time to initial findings is more impactful than increasing the accuracy of the finding or data set. Leveraging planning engineering allows teams to evaluate the necessary time to build new capabilities, while jointly looking at the impact of building pieces along the way. This evaluation allows planning teams to build the minimum necessary capability for proving an outcome and beginning to provide user access, while building the necessary foundation for future iteration and improvements.

Key Action Items for This Phase

- Engage project or program managers for stakeholder management, SME engagement, and other communications issues.
  - Assist with coordinating human learning and training needs
  - Understand and guide the AI development cycle
  - Manage equality metrics and coordinate their evolution
  - Coordinate security and privacy tracking and reporting
- Determine the planning framework under which the AI program will operate.
- Develop a roadmap along with the AI R&D team.

4. Running A Lean Engineering Phase

In this section, we discuss the key considerations when taking a small AI program (as described in Section 3) and expanding it at scale.

4.1 Aligning Upstream and Downstream Impact

An AI project is not fully complete until it fits seamlessly into existing processes in the organization. Your initial research and development would have supported the primary user with their goals and environment in mind, but now you need to step back and consider the potential impact your AI may have outside that scope, on upstream or downstream processes. Should side effects be discovered, it is important to implement measures to address them. Once all loose ends are tied, the AI is more likely to be accepted and trusted for scaling.

Key Action Items for This Phase

- Interview users and their related colleagues to ask for significant changes to other areas of their workflow resulting from using AI.
- Suggest or build out solutions to impacted streams.

4.2 Applying Responsible AI

A responsible AI-driven organization requires a sustained effort to scale responsibility alongside their AI programs. Domains in consideration for new AI projects should be reviewed for the appropriate domain mindset, ethical sensitivity, and need, in addition to technical capability.

Teams working with AI should establish and document their responsible AI requirements and cadences to monitor and assess their systems with.

Accountability should be clearly outlined for each AI project. Predefining roles and responsibilities amongst teams and members, as well as procedures for mitigating issues that may arise, builds a responsible organizational framework to adhere to. Accountability promised and held by the team builds trust amongst all stakeholders.

Upholding a dedication to human-centricity, AI governance, and societal and legal compliance are ways teams can promote best practices and build a responsible and effective AI organization.

Key Action Items for This Phase

- Define responsible AI metrics and an established cadence at which to review them overtime.
- Define roles, responsibilities, and procedures for accountability.

4.3 Using Repeatable AI Ops Patterns

Operating AI systems at scale requires specialized platforms, skills, instrumentation, and processes to ensure complete visibility into the intended outcomes. This definition, instrumentation, and operation in a measured way, is AI Ops. This instrumentation is key to ensuring a reliable AIOps environment that alerts teams proactively when models drift from their intended recommendations, when performance could impact user experience or data is not of appropriate quality for decision making.

The organization's systems strategy must include platforms for the storage of data sets, the computational capability to analyze them, and the necessary feature stores and data catalogs to locate data. These components must all be integrated with proper instrumentation for AIOps for visibility into outcomes and early altering for engagement.

Key Action Items for This Phase

- Use a containerized workflow that helps developers collaborate across different platforms and minimizes delays in going to production (e.g., Orbyter [2]).
- Creation of platform architectures including data storage, compute & access.

4.4 Avoiding Black Boxes

The ability to describe the output and the reasoning behind analytical models is critical to both building trust in the AI application and to meet emerging regulatory standards. The ability to understand why models behave certain ways ensures that organizations have eliminated biases from the training sets and that specific segments of users will not be harmed or inadvertently affected by AI outputs.

Model governance standards for the organization should identify both online and offline methods for describing the actions and outcomes of models, as well as identifying the associated input data used for specific results.

Containerizing models allows for them to quickly be shared with outside auditors and regulators, as well as input data sets for validation of their actions and outcomes. Outside teams can be leveraged that unintentional model bias does not enter the organization.

Negative testing should be part of all model release activities to ensure that models fail in prescribed ways and adverse results are caught before being acted upon.

Key Action Items for This Phase

- Always use simple, interpretable models before moving to more complex ones.
- Understand and discuss tradeoffs between blackbox models before investing engineering time.

5. Procurement

Procuring capabilities that include exploratory data analysis, machine learning or other data science work will more closely align with specialized technical projects than with traditional technology or IT procurement.

Different skills will be needed for “zero to one” product development, organizational transformation, and operating and developing models at scale. Each requires a different approach to defining expectations that can be put into procurement contracts and different skill sets to effectively execute.

Agency teams will need to engage closely with AI R&D teams, partnering with them to define specific outcomes to target and the necessary skills and approaches that will most closely align with the current and target maturity levels of the organization.

5.1 FAR Regulation

While today's Federal Acquisition Regulation (FAR) does not call out specifics by the way of AI and data science, the expectation is that they will continue to include these categories of procurement. The expectation is that FAR will grow to include ethics of AI, applicable usage, impact planning for the federal workforce and a reminder on requirements for companies to operate within the United States for work that is highly specialized and impactful using AI.

5.2 Program Requirements

When procuring AI capabilities, agencies should ensure they specify early in the process the type of procurement being made. Categories can include technology platforms, analytical models, intellectual property and R&D services. R&D services are the most critical in early phases of agency adoption because they inform the areas of interest for an agency to focus on and the broad impact of AI to agency direction and operating model.

Procurement in these early stage and R&D projects should not be constrained to specific deliverables, but rather focus on types of data sets and logical domains with set periods of time for joint exploration with agency experts for identification of potential outcomes and agency operational changes to take advantage of the findings.

5.3 Market Conditions

The AI market is extremely fragmented and brings a wide diversity in capabilities, domain of focus, company scale and costs. When procuring capabilities in the AI space agency leaders should focus on companies that bring an R&D focus.

While the market for talent here is hot and leading to higher costs for resources, these types of individuals bring a unique combination of experience and capability to move quickly in small teams, to identify early impactful data opportunities, and to turn them into technology and patterns that are widely consumable across the agency.

This is a rapidly evolving market, with a wave of expected consolidation, new startups and new approaches in the next 18-24 months. This speed will enable new capabilities and approaches available to CMS from the open market. Procurement should focus on projects that have definable outcomes over quick periods of time, allowing for shifts in procurement strategy over time as market conditions change and new capabilities become available.

6. Enterprise AI Project Phases

In this section, we discuss the key phases of rolling out an AI program agency-wide, starting from a small point project and then moving to scale. These are the major activities that will enable growth of the AI program in a standard, scalable way. As an organization matures through experimentation, into proof-of-concept projects and finally into a product orientation running AI models at scale, there are specific structural constructs required. Each phase of maturity identifies specific elements of agency wide programs that, when rolled out, enable consistency and uniformity in approach and supporting technology.

6.1 Point Projects (Experimentation)

6.1.1 EDA - Techniques & Approaches

The immediate goal for exploratory data analysis (EDA) is to establish small collaborative teams with access to raw data sets, in partnership with data owners for ongoing, iterative review of findings, understanding of their meaning and joint agreement on next steps.

The use of EDA early in projects allows AI R&D teams to explore new data sets, identify the types of data they contain, their relative relationship to one another and quality concerns that may be present in the data set.

EDA is key to enabling rapid experimentation and determining if a data set has analytical value. Oftentimes EDA will not lead to predictive models, but rather business processes changes that are necessary to improve the quality and completeness of data to allow for later modeling.

EDA should be driven both by technical considerations (e.g., identifying the most interesting columns, particularly those with categorical, numerical, and temporal data) and by business stakeholder considerations (e.g., identifying columns and groups that are most likely to yield relevant findings).

Appendix A explores the techniques used in initial CMS projects for analysis of structured and unstructured operations data and begins to lay out findings for the most relevant types of techniques given data types, dimensions, quality and relationships.

6.1.2 Data Science Toolkit

The data science toolkit is a standard set of technologies, packaged for easy consumption and portability to enable Data Engineers and Data Scientists to collaborate and quickly explore new data sets. These data science toolkits, which are available from third party vendors or may be developed in-house, are containerized for easy portability and are deployed to ensure uniform tooling across teams to facilitate collaboration.

The data science toolkit provides a variety of tools for enabling data engineering, data science and collaboration in a uniform way. This toolkit begins with Orbyter [2], a set of Machine Learning tools in a containerized environment for rapid deployment.

The data literacy programs, further explored in 6.2.1, zero in on an organization's chosen data science toolkit and technologies to ensure wide awareness of their existence and deep skills for their usage, coupled with enablement programs to constantly elevate the skills level across the agency.

6.2 Automated Decisioning (Proof of Concept)

6.2.1 Data Literacy Program Definition

While early adoption of AI can be facilitated through third party contracting, incremental hiring and some training, the wide adoption of AI and transformation to a data-centric organization requires broad training, team empowerment and changes to organizational operating models.

The creation of a formalized data literacy program becomes an enabler for the organization to share best practices and organizational accepted techniques, train individuals on analysis techniques and technologies, and ensure that the culture of the organization is transitioned to one that first uses data and experimentation before executing on business process changes.

Data literacy programs should be short term in their thinking. An organization should look one to two years out to define the necessary skills to be successful and to enable the organization to develop them. As the Data literacy program matures, the horizon should begin to look five and ten years out to support hiring ahead of agency needs and alignment with future technology trends the organization would like to capitalize on.

6.2.2 Data Platform Definition

The enabling technology for any AI program is the shared data platform. The data platform provides an environment for location of data sets, storage of exploratory data outputs and a standardized toolkit for organizational use when analyzing data through a variety of techniques.

The data platform becomes the technical implementation of standards for privacy by design, AIOps instrumentation for visibility into model performance and a set of collaboration tools to allow data engineering and AI R&D teams to quickly pass off the result of work between one another for further consumption and refinement.

Data platforms provide a core set of functionality for data landing, linking, enrichment and query. The query layer is defined by the data products that are required by the organization for reporting, data science exploration and model operations. Data platforms commonly contain core services for feature stores, data catalogs and orchestration to ensure uniformity of services across the organization.

6.2.3 Modern Data Governance Programs

While we have talked about privacy by design, it is only one aspect of strong data governance programs. Effective data governance ensures that teams can find the right data at the right time and fully understand the quality and business context of the data. Many data governance programs will focus on injecting appropriate training (including cultural training) to the data literacy curriculum. They simultaneously build shared services and reusable technology assets to enable development teams to move quickly with reusable components that protect data and manage its lifecycle according to organizational policies.

6.3 Machine Learning at Scale (MLOps)

Over time, data and its environment may evolve. Input data (independent variables) may see shifts in schema, quality, and/or distributions that no longer retain the same relationships with the output (dependent variable) as when the model underwent training and testing. Similarly, new policies for transparency/privacy, data collection process changes, etc. require model oversight allowing for deeper understanding of its purpose and efficacy. Finally, as model deployment volume, complexity, and usage increases, data scientists and engineers are met with overcoming each of the above trials while maintaining continuity in their products. Overcoming these persistent challenges – model staleness, monitoring and oversight, and continuous integration (CI)/ continuous delivery (CD) are the objectives of MLOps.

Advanced MLOps applications require procedures for model development, autonomous validation & deployment, and performance monitoring alongside well-documented metadata, threshold notifications, and CI/CD. Not all model development exercises require this depth of capabilities for successful deployment.

In this section, general best practices are provided for MLOps processes, artifacts, and their integrations into deployments ranging from mostly manual to fully scalable infrastructures. Specific products, services, or architectural implementations are not within scope. Several of the underlying concepts in this section can be seen in further detail in Google's Cloud Documentation here: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning [3].

6.3.1 Processes

Whether a team is tasked with developing unrelated models or with the maintenance of a single product serving prediction services, there are typically two high level processes: Experimentation, MLOps. While their processes have significant overlap, it's important to draw distinctions in that Experimentation is the development aspect of a model where MLOps is the deployment and monitoring of the outcomes of said experiments. This sub-section explores these processes at a high level.

6.3.1.1 Experimentation

Within experimentation, Exploratory Data Analysis (EDA) should be employed to foster understanding of the data and its structures and conduct data transformation preparation exercises. This step permits your data scientists and engineers knowledge and features necessary to conduct manual or autonomous experimental evaluations of model developments to identify and implement the best strategies for feature engineering before promoting the resulting code and models into the MLOps pipeline for deployment.

6.3.1.2 MLOps

Similarly structured to the experimentation process, the MLOps process has both structural and procedural aspects designed for handling of promoted experiments' data ingestion transformation & validation, model training & validation, model deployment, and monitoring processes. MLOps utilizes data collected in the form of validation results, evaluation results, data provenance alongside performance monitoring and trigger-based notifications and actions to continuously measure the integrity of feature stores and model predictions.

As processes mature, they move from manual stepwise activities to autonomous deployments & monitoring of experiments promoted to the MLOps process. Consequently, a model already established in the MLOps and serving a production environment will undergo monitoring activities to evaluate data and model quality based on predetermined triggers / automations or manual involvement.

In addition to the above processes, there are several key artifacts and tools required for the success of an adequate MLOps deployment:

Table 1. Key Artifacts and Tools for MLOps

Service	Definition
Proper code sourcing techniques	Used for both experimentation (eg. Notebooks, scripts,etc) and their formalized, modularized counterparts promoted to the MLOps process are integral to the overall scalability of the processes.
Feature Stores	Permit the reuse of various data and structures created throughout the MLOps processes. This permits tracing models evaluations to these artifacts by analyst or autonomous means.
Metadata Stores	Provide valuable context around versioning of models and datasets, training parameters, evaluations, scores, creators, etc. This information allows users to understand model quality, performance, and impact at a point in time.
Model Registries	Data stores for models that provide users with additional methods for accessing models, their information and versioning models with tools such as model cards and other means of storing model metadata.
Triggers	Utilized in examples such as threshold notifications or scheduled events to take action. Whether this is to automate ETL on new data arrival or to notify the engineering team of a concerning evaluation of production performance, using triggers is paramount as model deployments scale

6.3.2 Considerations at Scale

The ability to scale ML products is highly correlated to the maturity an organizations MLOps implementation. It's important that organizations consider the level of maturity required for their circumstances. An academic researcher creating a class demonstration model to predict invasive Lionfish population growth in a region likely has less of a need for MLOps than organizations deploying autonomous vehicles. Google provides an excellent overview of the maturity levels of MLOps on their MLOps: Continuous delivery and automation pipelines in machine learning page [3].

Some models may require training on large or complex data sources. If data reduction practices cannot achieve optimal reductions in training time, utilizing distributed compute tools and libraries designed for them can substantially reduce the model training time.

To ensure that model service provides timely and costly predictions, it's important to consider the best method of ML inferencing. Batch inferencing uses asynchronously logged predictions based on a batch of independent variables and is effective when models don't require the most recent training. Conversely, real-time inference is best employed to predict as requests are received. This method most important when the model requires retraining often to prevent growing stale.

Where possible, containerization of services and modularity in design are recommended for the benefit of CI/CD in the MLOps pipeline as well as their maintenance.

6.3.3 Governance

As the number of models deployed in an environment grows, the need to govern them becomes more critical. Governance includes several dynamics around release management, retirement management, business process connections, and monitoring for unexpected events and outcomes. The objectives of a model governance program are to eliminate inaccuracy that comes from untested releases or biases in data and training. Model governance defines the necessary policies for ensuring that there is a balance to be made between the accuracy of outputs and the explainability of the models. Furthermore, it's important to always employ use of responsible AI techniques and to consider explainability AI techniques when 'black box' products are required.

7 Responsible AI

This section reintroduces responsible artificial intelligence (RAI) in the CMS context and breaks down the framework into 6 key domains to address when developing AI solutions.

7.1 Responsible AI Defined

RAI is a governance framework that defines the principles and practices an organization follows to address the societal, ethical, and legal impact of artificial intelligence. The framework calls for a human-centered approach to AI development that reflects the needs of a diverse set of users and cross-industry ethical standards.

7.2 Responsible AI Relevance

As a major federal actor in the healthcare space, CMS must highly consider RAI when developing and deploying an AI tool or program. The legal landscape surrounding AI is in early development, hence upcoming policies will be directly influenced by the standards set by governmental and regulatory bodies. Efforts at CMS are far-reaching and sensitive, and thus require established structures to monitor key issues, react in a responsible manner, and establish governance mechanisms to limit negative implications that may come with the introduction of something as powerful yet tricky as AI. Every AI endeavor within CMS should set the standard toward RAI practices that are trusted and are worth that trust.

7.3 Responsible AI Domains

RAI is an expansive framework that is still growing and being defined. Due to its subjective nature, there is no straightforward criteria for measuring responsibility.

We introduce 6 responsible AI domains to provide a glance into the complex considerations of RAI and give examples and considerations on their relevance to the development and human-centered design of AI.

7.3.1 Bias and Transparency

Bias in AI is the influence of prejudiced assumptions in data that carry into the algorithms outputting results. Consider: *Does the training data include any sensitive variables? Does the output perpetuate racial, gender, and/or other stereotypes?* Ideally, an AI tool would not tend towards any human, systematic, and/or institutional bias. It is the responsibility of data scientists, analysts, and designers, with the help of stakeholders, to recognize inherent bias, remove as much as they can, and provide transparency to users and stakeholders.

Relevant stakeholders must be made aware of the usage and function of an AI tool within the decision-making process. Best practices are to be upfront with the capabilities, limitations, and intent of the AI product, in addition to transparency into reliability levels and potential underlying biases in the AI's output.

7.3.2 Model Explainability

The decision-making process used by an AI model to come to an output must be contextualized and easily explained to the end-user. Explainability in AI models not only fosters trust and confidence in AI systems, but also allows for more effective AI integration and better informed analysis and feedback from stakeholders.

Moving away from the "black box" approach to AI development has been an important area of research, and tools that provide model explainability are constantly evolving. Consider tools like SHAP and LIME for identifying feature relevance, or DICE-ML for counterfactual explanations.

7.3.3 Interface Explainability

Being able to communicate the value and responsible framework efforts effectively requires purposeful interface explainability. UX designers must consider how to best display outputs provided by the AI algorithm to an end-user and ensure the user trusts the output to an appropriate degree through model explainability and transparency efforts. The user should be given all the information they will need to use the interface easily and intuitively, as well as understand and create value from the results of interacting with the AI.

In practice, creating intentional UI copies which prompt the user to experiment with the interface can strengthen trust and open opportunities for improvement. Consider utilizing tooltips, explanations, and evaluation measures to convey appropriate details and transparency. Suggesting users to take an action as a result of the AI's prediction allows users to better understand why the AI's prediction is useful. Finally, be sure to create channels for feedback for continuous improvement of the effectiveness of your AI tool's interface and overall UX.

7.3.4 Robustness and Security

As with any software subject to data dependence or with a stake in decision-making, it is a key responsibility for AI tools to be built robust and secure. Robustness refers to the ability to maintain performance and accuracy – an essential yet difficult feature to achieve under the highly variable conditions in the real world. Part of maintaining robustness calls for security of the AI tool, which may be vulnerable to attacks by adversaries. Attacks can include manipulating input data to lead to false predictions, poisoning the dataset provided to the learning algorithm, or interception and breaching of sensitive system or user information. Security measures must identify such vulnerabilities and strengthen the AI tool's ability to resist such attacks.

Implementing robustness and security may involve a multifaceted trade-off with other aspects of RAI, so formal and comprehensive decisions must be made to determine the balance appropriate for that AI tool's context and sensitivity.

7.3.5 AI Testing

Thorough AI testing and evaluation should be performed on all RAI domains and must be considered pre-deployment, as AI algorithms are dynamic and change overtime (in contrast to traditional software). Consider: *What tests are needed and at what frequency? Do testing methods align with technical standards and is there a process set for rollback if the AI tool does not behave as planned?*

In addition to unit tests on isolated components of the AI system, integration tests are needed to understand how individual ML components interact with other parts of the overall system (across, up, and downstream). System quality checks should go through exhaustive testing with built in mitigations. User feedback, input drift, and output accuracy should be monitored overtime with updates being made in response to any changes.

7.3.6 Governance and Compliance

AI governance refers to the legal framework that manages an organization's research, design, and use of AI. Existing standards mandate that organizations must monitor and document all attempts to minimize unintentional discrimination and generate good faith justifications for AI use. AI organizations should also make considerations for communicating privacy and security relating to user data within the AI and establishing accountability for the development and impact of AI use.

Laws around AI usage vary across organizations and are changing as AI becomes more widely used. AI tools must be developed with current and future legal landscapes in mind.

8 References

1) https://www.hhs.gov/sites/default/files/final-hhs-ai-strategy.pdf
2) https://www.manifold.ai/project-orbyter
3) https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning#mlops_level_1_ml_pipeline_automation

9 Appendix A – Completed Pilots

A.1 Ontology Development

A point project with Manifold and CPMS kicked off in June 2021 to explore two example data sets: one from the ServiceNow ticketing system and one from the knowledge base articles used for self-service support. The goal was to understand if an ontology could be built from concepts contained in these data sets.

As described in section 6.1.1 EDA - Techniques and Approaches, a small team of four Data Scientists and Machine Learning Engineers (MLEs) explored the data sets to see what information could be extracted from those data sets. We will focus here on the overall discovery process.

A.1.1 Initial data exploration

A.1.1.1 Goals

As is typical for most exploratory analysis, our goals in this phase were as follows:

- Develop an initial understanding of the data (scale/size, relevant columns, degree of missing data).
- Identify trends and patterns and potential future approaches.
- Evaluate feasibility of technical approaches to solve initial business problems.
- Create a set of findings and open questions to begin a dialogue with organizational stakeholders (e.g., business owners, etc.)

While the exact goals and mode of implementation may vary from use case to use case, these goals can serve as a general template for the initial exploration.

A.1.1.2 What We Did

- Examine quantitative variables (ticket duration / resolution time, number of tickets, etc.).
- Examine various data groupings (category, subcategory, assignee, contact type, customer, CMS subdivision, etc.).

- Examine temporal data
  - e.g., category prevalence over time, unusual events like sudden spikes in term frequency, seasonal trends, and obsolete vs. new terms.
- Natural language processing (NLP) analysis of unstructured text (short description, description, close notes, etc.) using term frequency/inverse document frequency (TF-IDF).

A.1.1.3 Outcomes and What We Learned

- Identified key trends and patterns, e.g.:
  - Password resets and account locks were the most common ticket type
  - Resolution time varied by category and sub-category
  - Many subcategories were rare
  - Among cases, financial/enrollment tickets took the longest total time
  - Among incidents, security tickets took the longest
  - Distribution of category/time varied greatly from account to account
  - De-duplicating tokens would be necessary to see more patterns
  - Several categories showed seasonal trends (e.g., training in the fall, remote work, and access at the start of the pandemic, etc.)
- Developed a set of action items for future phases.
  - Selected categorical variables useful for future analyses and comparisons (e.g., identifying category as a key signal).
  - Need for robust, de-duplicated concepts or entities instead of just raw n-grams.

A.1.2 Refine the Findings: Concepts, Entities, and Vocabularies

A.1.2.1 Goals

Based on the results of the exploratory phase, we established the following goals:

- Develop a set of concepts and entities in order to:
  - facilitate future work on ontology learning and taxonomy representation
  - de-duplicate and refine results from exploratory phase

A.1.2.2 What We Did

- Resolve repeated phrases (n-grams) for the same concept (e.g., reset, password, reset password, etc.): the remaining high-frequency phrases capture “concepts” or “entities” in the data.
- Visualized concepts as a graph to show links between concepts based on co-occurrence in tickets.

- Used graph clustering to extract sets of related concepts and their associated categories.

A.1.2.3 Outcomes and What We Learned

- Developed vocabularies of key concepts and entities (one each for cases and incidents), which were useful for all future NLP analysis.
- Unsupervised graph clustering algorithms found concept groups relating to training/education, policy, access, technical assistance, and more.

After this phase, the team met with several subject matter experts (SME) and business owners to discuss findings to date, and to learn what specific use cases might be investigated in the time remaining on the project.

A.1.3 Explore additional Use Cases

A.1.3.1 Goals

Based on extended consultation and discussion with OIT project leads and business owners, the team established the following three areas as goals.

- Explore ability to predict ticket resolution time, reassignment, failure, etc.
- Continue ontology learning by attempting to learn entity relationships.
- Identify links between knowledge articles and cases that would value from the knowledge for accelerated resolution.

A.1.3.2 What We Did

Near the start of this phase of the project, the team acquired an additional dataset: approximately twohundred “tier zero” knowledgebase articles.

- Attempt relationship extraction by searching for sentences with multiple concepts.
- Identify connections between articles and tickets (e.g., suggest articles based on ticket descriptions).
- Developed a new vocabulary based on articles, using TF-IDF (the article dataset was too small to reliably use the entity and concept process described above).
- Build interpretable decision tree prediction models to predict resolution time (binarized) and reassignment.

A.1.3.3 Outcomes and What We Learned

- In entity relationship learning for ontology learning, structured data is ideal. ServiceNow ticket data was found to be too noisy, unstructured, and informal to extract relationships between different entities.
- Prediction performance was generally good, and the team's use of interpretable decision trees highlighted key variables that were associated with fast ticket resolution (common descriptions for known categories, etc.).
- The knowledgebase articles did not cover the most common use cases for incidents and cases. Therefore, most tickets did not have a reliable "best article" suggestion (some of these examples were due to the need for help desk staff intervention, e.g., password reset).

A.1.4 Wrapping up

As is common in point projects, the results were both a body of knowledge developed from the exploration and a set of potential areas for future focus.

The primary project learnings include:

1) Methodologies to extract taxonomy/ontology from unstructured data. This methodology could be generalized in the future to other unstructured datasets.
2) Validated the methodology with a second unstructured dataset to determine our ability to detect relationships between multiple datasets for the creation of future ontologies.
3) Generated several preliminary predictive models with actionable results.

Future focus areas can include:

- Expanding the number of datasets to allow for development of a true ontology.
- Developing automatic outlier/anomaly detection to understand users' needs.
- Pre-incident alerting and user notification (e.g., "We've seen an increase in VPN issues today; if that is your issue, can we point you to a given resource?").

A.2 OHC AI Pilot for Time to Hire Prediction

A.2.1 Introduction

The Office of Human Capital (OHC) AI Pilot team built a proof-of-concept "Time-To-Hire Calculator" that provides clarity into the hiring process for Hiring Managers and supports shorter hiring timelines. Using open-source tools and data from USA Staffing reports we processed the data, trained machine learning (ML) models, designed an explainable and user-friendly interface, then deployed these interwoven components securely to the AWS cloud — all while, drawing on human-centered design processes to verify that the product fit the needs of users.

A.2.1.1 Challenge

The OHC Division of Workforce Analytics and Accountability has access to a variety of raw datasets and wanted to explore ways to make those datasets accessible to CMS employees and support more effective work. With the AI Explorers Program, the Division wanted to address the following challenges.

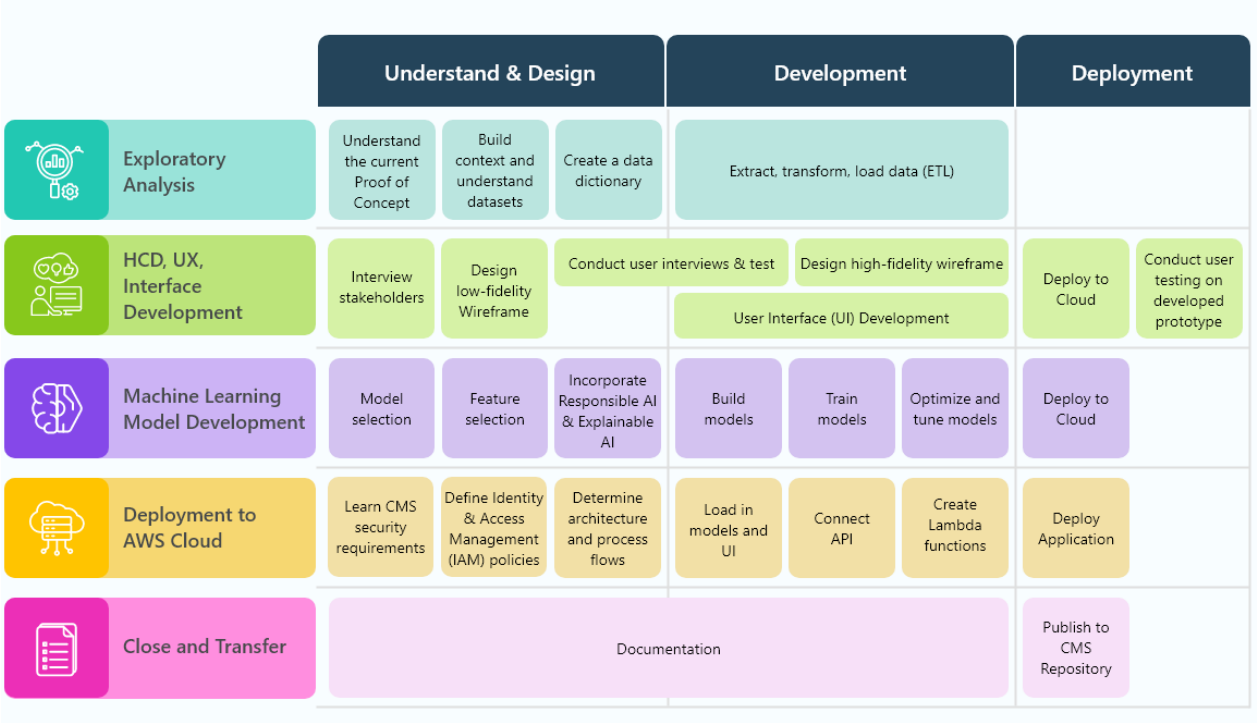
1) Develop a user-centered prototype from a raw dataset by processing the data through ML models and deploying the prototype in the cloud.
2) Ensure the process taken to develop the solution is repeatable and applicable to future datasets.

A.2.1.2 Solution

Our solution to these challenges consisted of the following steps:

1) **Exploratory Analysis**: Understanding the datasets and initial Proof of Concept (PoC) prototype, and building a new extracting, transforming, and loading (ETL) process.
2) **Human Centered Design (HCD), User Experience (UX) and Interface Development**: Identifying business questions, conducting user interviews and testing, iterating on wireframes, and developing the interface prototype.
3) **Machine Learning (ML) Model Development**: Building, training, testing, and evaluating various Python ML models, while incorporating Responsible AI (RAI) and Explainable AI (XAI).
4) **Deployment to the AWS Cloud**: Deploying the tool to the AWS Commercial Cloud in a CMS Cloud environment to provide a path to making the application accessible to users. The team utilized AWS managed services and documented CMS Cloud configuration and CMS security requirements to provide a secure deployment model and a path towards achieving Authority to Operate (ATO).

Figure 1. Process to create a product solution



A.2.1.3 Impact

As a result of this project:

1) OHC has a technically functional and user-centered “Time-to-Hire Calculator” prototype that allows Hiring Managers to assess how long it would take to hire a candidate for a specific job description. Along with the product, we are providing documentation, research, and recommendations for future versions of the product.
2) OHC and the AI Explorers program were able to participate in and learn from the solutioning process. The process is well-documented and can be repeated for future projects involving large datasets, AI and ML modeling, and human-centered design.

A.2.2 Exploratory Analysis

A.2.2.1 Goals

- Understand OHC’s project goals, current processes, data collection, and data insights.
- Assess the initial Proof of Concept (PoC), datasets, and user interface and determine what improvements could be made.
- Streamline and automate the Extract, Transform, and Load (ETL) process.

A.2.2.2 What We Did

- Reviewed the current PoC and related materials (e.g., datasets, ETL flows, ML and UI scripts) to establish a full understanding of the baseline functionality and resources.
- Produced a new data dictionary for the project, and a new ETL process comprised of three Python-based scripts: 'generate_certificate_files.py', 'generate_time_to_hire_regression_data_file.py', and 'TTH_Summary_Stats.ipynb'.

A.2.2.3 Outcomes and What We Learned

- **Understanding the PoC**: Familiarization of the initial PoC's design, structure, and process flows supported the re-engineering process necessary for enhanced pilot functionality.
- **Standardization**: Creation of a data catalog, to include both available and utilized elements, helped track all applicable features and data provenance. For example, standardization of terms utilized under 'Job Title' supported the transition to statistical-based input features.
- **Designing a New ETL**: Consolidation of data processing into a single Python-based platform to perform all ETL actions and generate required output files allowed for easier transition to AWS-based platform for pilot.

A.2.3 HCD, UX, Interface Development

A.2.3.1 Goals

- Understand how potential users of the product (Hiring Specialists and Hiring Managers) go through the hiring process, what tools they use, tasks they need to complete and what their pain points are.
- Understand to what extent the data collected and product interface meets users' needs.
- Develop a user-centered working prototype that integrates data, models, and user interface.

A.2.3.2 What We Did

- **Context Building**: Conducted initial review of material and interviewed a PoC stakeholder to define design and business questions and corresponding interface features.
- **Wireframe Iterations and Testing**: Conducted user interviews/testing and XAI/RAI research, then used the feedback and insights to iterate on interface wireframes.

- **Product Interface Development**: Followed a structured control system for code tracking and collaboration between design experts and interface developers to combine user feedback, interface design, backend integration.
- **Internal Prototype Testing**: Conducted user testing on the developed prototype using members of our internal team to identify any remaining general, functional, or stylistic pieces of feedback, then prioritized or backlogged items for the remainder of the pilot.

A.2.3.3 Outcomes and What We Learned

- **Collaboration is Key**: Scheduled time for collaboration between developers and designers is key in ensuring that all team members understand any constraints and challenges that other team members may face in their work stream and how that may impact the product.
- **Robust Datasets to Enhance the User Experience**: Asking users to input their own data into the application to make up for a lack of data and resultant low-accuracy predictions can decrease user confidence in the tool's usefulness, as well as the data's accuracy.
- **Provide Logical Context when User Testing**: When conducting user testing, the logistics of the wireframe and testing should be clearly scripted out in the research plan and shared with the user, including any limits within the prototype or wireframe.

A.2.4 Model Development

A.2.4.1 Goals

- Analyze the original PoC and the initial model development.
- Provide baseline analysis of current model evaluations for future comparison as the models become further developed and altered.
- Utilize ML best practices and tools to develop thorough models and choose the one with the highest yielding results, that also provide value to users.

A.2.4.2 What We Did

1) **Analyze and Evaluate Initial ML Models**: Created reusable processes and templates in our approach to analyze and evaluate the initial ML models. These include a model selection process, Notebook template and Model cards.
2) **Model Iterations**: Used initial analysis and evaluations to iterate and improve on the models, including efforts toward feature selection and hyperparameter tuning.

3) **Model Selection and Deployment**: Made a final selection of models and processes to use based on evaluation results, client recommendations, and pilot time constraints.

A.2.4.3 Outcomes and What We Learned

- **Model Selection**: A Linear Regression model was deployed for predicting time to hire. This model performed well against the other models tested (on metrics R^2, adj. R^2, MSE, RMSE) and accommodated the client's preference for a simpler and more interpretable model. XGBoost, our chosen classification model, was backlogged and excluded from our prototype.
- **Feature Selection**: The time to hire model is trained using the five Hiring Phases. The application uses statistical averages filtered by position title and grade, rather than ML, for model inputs due to time constraints of the pilot. While a sufficient workaround for the pilot, this highlights the importance of feature selection and its impact on model performance.
- **Creating ML Resources for Development and Comprehension**: The model team created reusable templates and structured workflows for selecting, training, and iterative logging and evaluation of model performance to streamline future model development processes.

A.2.5 Cloud Development

A.2.5.1 Goals

1) Provide a secure deployment model for the OHC AI Pilot Hiring Assessment Tool.
2) Deploy Version 1 of the OHC AI Pilot Hiring Assessment Tool on AWS.

A.2.5.2 What We Did

- **Accessing the CMS AWS Environment**: Collaborated with CMS Support to obtain access to the AWS sandbox environment, then set up Identity and Access Management (IAM) Roles (demo provided) within the environment..
- **Designing the Model Architecture**: Surveyed AWS services approved for use in the CMS cloud system and modeled the architecture of our pilot application.
- **Deployment to AWS**: Transferred our application to the cloud by deploying the data ETL process, web app server hosting, and data transmission process. The deployed application successfully populates with model results from user input and access requires the user to be signed into the CMS Cloud VPN.

- **ATO & Security**: Researched the process for obtaining Authority to Operate (ATO) in the CMS Cloud, with emphasis on security considerations laid out within the Security Impact Analysis (SIA) and the intent to leverage an existing ATO boundary. See: provided SIA and ATO Process Overview

A.2.5.3 Outcomes and What We Learned

- **Cloud and Security Capabilities**: The team was not as experienced in CMS Cloud expertise resulting in a slow start to the cloud deployment phase. As a result, several tasks had to be backlogged for future recommendation. Nonetheless, we gained significant capabilities and outlined a security plan that strengthened our competence in the cloud field.
- **Putting the Pieces Together**: Deployment to the cloud required full collaboration between CloudOps and the rest of the team to connect the previously generated and researched content together as one working system. We were successful by engaging in high responsiveness amongst the team and conducting focused working sessions as needed.
- **Technical Grievances**: There were several challenges and technical grievances that we faced while working in the CMS AWS environment which required additional efforts to work around. We endorse prompt communication with CMS Cloud Support regarding any roadblocks, especially where there are significant time restraints such as during a short pilot.

A.2.6 Future Considerations

This pilot produced a working prototype that is both technically functional and meets users' needs. In order to continuously improve the product and prepare for it to be used at scale by CMS employees, we provide an extensive list of future enhancements for consideration.

As with any production, all steps making up the solution to this pilot have room for improvement and expansion for the next iteration of the deployed application. Key suggestions that will support the serviceability and scaling of the application include improving the models and datasets, building out our proposed interface and cloud architecture, securing the application in line with an ATO, and reviewing our HCD findings regarding the user need for a real-time tracking tool.

A.2.7 Resources

Unabridged documentation of the OHC AI Pilot and Time-To-Hire Calculator, including implementation details, artifacts, and future considerations, can be found on the CMS AI Explorer Program Confluence site under Awarded Projects and the CMS Github repository.