AI ANALYTICS & AUTOMATION

Episode 2: Training a Pre-trained Model

Nick White

June 2024

Objectives of this Session

As a business user you can describe what you need and write good prompts.

As a data professional you understand how using pre-trained models differs from traditional data science.

As an engineer, you understand how cloud services, integrations and good User Interfaces are critical.

Key Concepts

Pre-Trained Models: Models trained on extensive datasets for general tasks (e.g., GPT-4, Gemini, Llama, etc.)

User Prompting: The actions taken by end users when they interact with an AI model within an application. This involves users entering queries, commands, or inputs that the AI system processes to generate responses

Prompt Tuning: Crafting and optimizing the initial set of prompts that guide the interactions between users and an AI model within an application

RAG (Retrieval-Augmented Generation) Data Preparation: The process of managing, integrating, and optimizing external data sources to enhance the responses generated by an AI model

Fine-Tuning: Adjusting model parameters with specific datasets to enhance performance

ORIGIN®

Framework for Discussion

People

Process

Pitfalls

1/ People

Pre-Trained AI Model Best Practices

Key Roles for AI Success

Roles in AI projects do not equate to a specific position or a single individual; rather, they represent a set of capabilities and responsibilities that are required to achieve objectives. In fact, AI will continue to blur the lines of who does what.

Business Users

Data & Analytics

Design & Engineering

Business User Responsibilities

Define Use-Cases

Identify Pain Points and Opportunities

- Problem Identification: Highlight specific challenges and inefficiencies in current workflows or processes.
- Opportunity Spotting: Suggest areas where AI could improve productivity, reduce costs, or enhance customer satisfaction.

Define User Requirements

- User Stories: Help create user stories and scenarios that describe how AI solutions should interact with users.
- Feature Prioritization: Prioritize features and functionalities based on user needs and business impact.

Provide Domain Expertise

- Contribute to Data Understanding: Identify valuable data sources and help determine the quality of the data.
- Key Performance Indicators (KPIs): Identify relevant KPIs to measure the success of AI implementations.

Create Effective Prompts

Provide Detail for Effective Prompt-Tuning

- User Needs: Clearly understand the specific needs and objectives of the users. Know what users are trying to achieve with their prompts.
- Contextual Relevance: Ensure prompts are contextually relevant to the situation or task at hand.

Craft Clear and Specific Prompts

- Clarity and Precision: Write prompts that are clear, concise, and free from ambiguity. Avoid vague or overly complex language.
- Specific Instructions: Provide specific details or constraints within the prompt to guide the AI model towards the desired outcome.

Iterate and Refine Based on Feedback

- Testing and Feedback: Regularly test the prompts with real users and gather feedback on their effectiveness.
- Continuous Improvement: Use the feedback to refine and improve the prompts, ensuring they become more effective over time and better meet user needs.

The Good, the Bad with the Prompting

Bad Prompt

"Explain why a customer's order is late."

Result

"Your order is late due to unforeseen circumstances."

Why It's Bad

- Vagueness: The prompt is too vague and does not provide specific details about the order, making it difficult for the LLM to generate a detailed and helpful response.
- Lack of Context: The prompt does not specify which order it refers to or the relevant details, resulting in a generic and uninformative response.
- Missing Information: The response lacks actionable information and does not provide the customer with a clear update or expected resolution.

Good Prompt

"Respond to a customer asking why their order #12345, placed on May 1, 2023, has not been delivered yet. Provide an update on the current status and expected delivery date."

Result

"Dear [Customer Name],

Thank you for reaching out regarding your order #12345 placed on May 1, 2023. We apologize for the delay in delivery. Currently, your order is being processed and is expected to be shipped within the next 48 hours. You can expect delivery by May 10, 2023. We appreciate your patience and understanding.

Best regards,

[Your Company Name]"

Why It's Good

- Clarity and Specificity: The prompt clearly asks for a response to a specific order inquiry, including the order number and date.
- Contextual Relevance: The prompt provides the necessary details (order number, date, and status) to generate a relevant and informative response.
- Actionable Information: The response includes specific updates and a clear expected delivery date.

Data & Analytics Responsibilities

Data Preparation (for RAG & Tuning)

Identify and Collect Relevant Data Sources

- Diverse Data Collection: Gather data from various sources, such as internal databases, external APIs, web scraping, and publicly available datasets.
- Relevance Assessment: Ensure that the collected data is pertinent to the specific AI application and use case, focusing on quality and context.

Clean and Preprocess the Data

- Data Cleaning: Remove duplicates, outliers, and inconsistencies to ensure data quality. This step may involve filtering out irrelevant data points and correcting errors.
- Preprocess Data: by tokenizing text, normalizing numerical values, and handling missing data.

Structure the Data

- Data Organization: Put data into a format that is compatible with the use-case.
- Integrate and Annotate Data (RAG): Link queries with relevant context passages and accurately label data points to aid the retrieval process.
- Training and Validation Setup (Fine-Tuning): Split the dataset into training and validation subsets and organize into batches for model training.

Tuning (Prompt & Fine)

Prompt Tuning and Optimizing RAG Data

- Clear and Specific Prompts: Ensure prompts are clear, concise, and specific to guide the pre-trained model effectively.
- Contextual Data Integration: Optimize RAG (Retrieval-Augmented Generation) data by integrating relevant and high-quality context information to enhance the accuracy of the model's responses.

Iterative Testing and Refinement

- Continuous Feedback Loop: Regularly test the prompts and RAG data with real users and gather feedback to identify areas for improvement.
- Data-Driven Refinement: Refine the prompts and RAG data based on user feedback and performance data to enhance the effectiveness of the pre-trained model.

Fine-Tuning (When Needed)

- Evaluate Necessity: Consider fine-tuning only after exhausting prompt tuning and RAG data optimization. Assess whether another model is needed.
- Targeted Fine-Tuning: Fine-tune the pre-trained model using a dataset of high-quality input-output pairs, ensuring to split the data into training and validation sets to address specific gaps in performance.

Design & Engineering Responsibilities

Build AI Experiences

User-Centric Prompt Design

- Define User Personas and Needs: Create detailed profiles of the end-users to understand their needs, behaviors, and pain points, focusing on how they will interact with AI prompts.
- Prompt Usability Testing: Conduct testing sessions to ensure that prompts are clear, intuitive, and easy for users to understand and use effectively.

Data-Driven Interactions

- Integrate Data Seamlessly: Ensure that data is integrated into the user interface in a way that enhances the AI experience, providing relevant context and information to users.
- Dynamic Data Display: Design interfaces that dynamically present data based on user interactions and AI responses, making data a central part of the user experience.

Intuitive Interface Design

- Simplified Prompt Entry: Design the interface to make it easy for users to write effective prompts, with features like auto-suggestions, examples, and clear instructions.
- Accessible Data Output: Ensure that the results provided by the AI are easy to interpret and access, using visual aids like charts, graphs, and summaries to present data in a user-friendly manner.

Configure & Integrate Services

Service Selection and Configuration

- Choose Appropriate AI Service: Select the most suitable pre-trained model services based on the use case requirements.
- Service Configuration: Configure the chosen AI services to align with the specific needs of the application, such as setting parameters and integration points.

Seamless Integration

- API Integration: Integrate AI services into the application using APIs, ensuring smooth communication between different components.
- Data Pipeline Setup: Establish data pipelines to feed the AI services with relevant data for processing and analysis.

Performance Monitoring and Optimization

- Continuous Monitoring: Implement monitoring tools to track the performance of AI services in real-time.
- Optimization and Scaling: Regularly optimize the AI services for performance improvements and scale them to handle increased loads as the application grows.

ORIGIN®

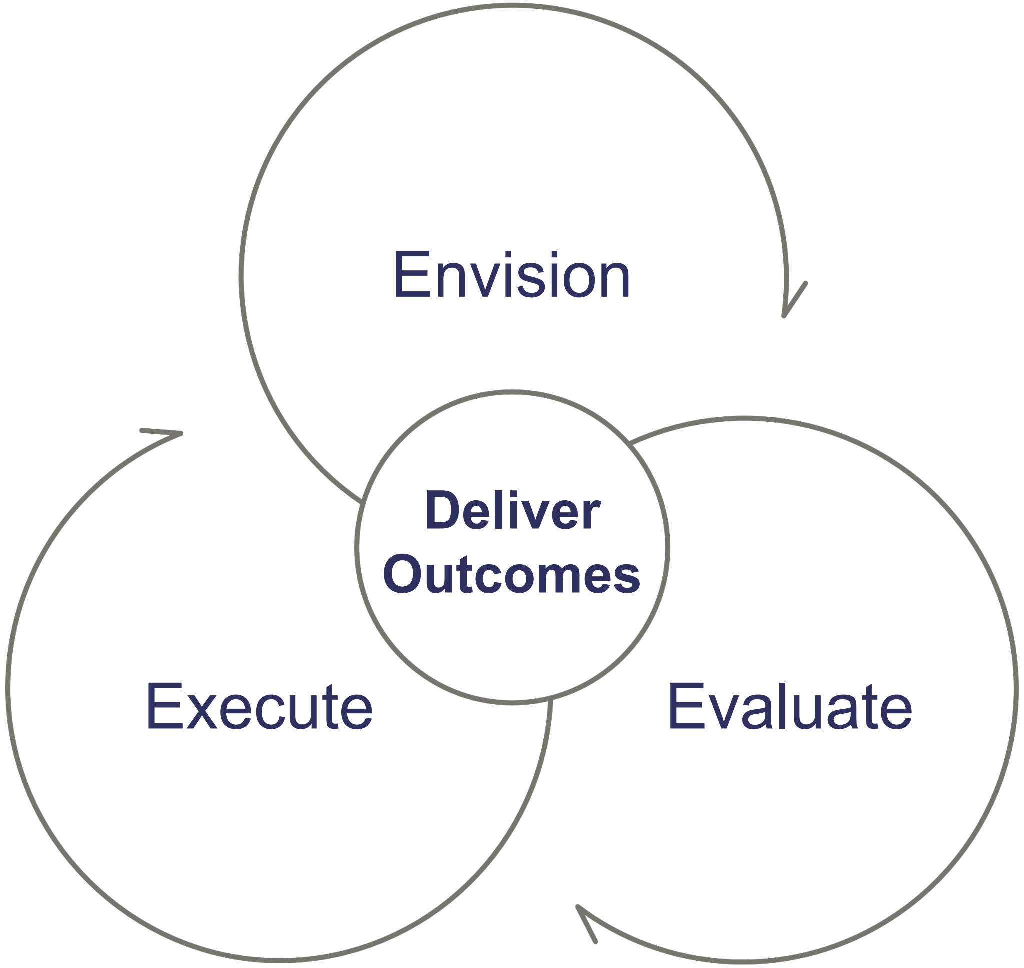
2/ Process

Optimizing for Pre-Trained AI Models

The AI Virtuous Cycle

Achieving success with AI requires adherence to a virtuous cycle encompassing three critical stages: envision, evaluate, and execute.

Organizations can maximize the value and impact of AI while mitigating potential challenges using a pragmatic, iterative and outcome –focused approach.



Envision

Maturity Assessment

- Conduct interviews & surveys with key stakeholders to understand their challenges and expectations.
- Review any documentation to understand organization, process & technical current state of marketing.
- Review systems, data and future initiatives to understand the current state of marketing.
- Analyze marketing KPIs and OKRs to identify areas for improvement.
- Define strategic initiatives and tactical steps to achieve each level of maturity.

Strategic Roadmap

- Identify and prioritize use-cases against strategic initiatives based on urgency, impact, feasibility, and resource requirements.
- Define the requirements for each initiative and use-case, including timelines, resources, and dependencies.
- Define success metrics and monitoring mechanisms to track progress.
- Create a dual-track (maturity & value delivery) roadmap, beginning with MVP use-cases to inform technical recommendations and POCs.

Target Architecture

- Cloud platform
- Content management
- Data ingestion, integration and governance platforms
- Data visualization & insight delivery platforms
- Data science & analytics platforms
- Automation platforms
- Processes, organization and workflows
- Collaboration platforms

Evaluate

Product Definition

- Define a clear product vision that aligns with the selected use case and business objectives.
- Communicate the vision effectively to all stakeholders.
- List the key features needed to support the use case.
- Prioritize features based on user needs and business impact.
- Develop detailed user personas representing the target audience.
- Validate the personas with real user data and feedback.
- Identify the minimum viable product (MVP) that includes essential features to address the use case.
- Ensure the MVP is scoped to deliver maximum value with minimal effort.
- Investigate the quality and availability of data required for the use case.
- Ensure data is clean, relevant, and accessible for AI model training and deployment.
- Evaluate available AI models to determine the best fit for the use case.
- Choose the AI model that aligns with data characteristics and use case requirements.

Proof of Concept (POC)

- Define success criteria and align stakeholders on objectives.
- Choose a suitable test environment and deploy pre-trained AI models.
- Train users on AI model interaction and gather initial feedback.
- Collect usage data and implement continuous feedback loops.
- Refine prompts, data staging, and UI based on feedback.
- Measure against success criteria and identify improvement areas.
- Assess scalability and develop a detailed deployment plan.
- Compile findings and recommendations for deployment.

Execute

Prepare the Data

- Collect and organize relevant data.
- Clean and preprocess the data to ensure it's in a format compatible with the selected AI model.
- Split the data into training and testing sets to evaluate the model's performance.

Design the User Experience

- Create a user-centric design for the AI-powered application.
- Develop intuitive and user-friendly interfaces that allow seamless interaction with the AI model.

Train and Deploy the Model

- Train the AI model on the prepared data using the selected platform.
- Monitor the training process and make adjustments as needed.
- Deploy the trained model to a production environment for real-world use.

Integrate with Existing Systems

- Integrate the deployed AI model with existing systems and processes to enable seamless data flow and decision-making.
- Ensure secure and reliable communication between the AI model and other components.

Monitor & Evaluate

- Continuously monitor the performance of the AI model in production.
- Track key metrics like accuracy, latency, and user satisfaction.
- Identify areas for improvement and make necessary adjustments.

Improve & Scale

- Regularly update and improve the AI model based on new data and emerging insights.
- Scale the AI deployment as needed to handle increasing user demand or new use cases.

3/ Pitfalls

Common Traps to Avoid with Pre-Trained AI Models

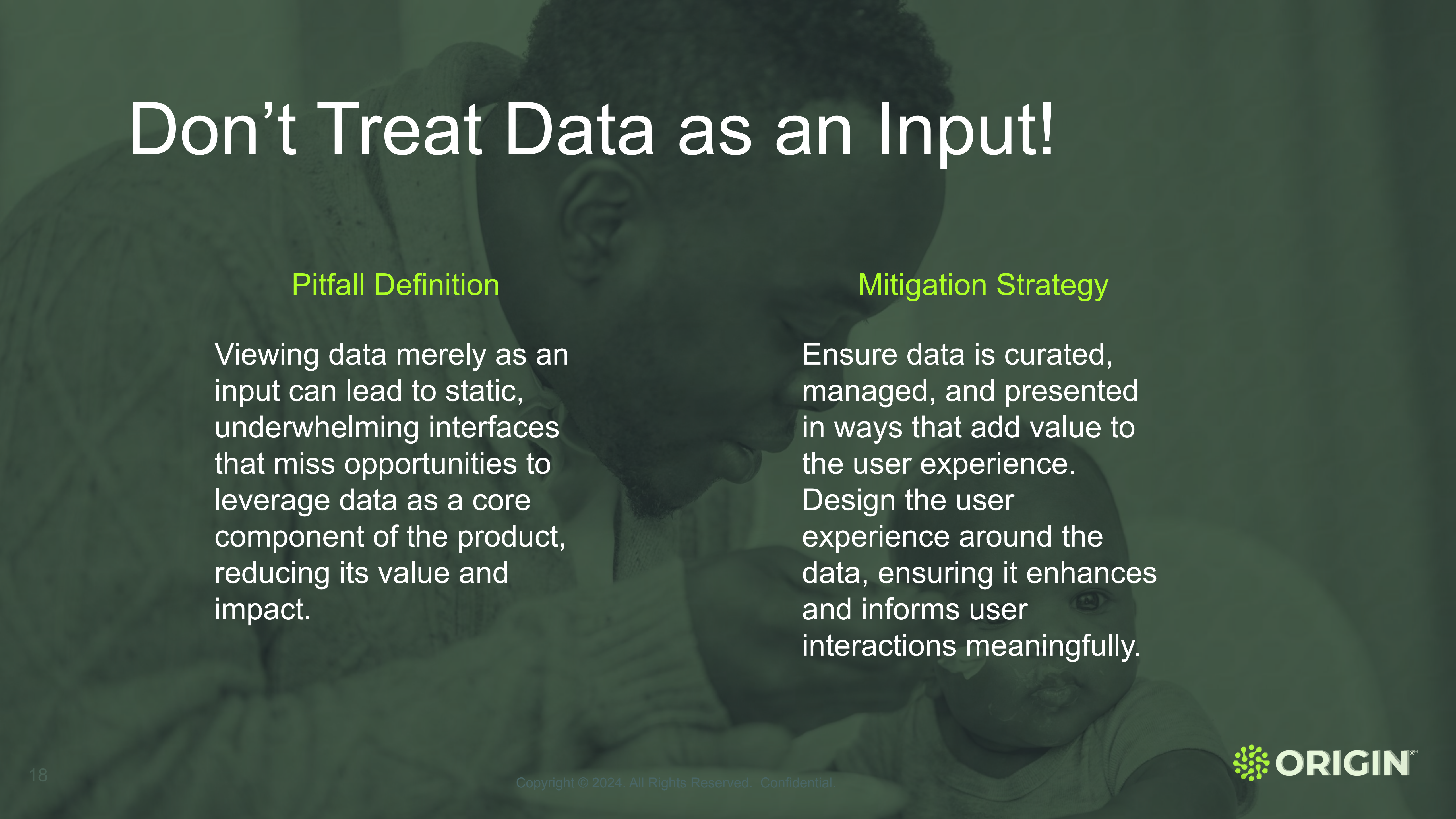
Don’t Overengineer!

Pitfall Definition

Adding excessive features, layers, or customizations that do not significantly improve performance but increase the difficulty of maintenance, deployment, and scalability.

Mitigation Strategy

Utilize pre-trained models as a base and adopt an iterative development approach to only make minimal, necessary adjustments (prompt-turning or fine-tuning) to suit the specific use case.



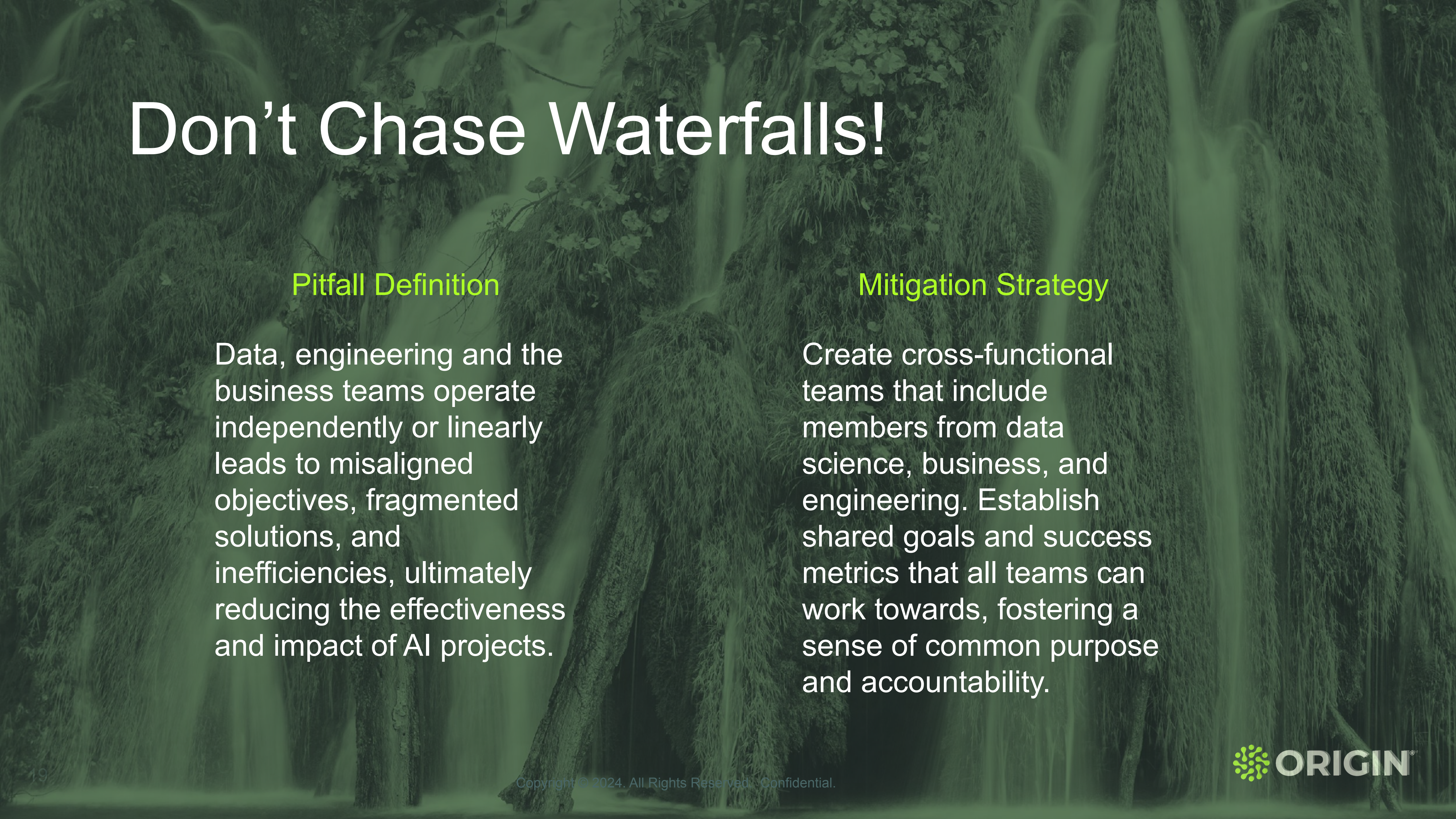
Don’t Treat Data as an Input!

Pitfall Definition

Viewing data merely as an input can lead to static, underwhelming interfaces that miss opportunities to leverage data as a core component of the product, reducing its value and impact.

Mitigation Strategy

Ensure data is curated, managed, and presented in ways that add value to the user experience. Design the user experience around the data, ensuring it enhances and informs user interactions meaningfully.



Don’t Chase Waterfalls!

Pitfall Definition

Data, engineering and the business teams operate independently or linearly leads to misaligned objectives, fragmented solutions, and inefficiencies, ultimately reducing the effectiveness and impact of AI projects.

Mitigation Strategy

Create cross-functional teams that include members from data science, business, and engineering. Establish shared goals and success metrics that all teams can work towards, fostering a sense of common purpose and accountability.

ORIGIN®

Simple takeaways

Using pre-trained models is a combination of current approaches.

Prompting excellence by business users is critical to success.

Garbage in, garbage out still applies with pre-trained models.

Thank you!

Nick White
Head of Decision Science & Experience
nwhite@origindigital.com
origindigital.com