TRAINING, FINE-TUNING, AND EVALUATING FOUNDATION MODELS

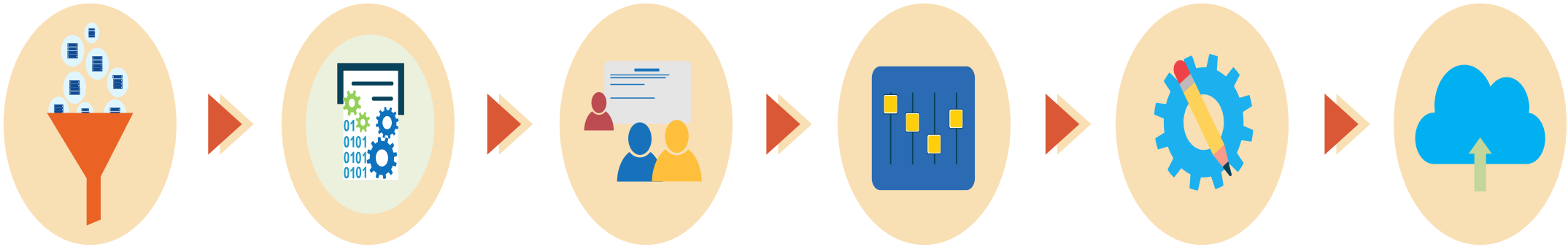
Objectives

- Describe key elements of training a foundation model (FM)
- Prepare data to fine-tune a foundation model
- Evaluate foundation model performance metrics to assess performance
- Determine if a foundation model meets business goals

KEY ELEMENTS OF TRAINING A FOUNDATION MODEL (FM)

- The foundation model (FM) **training** process ensures that the foundation model is well-prepared and optimized for specific use cases, improving its performance and creating a better user experience
- This important process involves several steps

TRAINING A FOUNDATION MODEL



Data collection

Data preprocessing

Continued pre-training

Fine-tuning

Model customization

Deployment

INSTRUCTION TUNING
FINE-TUNING

- Instruction tuning is a method of fine-tuning a foundation model using labeled examples formatted as prompt-response pairs
- This process is designed to improve the model's performance on specific tasks by providing clear instructions and expected responses

INSTRUCTION TUNING FINE-TUNING

Prepare instructions

Format data

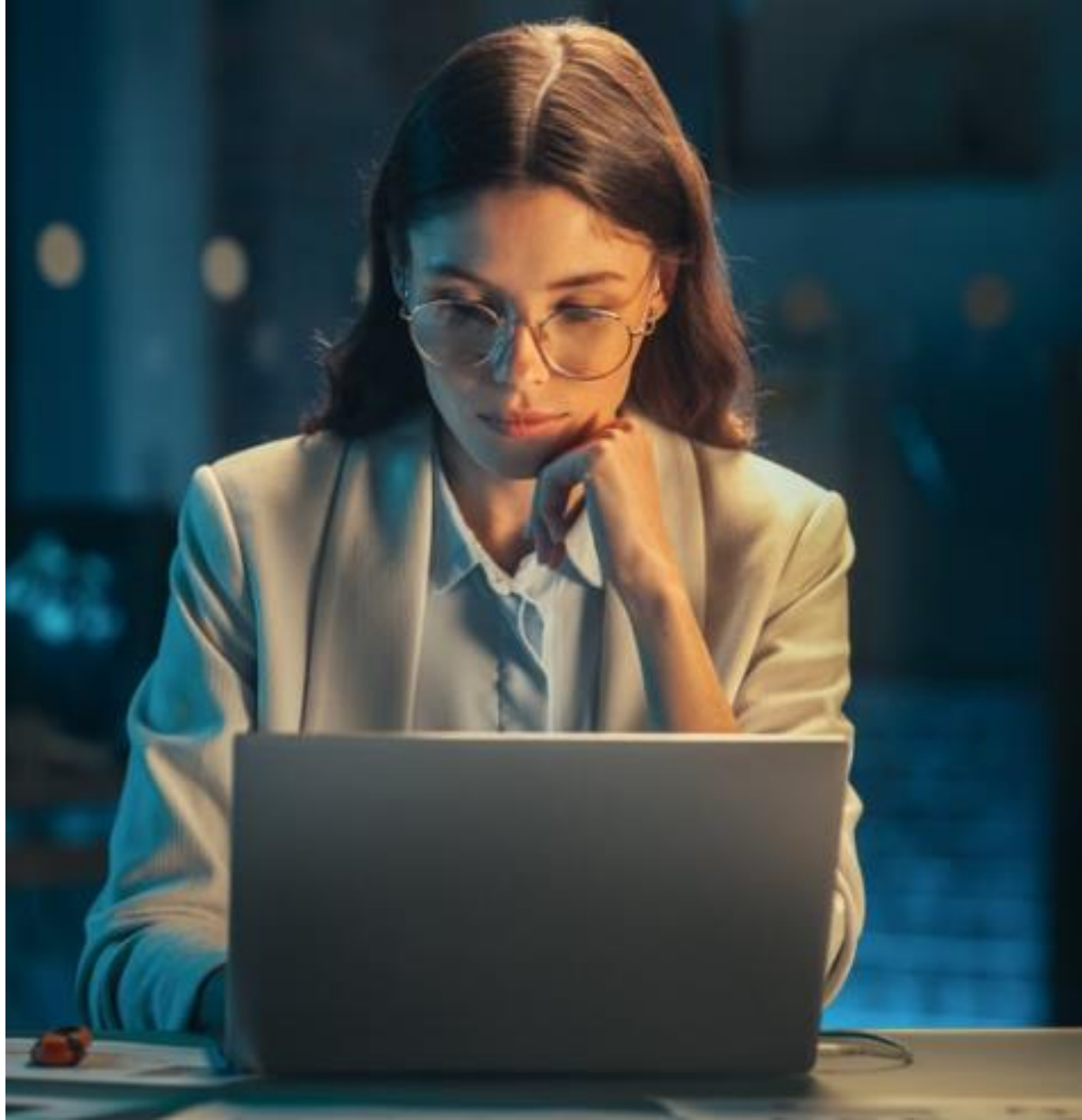
Create training job

Fine-tuning process

Evaluate and adjust

DOMAIN ADAPTATION

- Fine-tuning a foundation model for specific domains uses a process called domain adaptation
- This method enables the leveraging of pre-trained foundation models and their adaptation to specific tasks with limited domain-specific data
  - Helps make models more agreeable to solving general downstream NLP tasks
  - Can be further adapted to more specific tasks with additional data and fine-tuning



DOMAIN ADAPTATION

- For example, imagine the goal is to fine-tune a language model for financial text generation
- The process involves collecting a dataset of financial documents, such as SEC filings, financial reports, and news articles related to finance

TRANSFER LEARNING

- Fine-tuning a foundation model using transfer learning involves leveraging a pre-trained model and adapting it to a specific task using domain-specific data
- Imagine the goal is to fine-tune a language model for legal text analysis

Prepare training data

Upload data to Amazon S3

Create a fine-tuning job

Fine-tuning process

Evaluate and adjust

TRANSFER LEARNING

CONTINUOUS PRE-TRAINING

- Continuous pre-training, according to AWS, involves further training a foundation model using unlabeled data to improve its domain knowledge and adaptability
- Imagine the goal is to continuously pre-train a language model for medical text analysis

CONTINUOUS PRE-TRAINING

Prepare unlabeled data

Upload data to Amazon S3

Create continuous pre-training job

Pre-training process

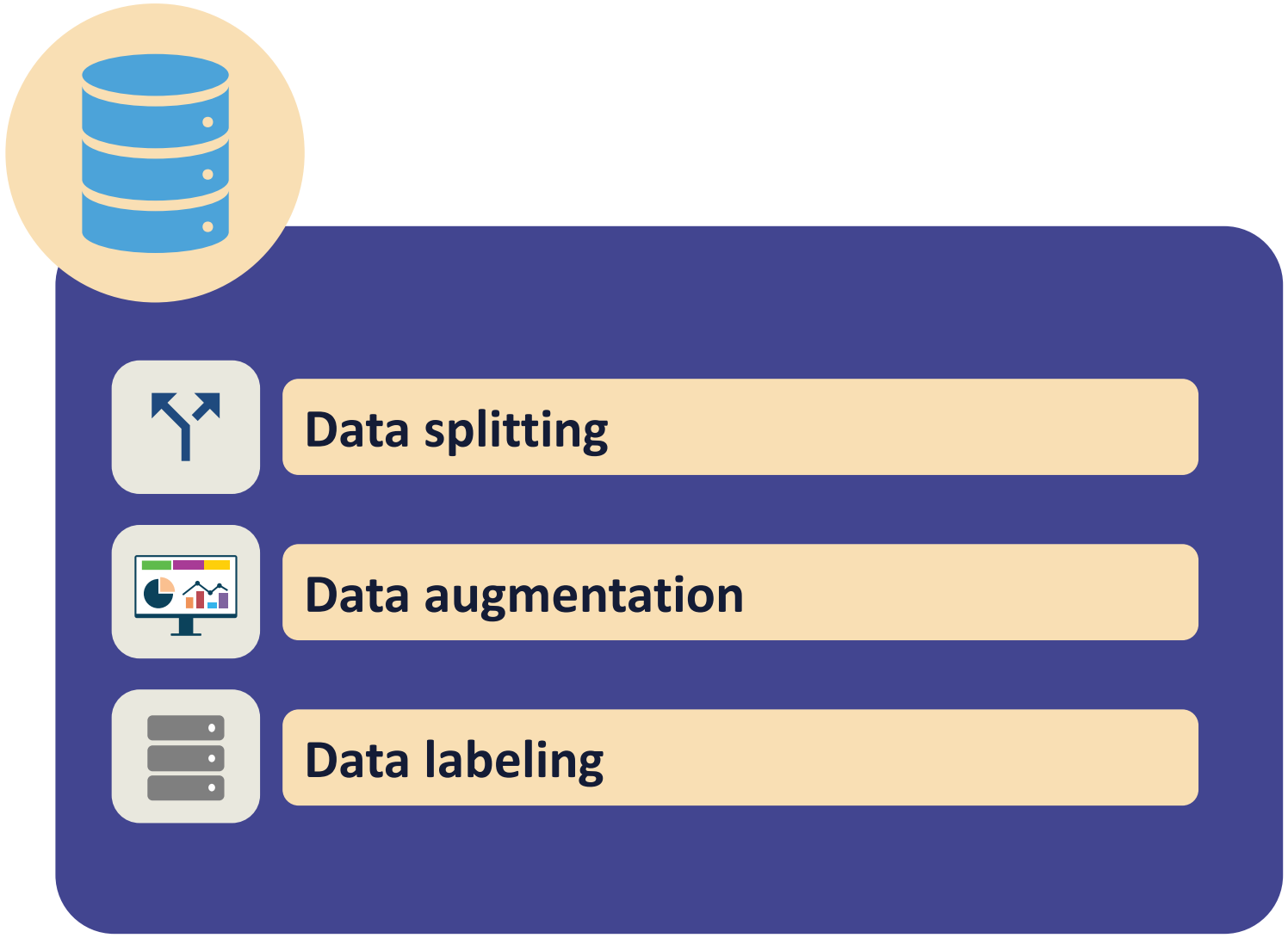
Evaluate and adjust

PREPARING DATA TO FINE-TUNE A FOUNDATION MODEL WITH DATA CURATION

- Fine-tuning a foundation model involves preparing the data meticulously to ensure the model adapts well to specific needs and results
- By taking the following steps, data can be well-prepared for fine-tuning a foundation model, resulting in better performance and more accurate results

- Data collection
- Data cleaning
- Data labeling
- Data formatting

DATA CURATION



DATA CURATION

PREPARING TUNING DATA WITH GOVERNANCE

- When preparing data to fine-tune a foundation model, governance plays a crucial role in ensuring data quality, privacy, and security
- Incorporating the following governance practices ensures that data is well-managed and secure, leading to more reliable and accurate fine-tuning of foundation models

PREPARING TUNING DATA WITH GOVERNANCE

Data quality

Data privacy

Data security

PREPARING TUNING DATA WITH GOVERNANCE

Data access control

Data lineage

Data governance framework

FINE-TUNING BASED ON SIZE

- Ensure that the dataset is large enough to provide the model with sufficient examples to learn from
- A larger dataset can help the model generalize better and improve its performance on the specific task
- If the dataset is too large to process efficiently, consider using data sampling techniques to create a representative subset

FINE-TUNING BASED ON SIZE

- For smaller datasets, data augmentation techniques can be used to artificially increase the size of the dataset
  - Involves creating variations of the existing data, such as adding noise, rotating images, or paraphrasing text

Fine-Tuning Based on Labeling

- Clearly define the labels needed for the specific task
  - This could include categories, tags, or annotations that are relevant to the model's purpose
- Create detailed guidelines for the labeling process to ensure consistency and accuracy
- Use automated labeling tools to pre-label the data when feasible

REPRESENTATIVENESS

- When preparing data to fine-tune a foundation model, ensuring representativeness is key to achieving accurate and reliable results
- Gather data from a wide range of sources to ensure that the dataset represents the full spectrum of scenarios the model will encounter
  - This helps the model generalize better to different contexts and reduces bias

REPRESENTATIVENESS

- Ensure that the dataset is balanced in terms of different classes or categories
  - For example, when training a model for sentiment analysis, it is important to have an equal number of positive, negative, and neutral examples

REPRESENTATIVENESS

- Use sampling techniques to create a representative subset of the data if the full dataset is too large to process
  - This helps maintain the diversity and balance of the dataset while reducing computational costs

FINE-TUNING WITH REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF)

- When preparing data to fine-tune a foundation model, Reinforcement Learning from Human Feedback (RLHF) is a valuable approach
- Incorporating RLHF ensures that the model's outputs are more aligned with human preferences, leading to better performance and more accurate results

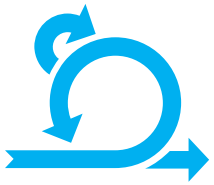
REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF)

Human feedback collection

Reward signal design

Training data generation

Policy optimization



Iterative refinement

ASSESSING FOUNDATION MODEL PERFORMANCE WITH HUMAN EVALUATION

- Evaluate the performance of foundation models using human evaluation is a critical step to ensure the model meets the desired quality and aligns with people’s preferences
- Incorporate human evaluation ensures that the model's outputs are better aligned with human inclinations, leading to better performance and more accurate results

ASSESSING FOUNDATION MODEL PERFORMANCE WITH HUMAN EVALUATION

- Gather feedback from human evaluators on the model's outputs
  - This feedback can be in the form of ratings, rankings, or detailed comments on the quality and relevance of the model's responses
- For subjective or custom metrics – such as friendliness, style, and alignment to brand voice – set up human evaluation workflows
  - This allows for a more nuanced assessment of the model's performance

ASSESSING FOUNDATION MODEL PERFORMANCE WITH HUMAN EVALUATION

- Continuously refine the model by incorporating human feedback into the training process
  - This iterative approach helps improve the model's performance over time
- Use human evaluators to assess the model's performance on specific tasks, such as text generation, text classification, question answering, and text summarization

EVALUATING FOUNDATION MODEL PERFORMANCE WITH BENCHMARK DATASETS

- Assessing the performance of foundation models using benchmark datasets is another critical step to ensure the model's effectiveness and reliability
- Choose appropriate benchmark datasets that are relevant to the specific task the model is being fine-tuned for
  - These datasets should be well-established and widely recognized in the field

EVALUATING FOUNDATION MODEL PERFORMANCE WITH BENCHMARK DATASETS

- Establish a baseline performance by evaluating the model on the selected benchmark datasets before fine-tuning
  - This helps in understanding the initial capabilities of the model
- Fine-tune the model using the curated dataset and then evaluate its performance on the benchmark datasets
  - This step helps in assessing the improvements made through fine-tuning

EVALUATING FOUNDATION MODEL PERFORMANCE WITH BENCHMARK DATASETS

- Compare the model's performance with state-of-the-art models on the same benchmark datasets
  - This provides a clear indication of where the model stands in terms of performance
- Continuously refine the model based on the evaluation results and re-evaluate it on the benchmark datasets
  - This iterative process helps in achieving optimal performance

Real-World Benchmark Datasets at AWS

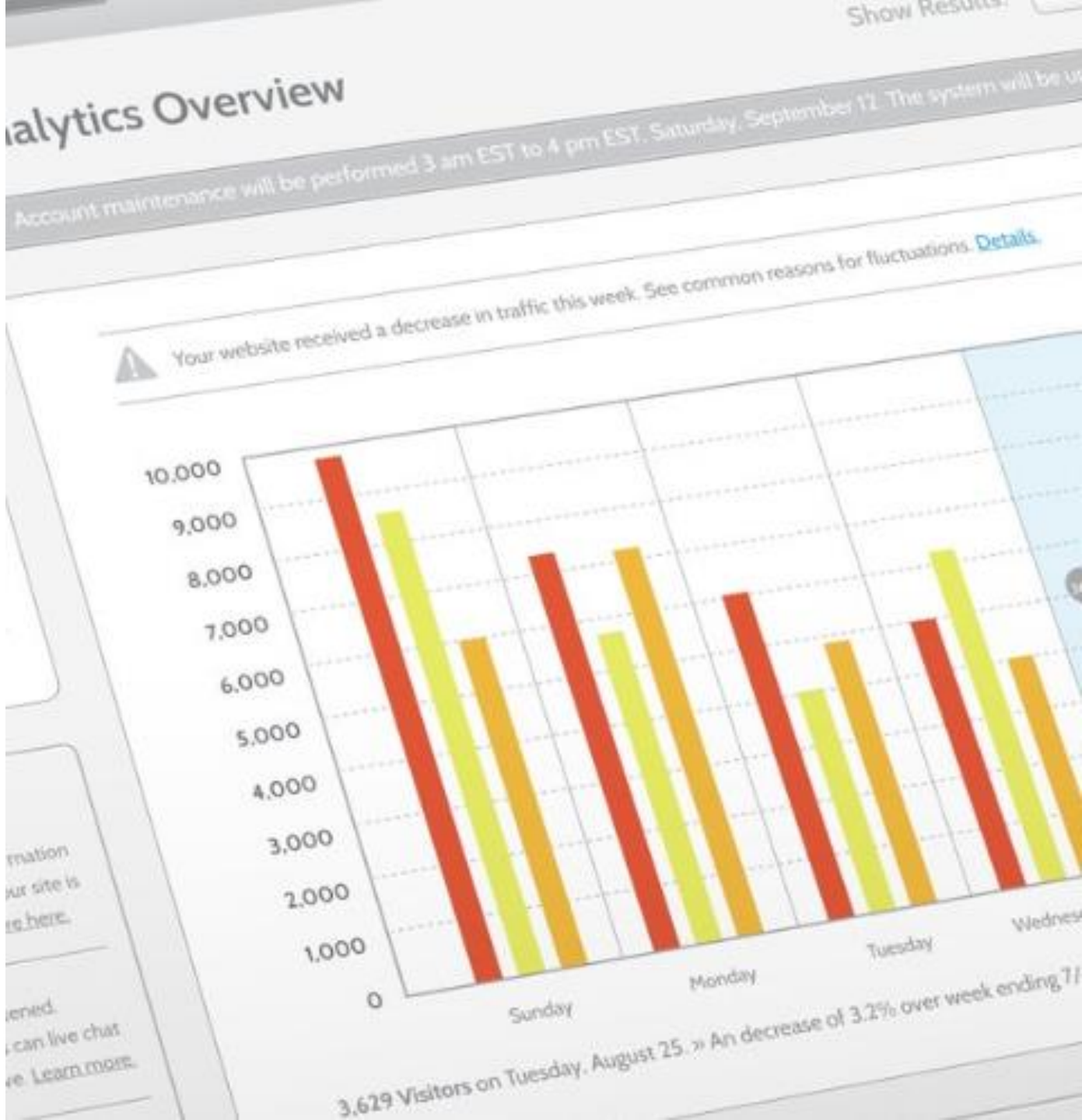
GLUE (General Language Understanding Evaluation)	SQuAD (Stanford Question Answering Dataset)	ImageNet	MS MARCO (Microsoft MAchine Reading COmprehension):
A collection of diverse natural language understanding tasks designed to evaluate models on a wide range of linguistic phenomena	A reading comprehension dataset consisting of questions posed on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding article	A large-scale dataset of labeled images used for image classification and object detection tasks, widely used to benchmark the performance of computer vision models	A dataset for evaluating models on machine reading comprehension tasks, including question answering and passage ranking

THE ROUGE METRIC FOR ASSESSING FM PERFORMANCE

- The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric is a set of metrics used to evaluate the quality of text summaries by comparing them to reference summaries
- ROUGE is commonly used to assess the performance of a foundation model (FM), especially in tasks like text summarization

THE ROUGE METRIC FOR ASSESSING FM PERFORMANCE

- **ROUGE-N**: measures the overlap of n-grams (contiguous sequences of *n* items) between the generated summary and the reference summary
  - For example, ROUGE-1 measures the overlap of unigrams (single words), while ROUGE-2 measures the overlap of bigrams (two-word sequences)



THE ROUGE METRIC FOR ASSESSING FM PERFORMANCE

- **ROUGE-L**: measures the longest common subsequence (LCS) between the generated summary and the reference summary
  - This metric captures the longest sequence of words that appear in both summaries in the same order

THE ROUGE METRIC FOR ASSESSING FM PERFORMANCE

- **ROUGE-W**: a weighted version of ROUGE-L that gives more importance to longer subsequences
- **ROUGE-S**: measures the overlap of skip-bigrams, which are pairs of words that appear in the same order in both summaries but may have other words in between them

THE BLEU METRIC FOR ASSESSING FM PERFORMANCE

- The BLEU (Bilingual Evaluation Understudy) metric is a widely used method for evaluating the quality of text generated by a model, particularly in machine translation tasks
- BLEU measures the similarity between the generated text and one or more reference texts

THE BLEU METRIC FOR ASSESSING FM PERFORMANCE

- **N-gram precision**: BLEU calculates the precision of n-grams (contiguous sequences of *n* items) in the generated text compared to the reference text
  - For example, BLEU-1 measures the overlap of unigrams (single words), while BLEU-2 measures the overlap of bigrams (two-word sequences)

THE BLEU METRIC FOR ASSESSING FM PERFORMANCE

- **Brevity penalty**: to avoid favoring shorter generated texts, BLEU includes a brevity penalty
  - This penalty reduces the score if the generated text is significantly shorter than the reference text
- **Weighted average**: BLEU computes a weighted average of the n-gram precisions, typically up to four grams (BLEU-4)
  - This provides a balanced measure of the model's performance across different levels of granularity

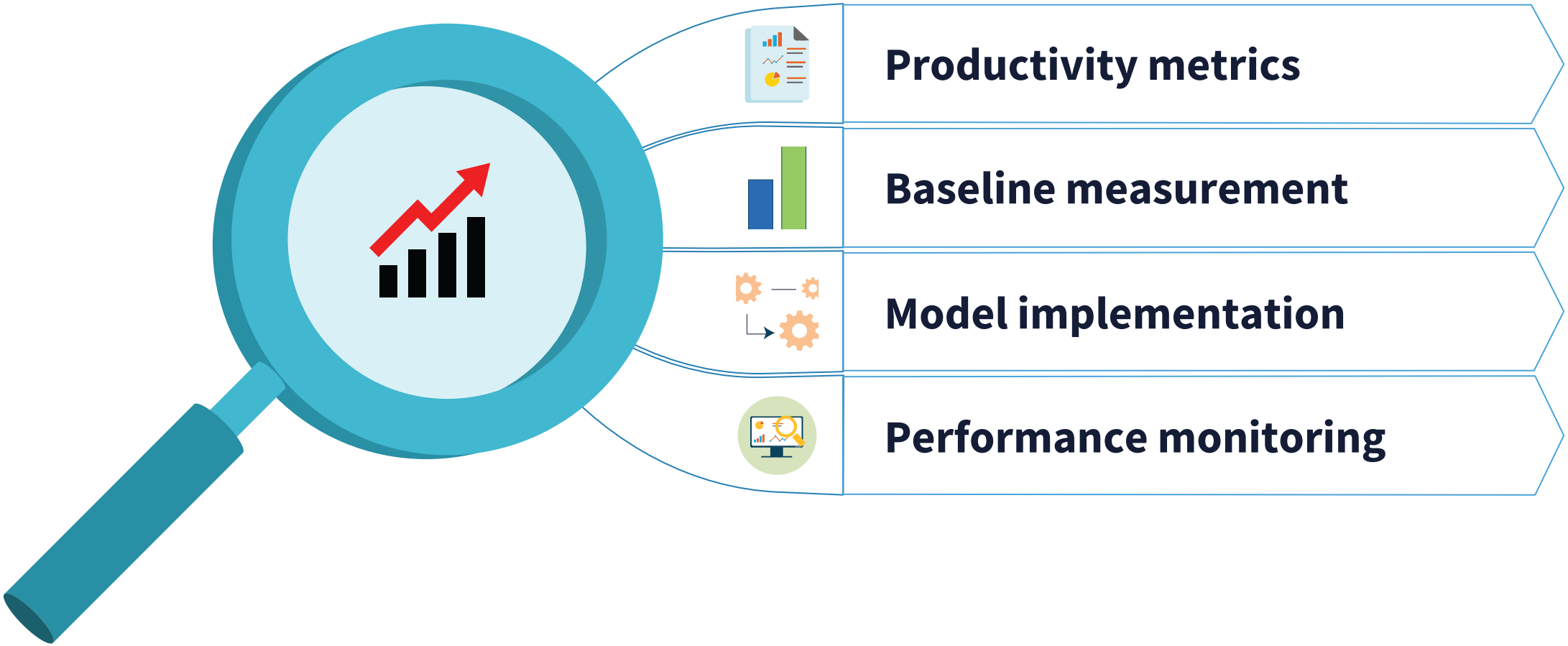
THE BLEU METRIC FOR ASSESSING FM PERFORMANCE

- **Score range**: BLEU scores range from 0 to 1, with higher scores indicating better performance
- A score of 1 means the generated text is identical to the reference text

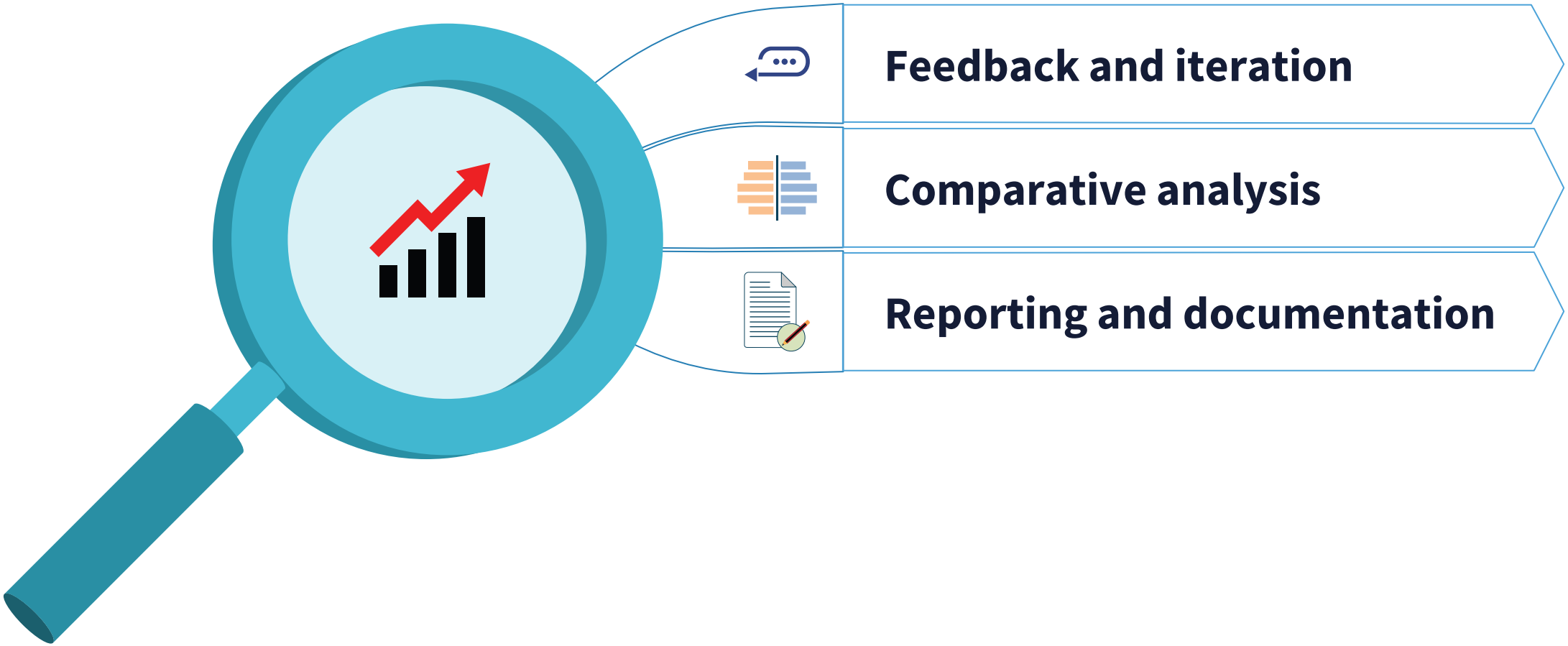
DETERMINING IF AN FM MEETS BUSINESS GOALS BASED ON PRODUCTIVITY

- Businesses can effectively evaluate whether a foundation model (FM) meets their **productivity** objectives and make data-driven decisions to optimize its performance
- According to AWS, determining whether a foundation model effectively meets business objectives based on productivity involves several steps

DETERMINING IF AN FM MEETS BUSINESS GOALS BASED ON PRODUCTIVITY



DETERMINING IF AN FM MEETS BUSINESS GOALS BASED ON PRODUCTIVITY



DETERMINING IF AN FM MEETS BUSINESS GOALS BASED ON USER ENGAGEMENT

- Engagement metrics
- Performance monitoring
- Baseline measurement
- Feedback and iteration
- Model implementation
- Reporting and documentation

GUIDELINES FOR RESPONSIBLE AI

OBJECTIVES

- Describe the features and tools of responsible AI
- Explore responsible practices and legal risks of generative AI
- Define characteristics of datasets
- Compare effects and tools for bias and variance
- Describe transparent, explainable, non-transparent, and non-explainable models and tools to identify them
- Explain the tradeoffs between model safety and transparency
- Explore the principles of human-centered design for explainable AI

BIAS FEATURE OF RESPONSIBLE AI

- The "**bias**" feature of responsible AI is a way to find and mitigate potential prejudices in AI models
- Biases are imbalances in data or disparities in the performance of a model across different groups
- AWS offers tools like SageMaker Clarify to help identify and address these biases during data preparation, after model training, and in deployed models
- This ensures that AI systems are fair and perform as intended across diverse groups

FAIRNESS FEATURE OF RESPONSIBLE AI

- The "**fairness**" feature of responsible AI is a core dimension that considers the impacts of AI systems on different groups of stakeholders
- Fairness in AI aims to ensure that AI models perform equitably across diverse groups and do not perpetuate or amplify biases
- AWS provides tools like Amazon SageMaker Clarify to help detect and mitigate potential biases during data preparation, model training, and deployment

AMAZON BEDROCK GUARDRAILS

- **Inclusivity** in AI ensures that AI systems are designed and deployed to be accessible and beneficial to all users, regardless of their background, abilities, or circumstances
- This involves engaging with a broad range of stakeholders, including underrepresented groups, to ensure that AI technologies are developed and used in ways that promote equity and inclusion

ROBUSTNESS FEATURE OF RESPONSIBLE AI

- The "**robustness**" feature of responsible AI focuses on ensuring AI systems produce correct outputs, even when faced with unexpected or adversarial inputs
- Robustness in AI aims to make models resilient and reliable, minimizing errors and vulnerabilities

SAFETY FEATURE OF RESPONSIBLE AI

- The "**safety**" feature is a core dimension that focuses on preventing harmful system outputs and misuse
- AWS aims to ensure that AI systems are designed and deployed in a way that minimizes risks and protects users from potential harm
- This involves safeguards, conducting thorough testing, and continuously monitoring AI systems to detect and address any issues

VERACITY FEATURE OF RESPONSIBLE AI

- **Veracity** is closely related to "robustness" and focuses on ensuring that AI systems produce correct outputs, even when faced with unexpected or adversarial inputs
- Veracity in AI aims to achieve accurate and reliable system outputs, minimizing errors and vulnerabilities
- SageMaker Clarify can help evaluate and improve model performance by detecting potential biases and ensuring accurate predictions

AMAZON BEDROCK GUARDRAILS

- Amazon Bedrock Guardrails are safeguards designed to ensure the safety and reliability of generative AI applications
  - They help prevent harmful content, protect user privacy, and maintain the accuracy of AI outputs
- Guardrails can be configured to filter user inputs and model responses and block or mask undesirable content such as hate speech, insults, sexual content, violence, and sensitive information

AMAZON BEDROCK GUARDRAILS

- Guardrails can be tailored to specific use cases and applied across multiple foundation models, providing a consistent user experience and standardizing safety and privacy controls
- They also support contextual grounding checks to detect and filter hallucinations in model responses, ensuring that the AI outputs are factually accurate and relevant to the user's query

RESPONSIBLE ENVIRONMENTAL CONSIDERATIONS WHEN SELECTING A FOUNDATION MODEL (FM)

Sustainability goals

AWS Region selection

Right-sizing consumption

Use of managed services

Optimizing software development

LEGAL RISKS OF INTELLECTUAL PROPERTY INFRINGEMENT CLAIMS WHEN WORKING WITH GENERATIVE AI

LEGAL RISKS ASSOCIATED WITH BIASED MODEL OUTPUTS

Discrimination claims

If a generative AI model produces biased outputs that discriminate against certain groups, it can lead to legal claims of discrimination

This is particularly relevant in areas like hiring, lending, and law enforcement

Reputational damage

Biased outputs can harm an organization's reputation, leading to loss of customer trust and potential legal action

Regulatory compliance

Many industries are subject to regulations that require fair and unbiased decision-making

Biased AI outputs can lead to non-compliance with these regulations, resulting in fines and other legal consequences

Liability for harm

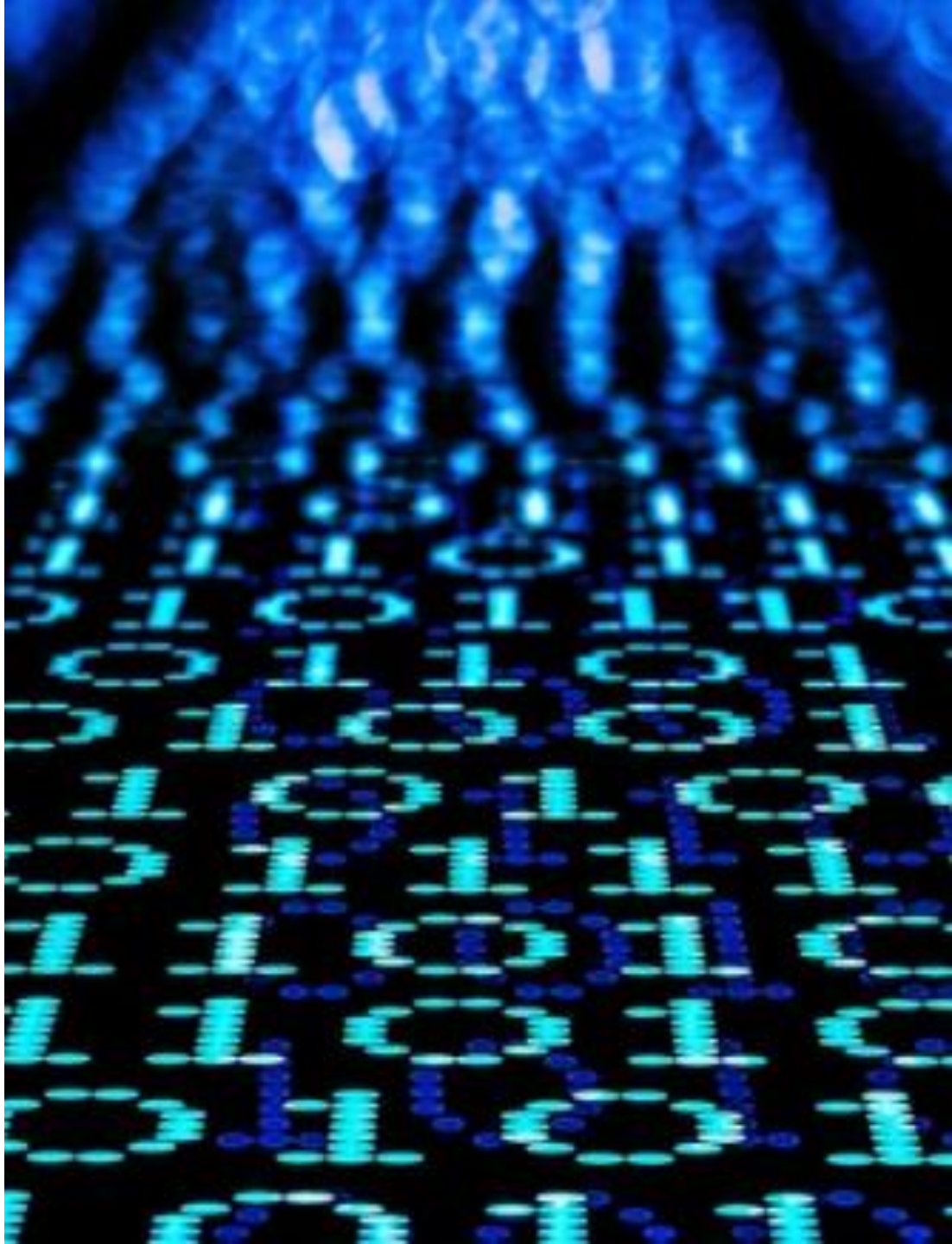
If biased AI outputs cause harm to individuals or groups, the organization using the AI could be held liable for damages

LEGAL RISKS OF LOSING CUSTOMER TRUST

- When working with generative AI, AWS highlights several legal risks associated with the loss of customer trust
- **Data privacy violations**: if generative AI systems mishandle or expose sensitive customer data, it can lead to significant privacy violations
  - This can result in legal actions and loss of customer trust

LEGAL RISKS OF LOSING CUSTOMER TRUST

- **Inaccurate or misleading outputs**: generative AI models can sometimes produce inaccurate or misleading information
  - If customers rely on this information and it leads to negative outcomes, it can damage trust and result in legal claims
- **Bias and discrimination**: if generative AI models produce biased outputs that discriminate against certain groups, it can lead to legal claims and damage the organization's reputation



LEGAL RISKS OF LOSING CUSTOMER TRUST

- **Non-compliance with regulations**: many industries have strict regulations regarding data handling and decision-making processes
  - Non-compliance due to generative AI outputs can lead to legal consequences and loss of customer trust

LEGAL RISKS ASSOCIATED WITH END USERS

Data privacy violations

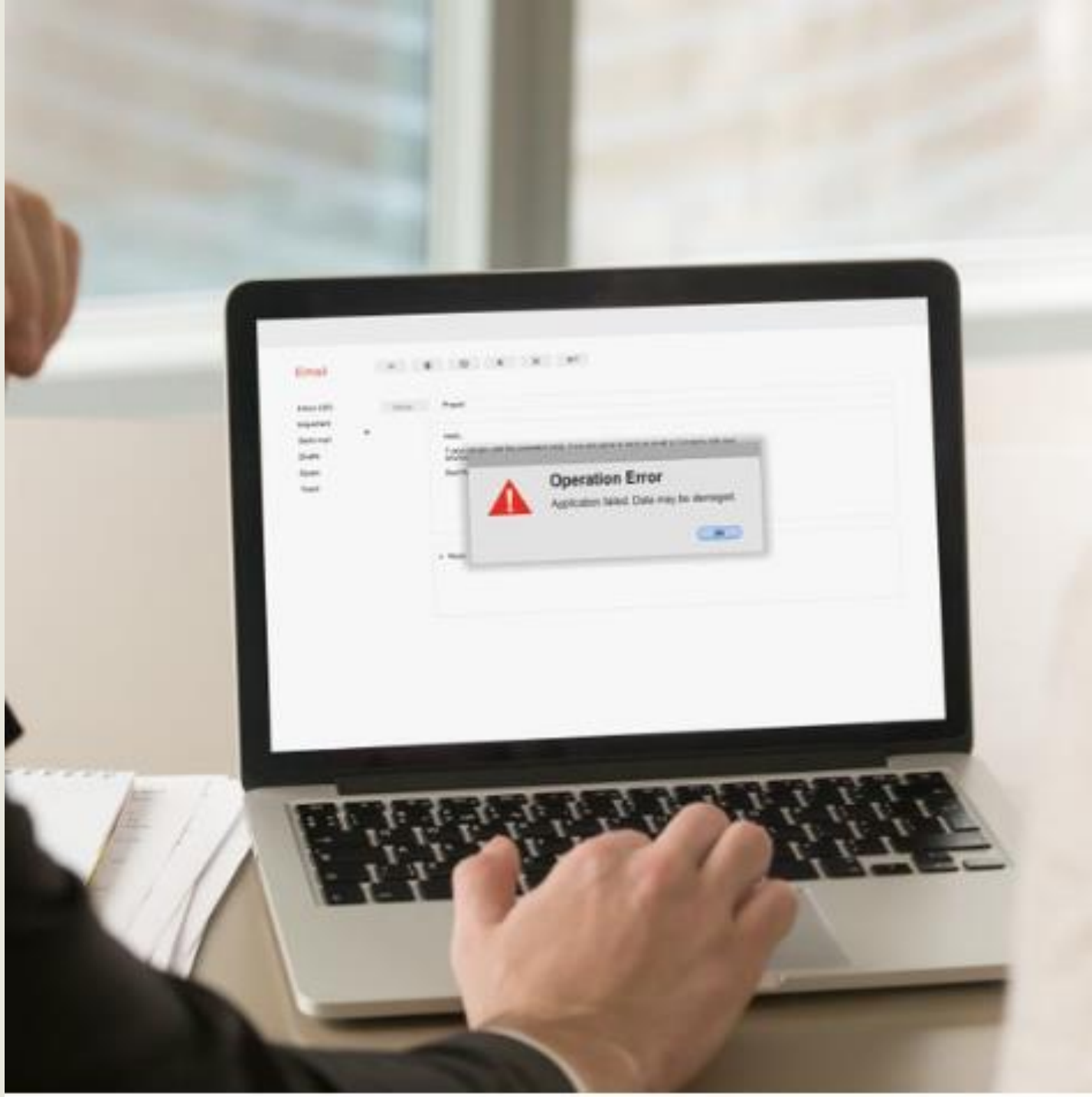
Non-compliance with regulations

Bias and discrimination

Inaccurate or misleading outputs

LEGAL RISKS ASSOCIATED WITH HALLUCINATIONS

- **Inaccurate information**: AI hallucinations can produce outputs that appear plausible but are factually incorrect
  - This can lead to the dissemination of false information, which can have serious legal implications, especially in fields like healthcare, law, and finance
- **Misleading content**: generative AI can create content that misleads users, potentially resulting in legal claims of fraud or misrepresentation



LEGAL RISKS ASSOCIATED WITH HALLUCINATIONS

- **Liability for harm:** if AI-generated hallucinations cause harm to individuals or businesses, the organization using the AI could be held liable for damages
- **Reputational damage:** the spread of incorrect information can damage an organization's reputation, leading to loss of customer trust and potential legal action

INCLUSIVITY IN DATASETS

AWS emphasizes the importance of inclusivity in datasets to ensure that AI models are fair and unbiased. Here are some key points:

- **Diverse representation**: datasets should include a wide range of data points that represent different demographics, cultures, and perspectives
  - This helps in creating models that are more inclusive and less likely to exhibit bias
- **Balanced data**: ensuring that the dataset is balanced and does not over-represent or under-represent any particular group
  - This helps in preventing the model from developing biases based on skewed data

INCLUSIVITY IN DATASETS

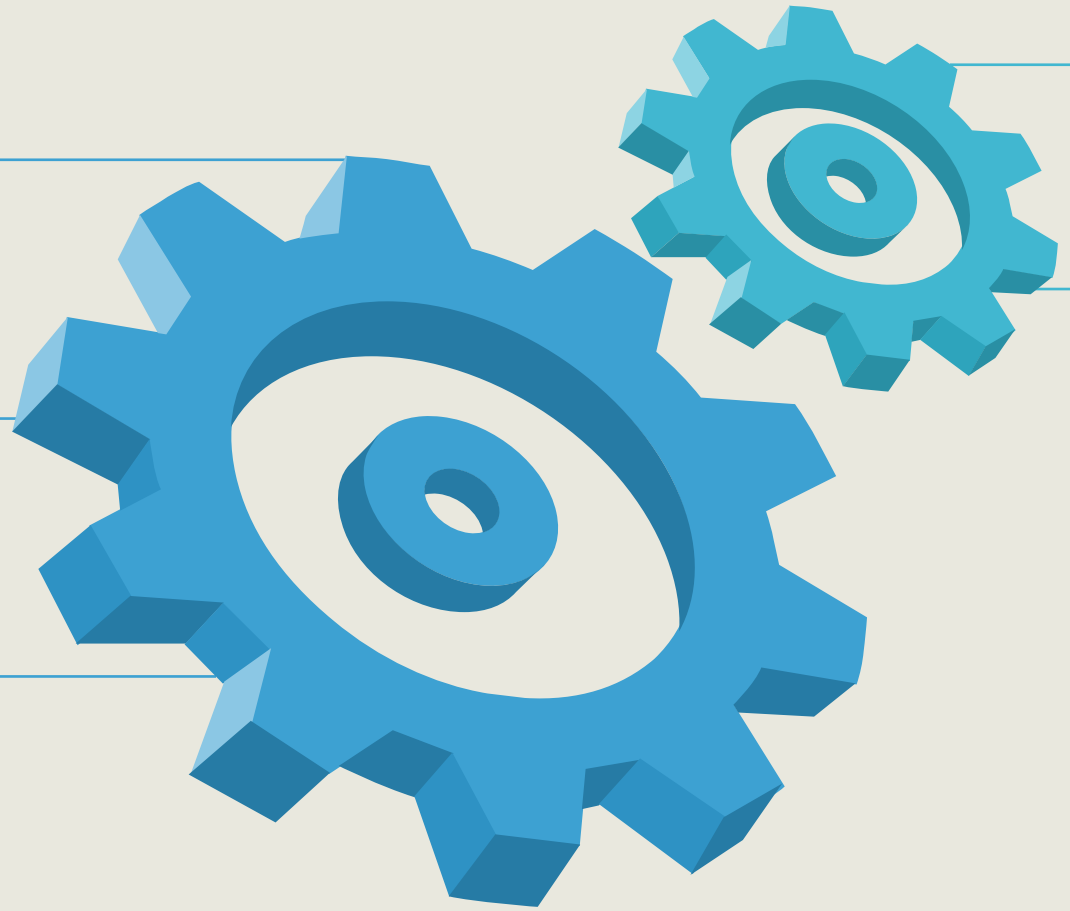
- **Bias detection and mitigation**: implementing techniques to detect and mitigate biases in the dataset
  - This includes using tools and methods to identify and correct any imbalances or biases in the data
- **Transparency**: providing clear documentation about the dataset, including its sources, collection methods, and any potential biases
  - This helps users understand the limitations and strengths of the dataset

DIVERSITY IN DATASETS

Wide range of data points

Balanced representation

Comprehensive coverage



Bias detection and mitigation

Transparency

CURATED DATA SOURCES IN DATASETS

- Curated data sources undergo rigorous **quality checks** to ensure accuracy, consistency, and reliability
  - This helps in building robust AI models that produce reliable outputs
- Curated data sources are selected based on their **relevance** to the specific use case or domain
  - This ensures that the data is contextually appropriate and enhances the model's performance

CURATED DATA SOURCES IN DATASETS

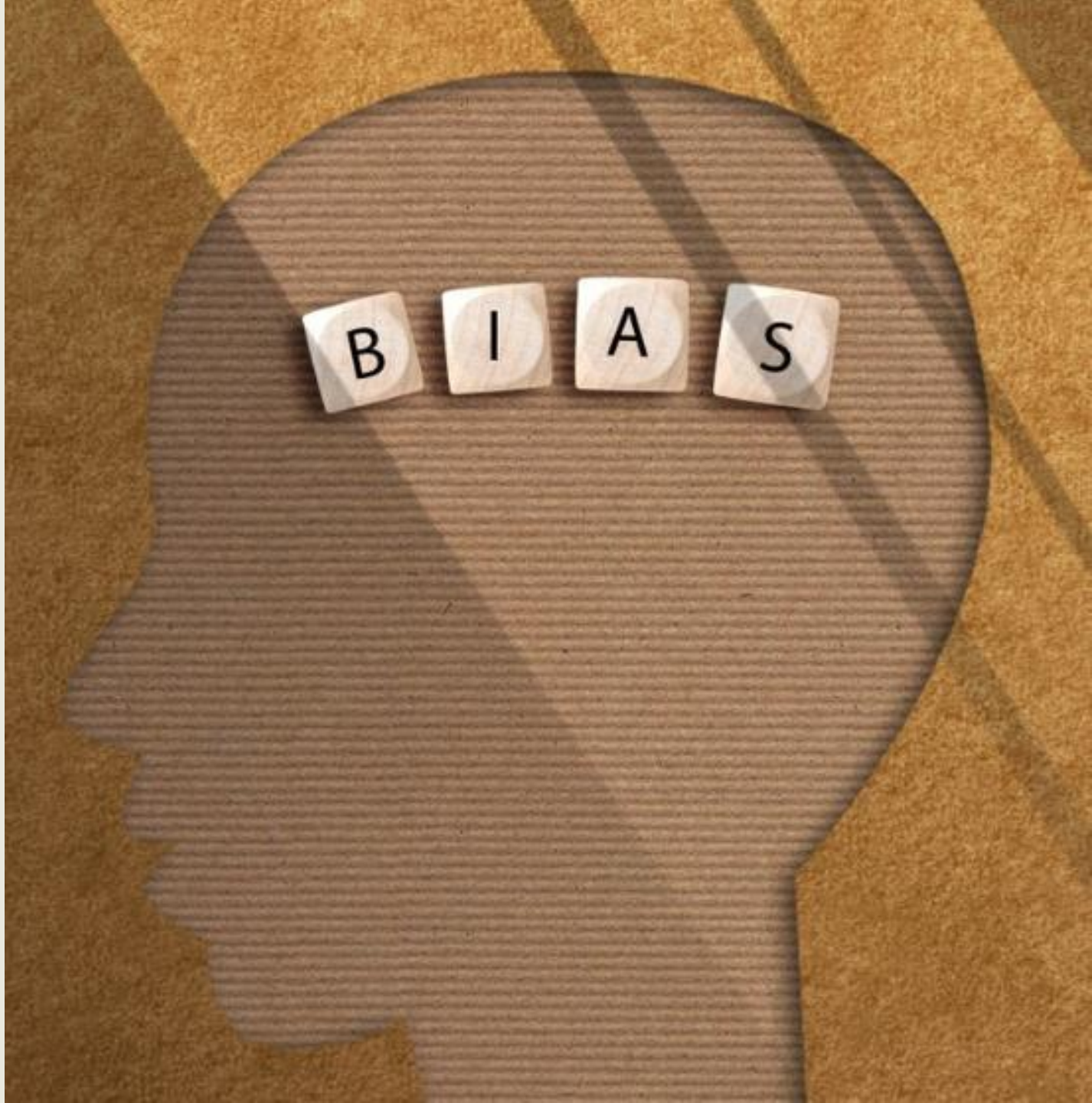
- By carefully selecting and curating data sources, AWS aims to **minimize biases** in the dataset
  - This helps in creating fair and unbiased AI models
- Curated data sources come with detailed **documentation** about their origin, collection methods, and any potential biases
  - This **transparency** helps users understand the strengths and limitations of the dataset

Balanced Datasets

- Improved model performance
- Fairness and equity
- Enhanced robustness
- Better generalization

EFFECTS OF BIAS AND VARIANCE

- AWS emphasizes the importance of addressing inaccuracies in AI models, particularly when it comes to bias and variance affecting demographic groups
- **Bias** in AI models can lead to unfair treatment of certain demographic groups
  - This can occur if the training data is not representative of the entire population, leading to skewed results that favor one group over another

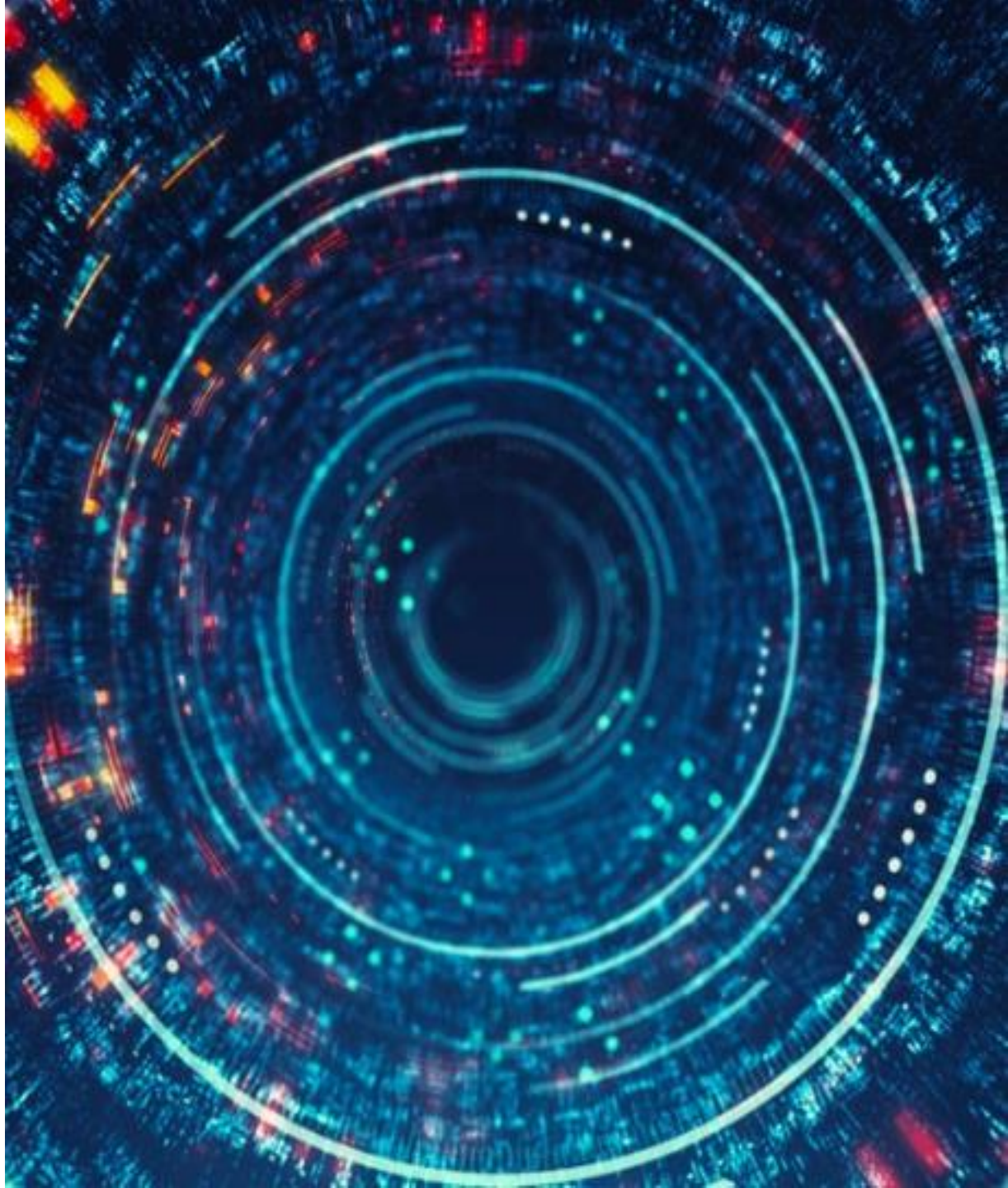


EFFECTS OF BIAS AND VARIANCE

- High **variance** in AI models can result in **overfitting**, where the model performs well on training data but poorly on new, unseen data
  - This can lead to inconsistent and unreliable predictions, which can disproportionately affect certain demographic groups if the model fails to generalize well across diverse populations
- **Inaccurate** AI models can produce misleading or incorrect outputs, which can have serious consequences for demographic groups

EFFECTS OF OVERFITTING

- **Overfitting** occurs when an ML model learns the training data too well, including the noise and outliers, and fails to generalize to new data
- This can have significant effects on bias and variance, particularly for demographic groups:
  - Overfitting results in a model with high variance
  - While overfitting typically results in low bias, the model may still exhibit biased behavior if the training data itself is biased
  - Overfitting can lead to inaccurate predictions for demographic groups that are not well-represented in the training data



EFFECTS OF UNDERFITTING

- **Underfitting** occurs when an ML model is too simple to capture the underlying structure of the data, resulting in poor performance on both the training and evaluation data
- This can have substantial effects on bias and variance, predominantly for demographic groups:
  - Underfitting results in a model with high bias
  - While underfitting typically results in low variance, the model's predictions are consistently inaccurate
  - Underfitting can lead to inaccurate predictions for all demographic groups

AMAZON SAGEMAKER CLARIFY

- SageMaker Clarify offers a comprehensive suite of tools to ensure the quality and fairness of ML models
- Label quality analysis: SageMaker Clarify evaluates the quality of labels by analyzing the consistency and accuracy of the labels in the dataset
- It uses various metrics to identify potential issues such as mislabeled data, label noise, and inconsistencies
  - This helps in improving the overall quality of the dataset, which is crucial for building reliable ML models

AMAZON SAGEMAKER CLARIFY

Human audits: human audits in SageMaker Clarify involve using human evaluators to review and validate model predictions

This is particularly useful for tasks that require nuanced judgment or where automated metrics might not be sufficient

- The results of these audits are saved in Amazon S3 and can be used to understand how human evaluators perceive the model's performance

AMAZON SAGEMAKER CLARIFY

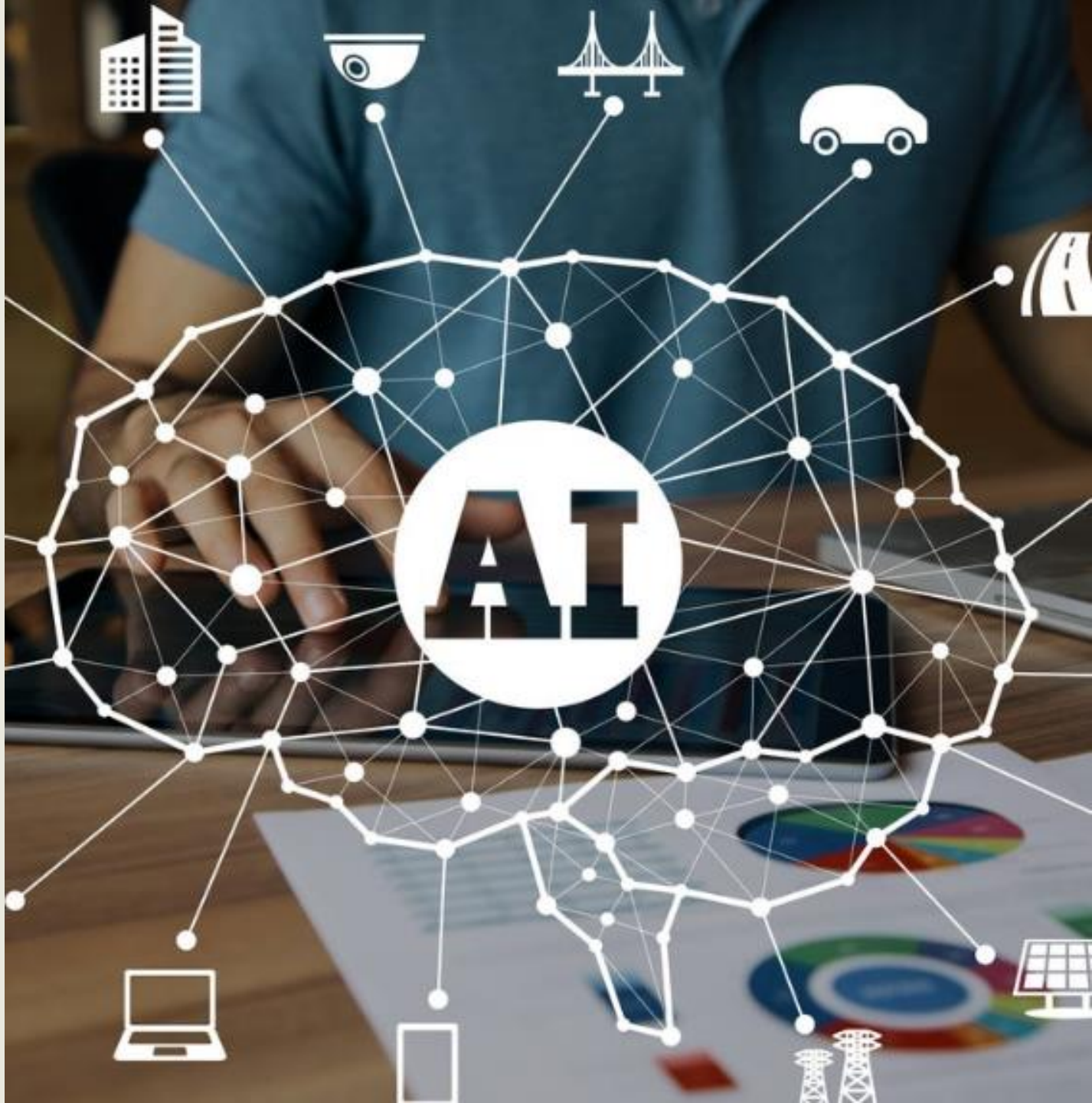
- **Subgroup analysis** in SageMaker Clarify involves examining the performance of the model across different subgroups within the dataset
  - This helps in identifying any biases or disparities in the model's predictions for different demographic groups
- The analysis is configured through an analysis file that specifies the parameters for bias detection and explainability
  - The results are aggregated into a report

AMAZON SAGEMAKER MODEL MONITOR

- Amazon SageMaker Model Monitor is a tool designed to continuously monitor the quality of machine learning (ML) models in production
- It helps ensure that models maintain their performance and reliability over time by performing:
  - Data quality monitoring
  - Model quality monitoring
  - Bias detection
  - Feature attribution drift

AMAZON AUGMENTED AI (AMAZON A2I)

- Amazon A2I is a fully managed service that makes it easier to incorporate human reviews into machine learning (ML) workflows
- It allows the integration of human judgment into ML applications to ensure higher accuracy and reliability, especially for tasks that require nuanced decision-making

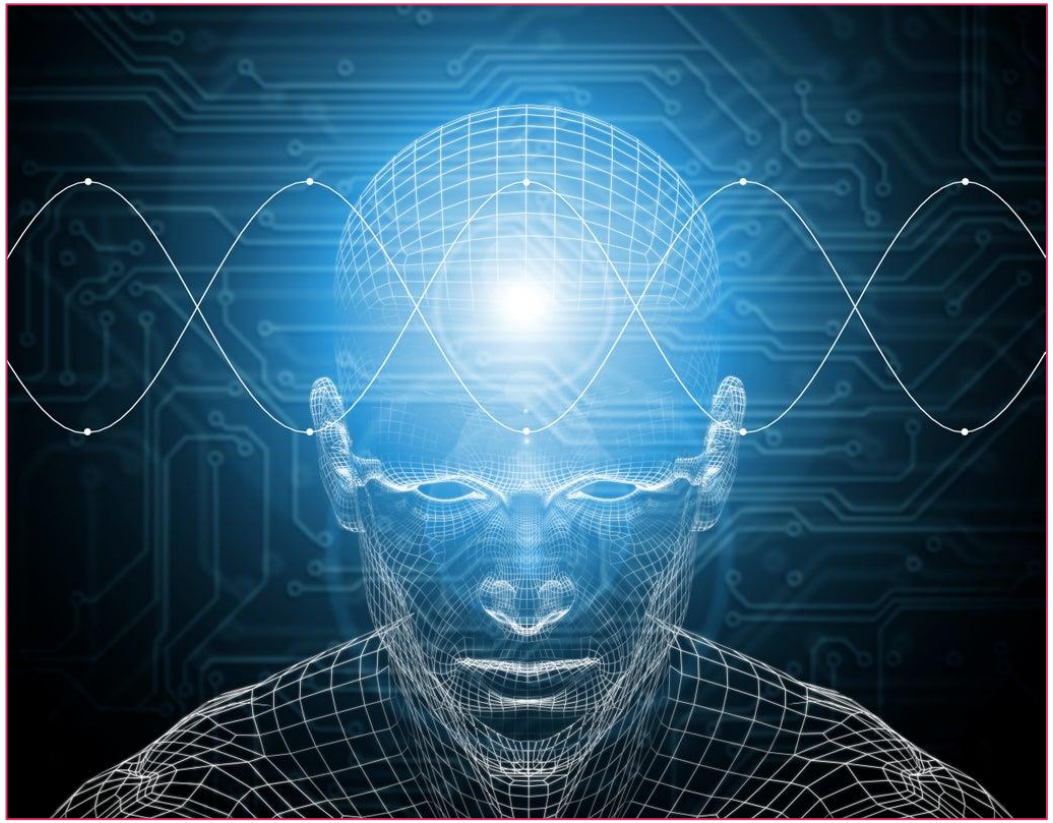


TRANSPARENT ML MODELS

- A transparent machine learning model is one whose decision-making process can be easily understood and interpreted by humans
- Transparent models are crucial for ensuring fairness, accountability, and trust in AI systems
- They allow users to understand how decisions are made, which is especially important in sensitive applications like healthcare, finance, and law

NON-TRANSPARENT ML MODELS

- Non-transparent ML models (or black-box) are those whose internal workings are not easily interpretable by humans
- These models can make accurate predictions, but understanding how they arrive at those predictions is challenging



EXPLAINABLE ML MODELS

- Explainable ML models are designed to make their decision-making processes understandable to humans
- These models provide insights into how they arrive at their predictions or classifications, which is crucial for building trust, ensuring fairness, and meeting regulatory requirements
- Explainable models can help identify and mitigate biases, improve transparency, and facilitate better decision-making

NON-EXPLAINABLE ML MODELS

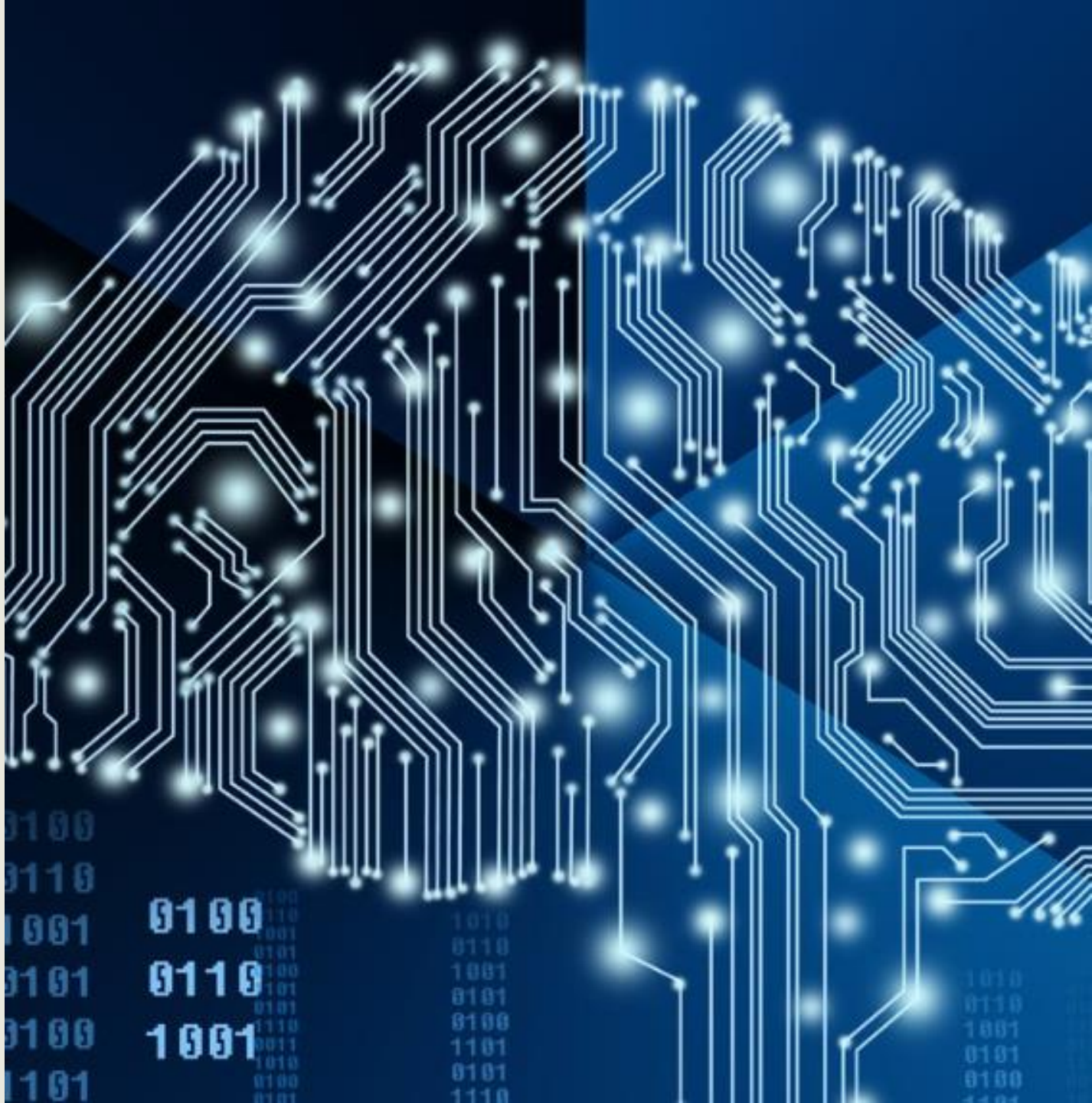
- Non-explainable machine learning models are those whose decision-making processes are not easily interpretable by humans
- These models can make accurate predictions, but understanding how they arrive at those predictions is challenging
- The lack of explainability can be a concern in applications where understanding the decision-making process is crucial, such as in healthcare, finance, and legal systems

TRADEOFFS WITH FOUNDATION MODELS

- **Interpretability**: this refers to how easily humans can understand and trust the decisions made by AI models
- **Performance**: often, more complex models (like deep neural networks) achieve higher accuracy but are less interpretable
  - Simpler models (like decision trees) are more interpretable but may not perform as well

TRADEOFFS WITH FOUNDATION MODELS

- Balancing interpretability and performance is challenging
- High-performing models may lack transparency and interpretability, making it difficult to understand their decisions
- Conversely, highly interpretable models may not perform as well



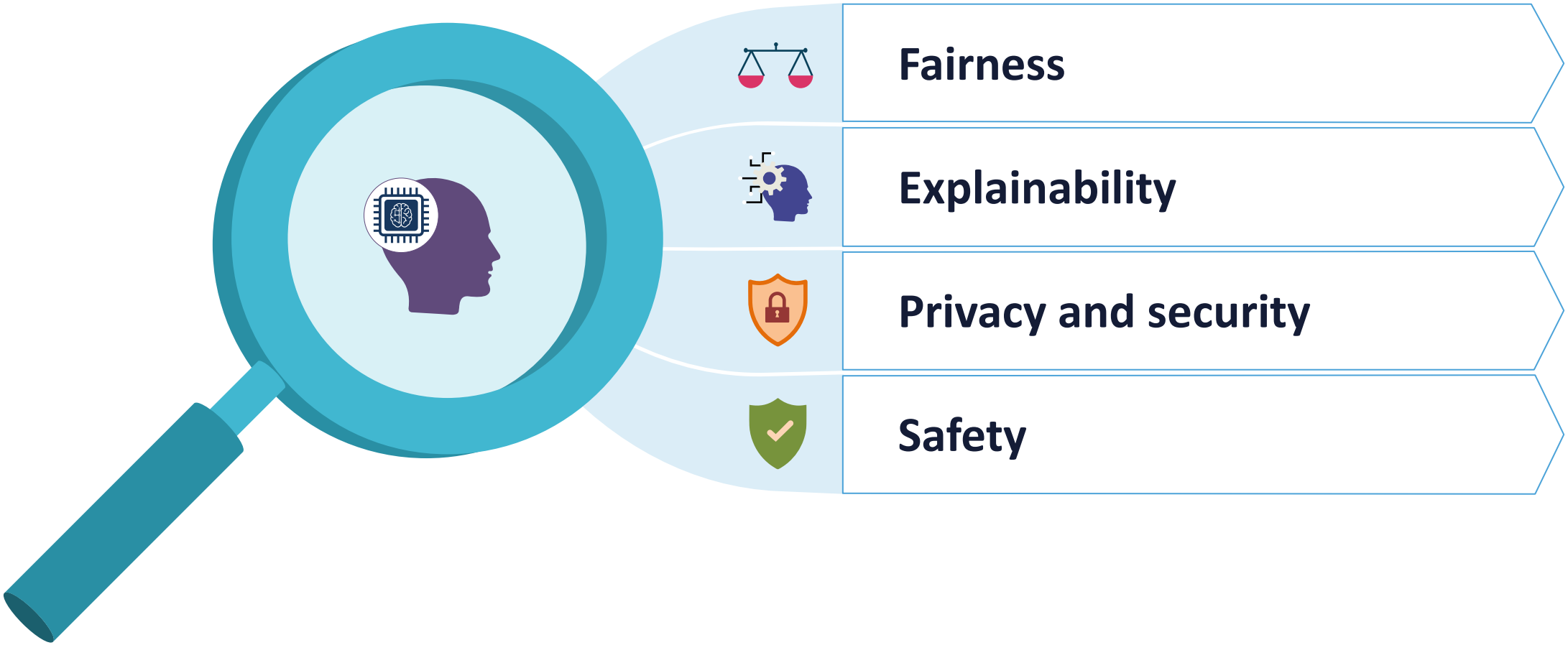
TRADEOFFS WITH FOUNDATION MODELS

- **Safety**: ensuring that AI models operate safely and ethically is paramount
  - Transparent and interpretable models help in identifying and mitigating risks
- **Accountability**: transparent models allow stakeholders to understand and trust AI decisions, ensuring accountability in AI systems

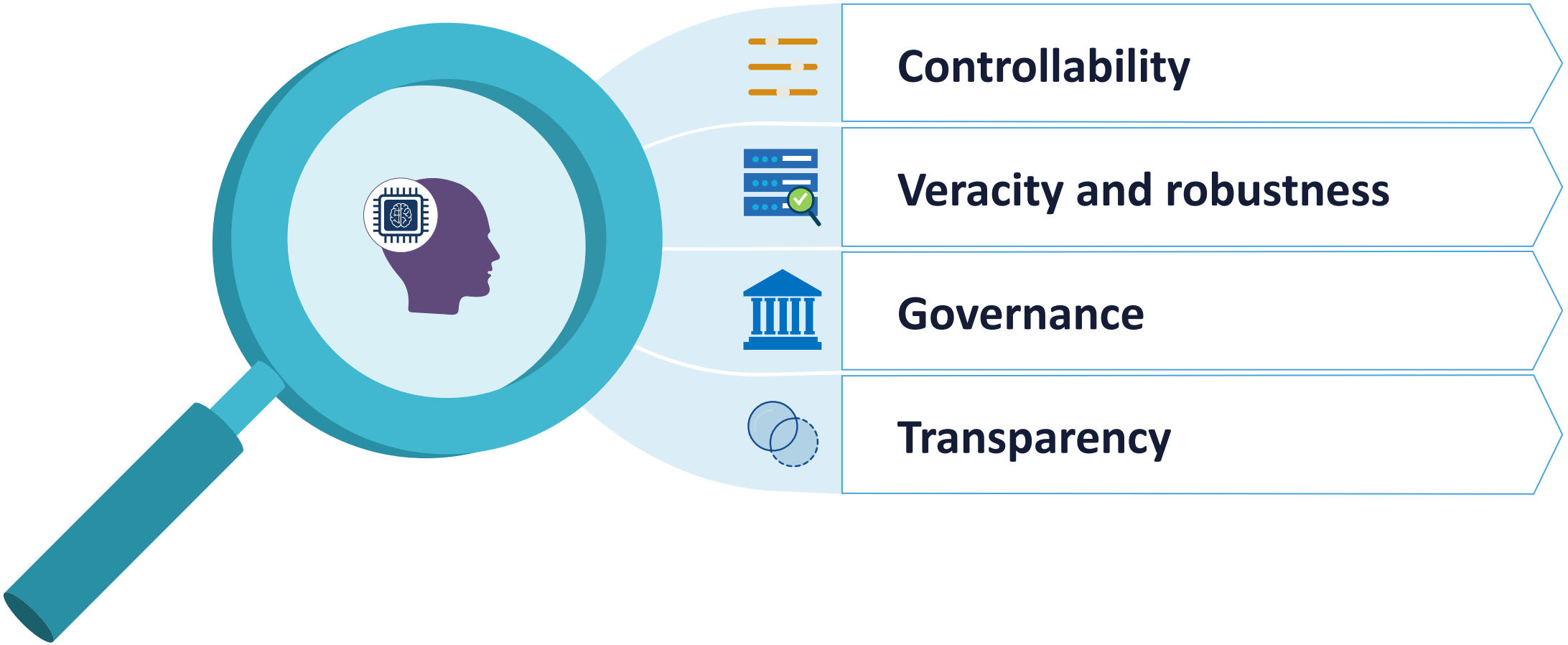
TRADEOFFS WITH FOUNDATION MODELS

- In summary, achieving a balance between interpretability, performance, safety, and transparency is essential for the responsible deployment of AI models
- This tradeoff requires careful consideration and the use of techniques that enhance both interpretability and performance

PRINCIPLES OF HUMAN-CENTERED DESIGN FOR EXPLAINABLE AI



PRINCIPLES OF HUMAN-CENTERED DESIGN FOR EXPLAINABLE AI



SECURITY, COMPLIANCE, AND GOVERNANCE FOR AI SOLUTIONS

Objectives

- Describe securing AI systems and services
- Examine source citation and documenting data origins
- Learn about secure data engineering
- Define security, privacy, and regulatory compliance standard considerations for AI systems
- Compare services to assist with governance and regulation compliance
- Describe data governance strategies
- Compare processes for following governance protocols

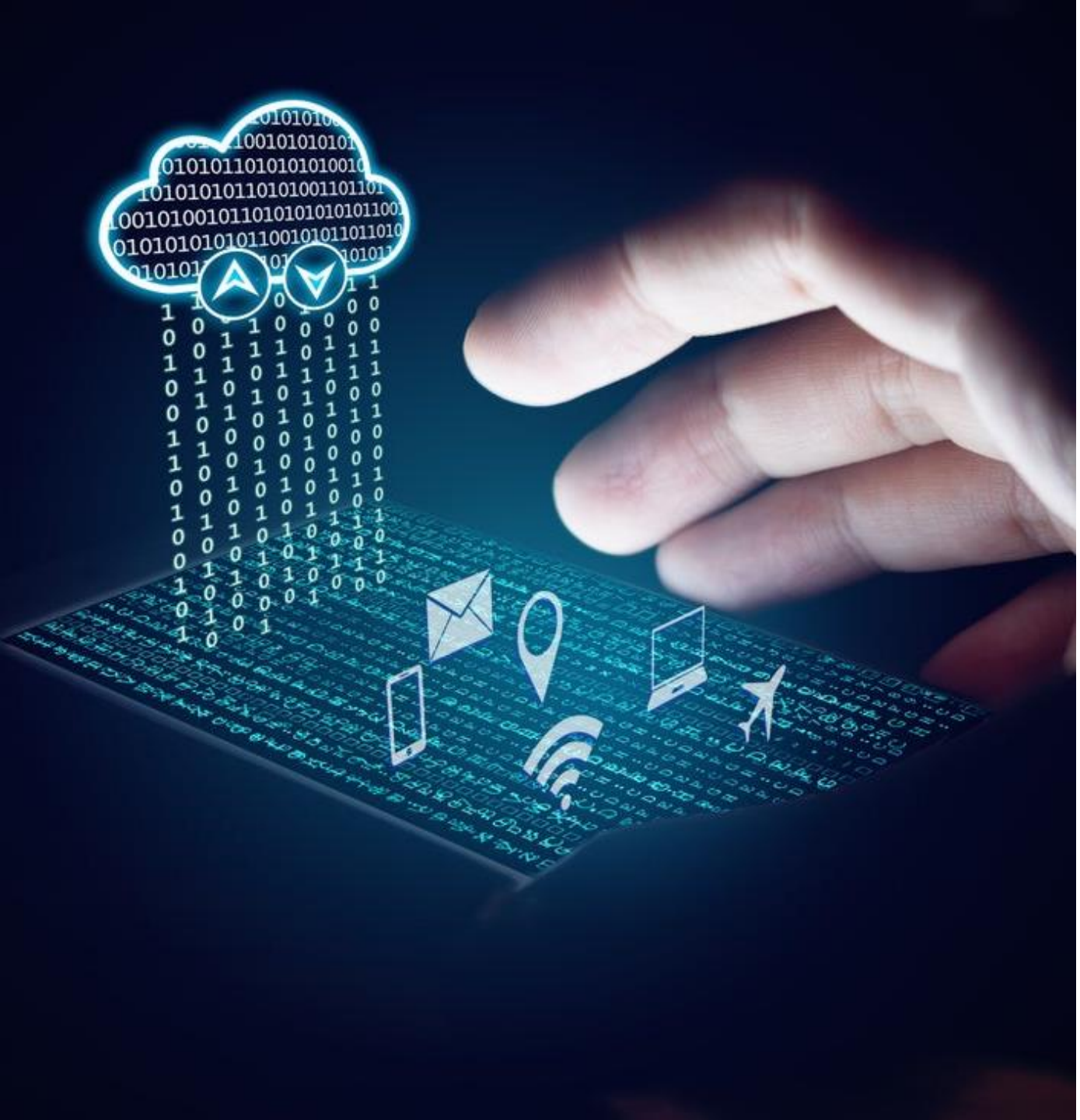
SECURING AI SYSTEMS AND SERVICES

In this demo…

We will explore AWS services and features to secure AI systems such as IAM roles, policies, permissions, encryption, Amazon Macie, AWS PrivateLink, and the AWS shared responsibility model.

SOURCE CITATION

- An AI source citation is a segment of a generated response that is based on a source in the knowledge base, along with information about that source
- This includes metadata about the sources cited for the generated response



SOURCE CITATION

- Source citations are important because they provide transparency and credibility to the information generated by AI
- They allow users to verify the accuracy of the information by checking the original sources
- This is especially crucial in ensuring that the AI-generated content is reliable and trustworthy

DATA LINEAGE

- AWS describes documenting data origins using data lineage to capture and visualize lineage events from OpenLineage-enabled systems or through APIs
- This helps trace data origins, track transformations, and view cross-organizational data consumption
- Data lineage provides an overarching view into data assets, allowing users to see the origin of assets and their chain of connections

DATA LINEAGE

- In Amazon DataZone, data lineage includes information on the activities inside the business data catalog, such as cataloged assets, subscribers, and activities captured programmatically using APIs
- This historical lineage helps in troubleshooting, auditing, and ensuring the integrity of data assets

DATA CATALOGING

- AWS describes documenting data origins as using data cataloging through the **AWS Glue Data Catalog**
- This catalog acts as a centralized repository that stores metadata about the organization's data sets
- It serves as an index to the location, schema, and runtime metrics of the data sources

DATA CATALOGING

- The AWS Glue Data Catalog can be populated using crawlers, which automatically scan the data sources and extract metadata
- These crawlers can connect to both internal (AWS-based) and external data sources
- Additionally, tables can be manually created in the Data Catalog by defining the table structure, schema, and partitioning structure according to specific requirements

DATA CATALOGING

- The Data Catalog integrates with other AWS analytics services, providing a unified view of data sources and making it easier to manage and analyze data
- For example, one can store and query table metadata in the Data Catalog for Amazon S3 data using SQL with Amazon Athena

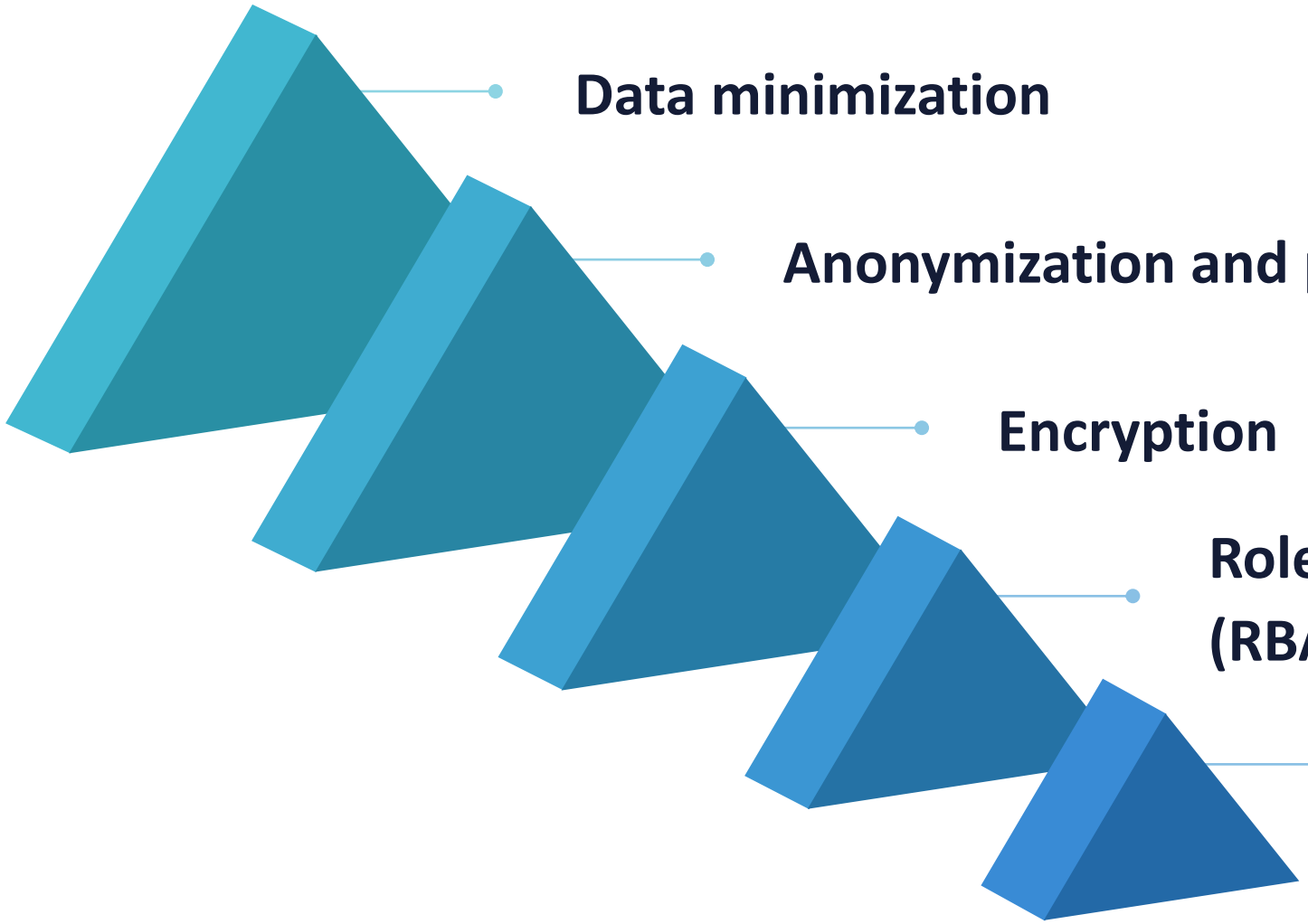
SAGEMAKER MODEL CARDS

- Amazon SageMaker Model Cards are a feature within Amazon SageMaker that allows users to document critical details about their ML models in a single place for streamlined governance and reporting
- These model cards help capture key information about models throughout their lifecycle and implement responsible AI practices

SAGEMAKER MODEL CARDS

- Model cards include details such as the intended use and risk rating of a model, training details and metrics, evaluation results and observations, and additional call-outs like considerations, recommendations, and custom information
- Users can provide guidance on how a model should be used and support audit activities with detailed descriptions of model training and performance

PRIVACY-ENHANCING TECHNOLOGIES IN SECURE DATA ENGINEERING



Data minimization

Anonymization and pseudonymization

Encryption

Role-based access control (RBAC)

Privacy by design

DATA ACCESS CONTROL IN SECURE DATA ENGINEERING

Identity and access management (IAM)

Data classification

Encryption

Role-based access control (RBAC)

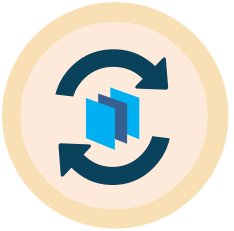
Monitoring and alerts

DATA ACCESS CONTROL IN SECURE DATA ENGINEERING

Data validation

Data auditing

Checksum and hashing



Version control

Data replication

SECURITY AND PRIVACY CONSIDERATIONS FOR AI SYSTEMS

- **Application security**: implement security measures during the software development lifecycle to detect and mitigate vulnerabilities in AI workloads
- **Threat detection**: continuously monitor AI systems to detect and mitigate potential security threats or unexpected behaviors

SECURITY AND PRIVACY CONSIDERATIONS FOR AI SYSTEMS

- **Vulnerability management**: identify, classify, remediate, and mitigate AI-specific vulnerabilities, such as prompt injection, data poisoning, and model inversion
- **Infrastructure protection**: secure the systems and services used to operate AI workloads, ensuring data protection at every layer of the AI stack

SECURITY AND PRIVACY CONSIDERATIONS FOR AI SYSTEMS

- **Prompt injection**: protect AI models from prompt injection attacks by implementing robust input validation and monitoring techniques
- **Encryption of data at rest and in transit**: use strong encryption methods to protect sensitive data both at rest and in transit, ensuring data confidentiality and integrity

REGULATORY COMPLIANCE STANDARDS

- AWS emphasizes the importance of ISO regulatory compliance standards for AI systems, particularly through the **ISO/IEC 42001:2023** standard
  - This standard outlines requirements for establishing, implementing, maintaining, and continually improving an AI Management System (AIMS) within organizations

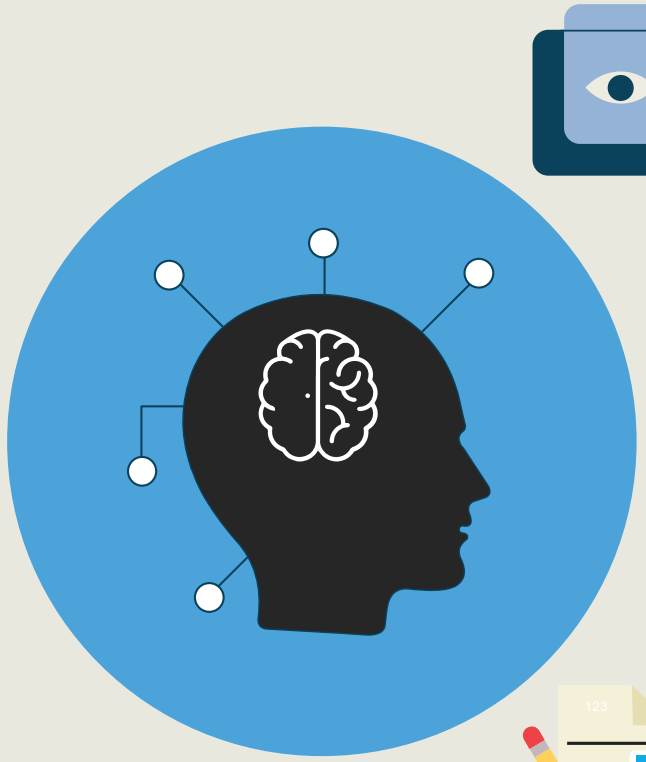
REGULATORY COMPLIANCE STANDARDS

- AWS also emphasizes the importance of System and Organization Controls (SOC) regulatory compliance standards for AI systems
- SOC 1 report: this report focuses on controls relevant to user entities' internal control over financial reporting
  - It helps customers understand AWS's control environment and how it supports their financial reporting requirements

REGULATORY COMPLIANCE STANDARDS

- **SOC 2 report: this report covers security, availability, confidentiality, and privacy**
  - **It provides an independent assessment of AWS's control environment and helps customers evaluate the effectiveness of these controls for their AI workloads**
- SOC 3 report: this is a public report that also covers security, availability, confidentiality, and privacy
  - It demonstrates AWS's commitment to meeting the AICPA Trust Services Criteria and provides assurance to customers and stakeholders

ALGORITHM ACCOUNTABILITY LAWS



TRANSPARENCY

FAIRNESS

ACCOUNTABILITY

AWS CONFIG FOR AI GOVERNANCE AND REGULATION COMPLIANCE

- Compliance validation
- Resource monitoring
- Automated compliance checks
- Audit trails
- Integration with other AWS services

AWS CONFIG FOR AI GOVERNANCE AND REGULATION COMPLIANCE

- Security testing
- Compliance validation
- Continuous monitoring
- Integration with other AWS services

AWS AUDIT MANAGER

- AWS Audit Manager assists with AI governance and regulation compliance by providing a structured approach to evaluating and adopting AI technologies
- IT can automate the collection of evidence, such as resource configurations and usage activity, to ensure compliance with regulatory requirements

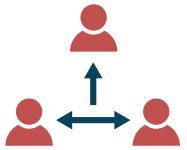
AWS AUDIT MANAGER

- Audit Manager generates customized assessment reports that help organizations monitor and demonstrate compliance with AI-specific controls
- It offers prebuilt frameworks, such as the AWS generative AI best practices framework
- It continuously monitors AI workloads to ensure they adhere to compliance requirements and best practices

AWS ARTIFACT FOR AI GOVERNANCE AND REGULATION COMPLIANCE

Compliance reports

Audit support



Third-party assessments

Agreement management

AWS CLOUDTRAIL FOR AI GOVERNANCE AND REGULATION COMPLIANCE

Audit trails

Compliance validation

Security monitoring

AWS service integration

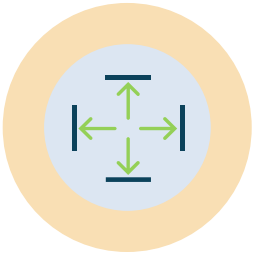
AWS CLOUDTRAIL FOR AI GOVERNANCE AND REGULATION COMPLIANCE

Security checks

Cost optimization

Performance improvement

Fault tolerance



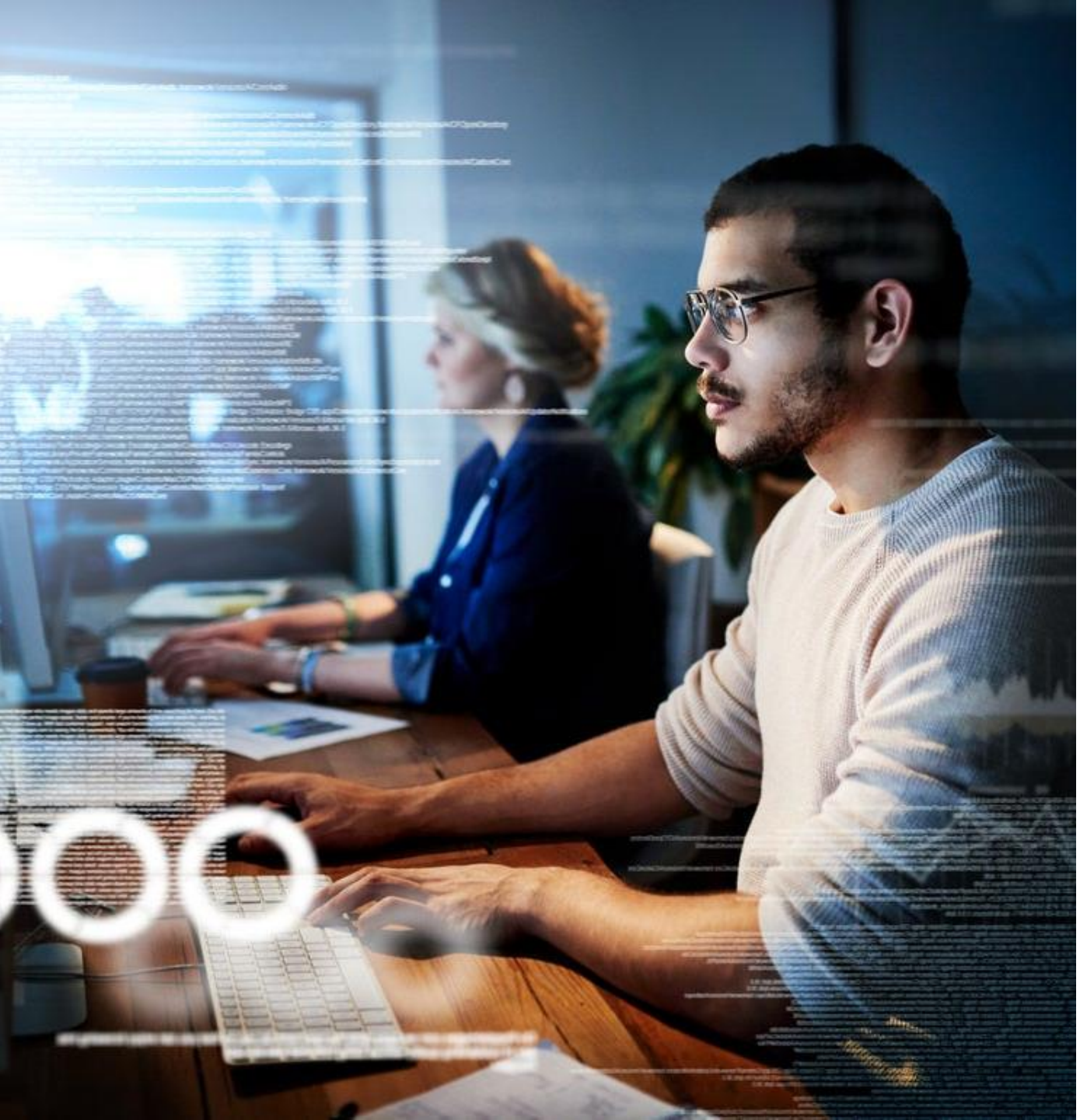
Service limits

DATA GOVERNANCE STRATEGIES

- **Data lifecycles**: AWS recommends establishing data retention policies to ensure data is retained for the appropriate period
  - AWS S3 lifecycle policies can automate data retention and deletion
- **Logging**: AWS CloudTrail provides detailed logs of all API calls made within an AWS account, creating a comprehensive audit trail essential for compliance and governance

DATA GOVERNANCE STRATEGIES

- **Data residency**: AWS offers various regions and availability zones to help organizations comply with data residency requirements by storing data within specific geographic locations
- **Monitoring**: AWS Config continuously monitors and records AWS resource configurations, allowing you to track changes and ensure they meet compliance requirements



DATA GOVERNANCE STRATEGIES

- **Observation**: Amazon CloudWatch provides monitoring and observability of AWS resources and applications, helping you gain insights into system performance and operational health
- **Data retention**: Amazon Glacier is recommended for long-term data archiving, and secure data disposal practices ensure data is properly deleted when no longer needed

AI GOVERNANCE PROTOCOLS

- AWS emphasizes the importance of AI governance protocols to ensure responsible and ethical use of AI technologies
- These processes help organizations maintain high standards of governance and security for their AI systems while ensuring compliance with regulatory requirements

AI GOVERNANCE PROTOCOLS

- **Policies**: develop comprehensive policies that cover data usage, transparency, responsible AI, and compliance
  - These policies should be aligned with organizational goals and regulatory requirements
- **Review cadence**: establish a regular review cadence to assess the performance, compliance, and ethical considerations of AI systems

AI GOVERNANCE PROTOCOLS

- The review cadence also includes periodic evaluations to ensure AI systems are functioning as intended and meeting governance standards
- **Review strategies**: implement review strategies that involve key stakeholders from multiple business units
  - This includes defining governance goals, monitoring AI systems for bias and compliance, and taking corrective actions based on predefined thresholds

AI Governance Frameworks

- **Generative AI Security Scoping Matrix**: this framework helps organizations assess and implement security controls throughout the AI lifecycle
- It breaks down security considerations into specific categories, enabling a focused approach to securing AI applications

AI Governance Frameworks

- **Transparency standards**: AWS emphasizes the importance of transparency in AI systems
- This includes providing clear and understandable explanations of AI decision-making processes, ensuring fairness, and promoting responsible AI

AI Governance Frameworks

- **Team training requirements**: AWS offers tailored training courses and learning plans to upskill teams for generative AI projects
- This includes courses for business decision-makers, data specialists, and developers to ensure they have the necessary skills to build and manage AI systems responsibly

Thank You for Attending

All the best in your Cloud Computing and IT future!

Michael J. Shannon
and
Eian Clair