AI in Production at Scale with AWS Sagemaker and Teradata Vantage

Dr. Chris Hillman,
Data Science Senior Director International, Teradata

Tomas Sykora,
Principal Solutions Architect, AWS

teradata.

AI at scale is a problem for many

All of AI… has a proof-of-concept-to-production gap… the full cycle of a machine learning project is not just modelling… it is finding the right data, deploying it, monitoring it, feeding data back… **doing all the things that need to be done [for a model] to be deployed**… [that goes] beyond doing well on the test set, which fortunately or unfortunately is what we in machine learning are great at."

– Andrew Ng, 2021

DATA:80%
of all project time is spent preparing data—not creating value

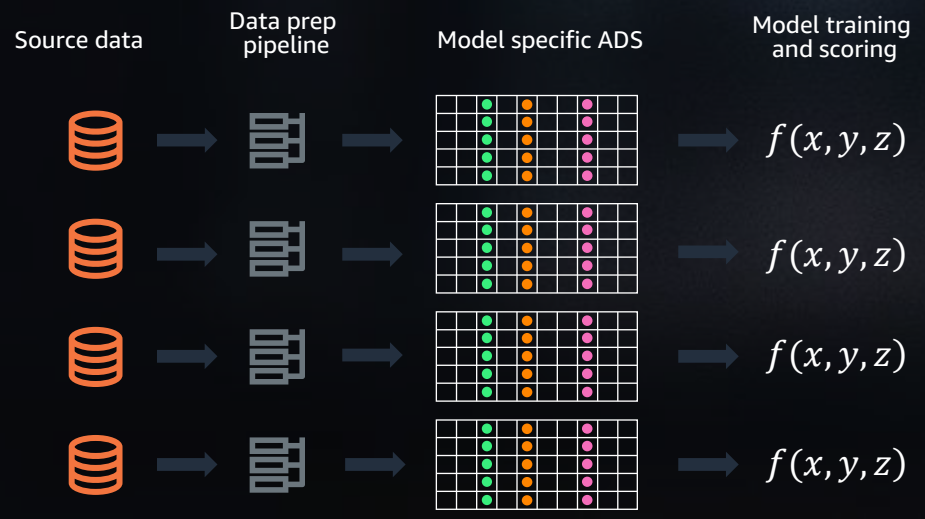
SCALE:100x
Increasing AI/ML adoption will require a 100x increase in the number of models and queries

DEPLOY:65%
of predictive models are never implemented in production

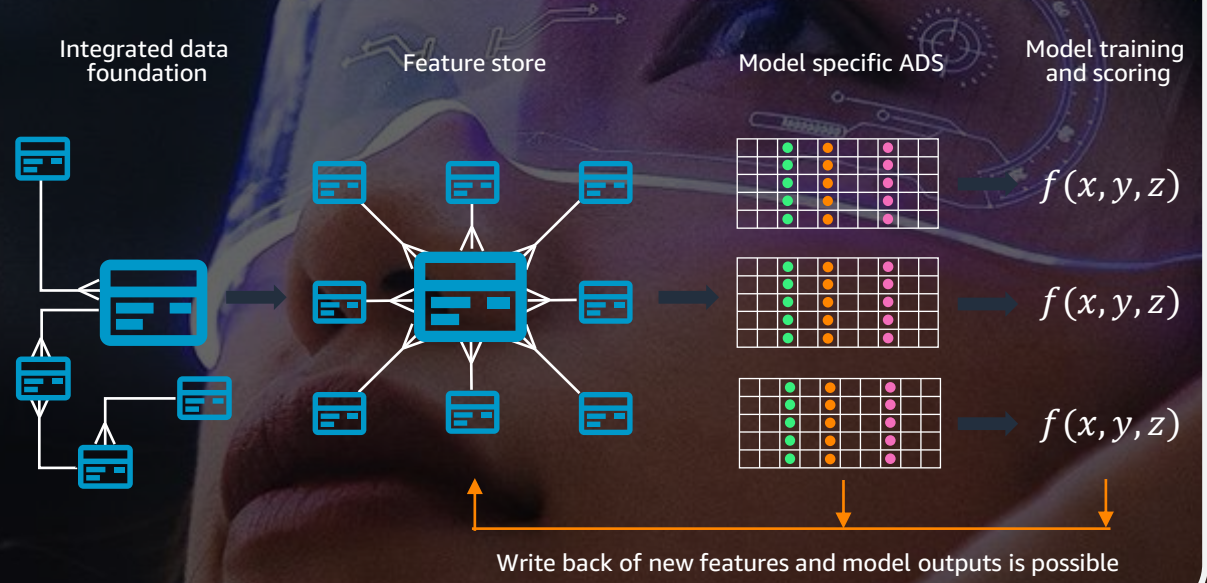
AI is a Data Problem

The Enterprise Feature Store

The one pipeline per model approach



The feature store approach



80%
of all project time is spent preparing data—not creating value

100x
Increasing AI/ML adoption will require a 100x increase in the number of models and queries

65%
of predictive models are never implemented in production

teradata.

AI is a Scale Problem

Millions of Models in Production

A **$6.6 billion national retailer** needed to manage hundreds of thousands of products across **multiple channels, including e-commerce and over 280 stores**.

To forecast product demand, ClearScape Analytics and the Python Prophet library were utilized directly within Teradata Vantage.

2.64M
demand-forecasting models trained in three hours

360K
seasonality-profiling models trained in 15 minutes

0kB
move processing to data, not the other way around

teradata.

AI is a Deployment Problem

AI for Inference

Using Large Language Models for real-time recommendations direct to the customer Shopping Cart

80%
of all project time is spent preparing data—not creating value

100x
Increasing AI/ML adoption will require a 100x increase in the number of models and queries

65%
of predictive models are never implemented in production

teradata.

ClearScape Analytics™ is designed to solve these problems

- Extensive Library of in-database functions
- Bring your own models to the data
- Full ModelOps Solution
- OpenAPIs and partner integration
- Leverage Tools of your choice
- Security, Privacy and Governance

Feature Engineering and Discovery at scale

Model Training – extensive native library and open to all tools

Automated model management and monitoring

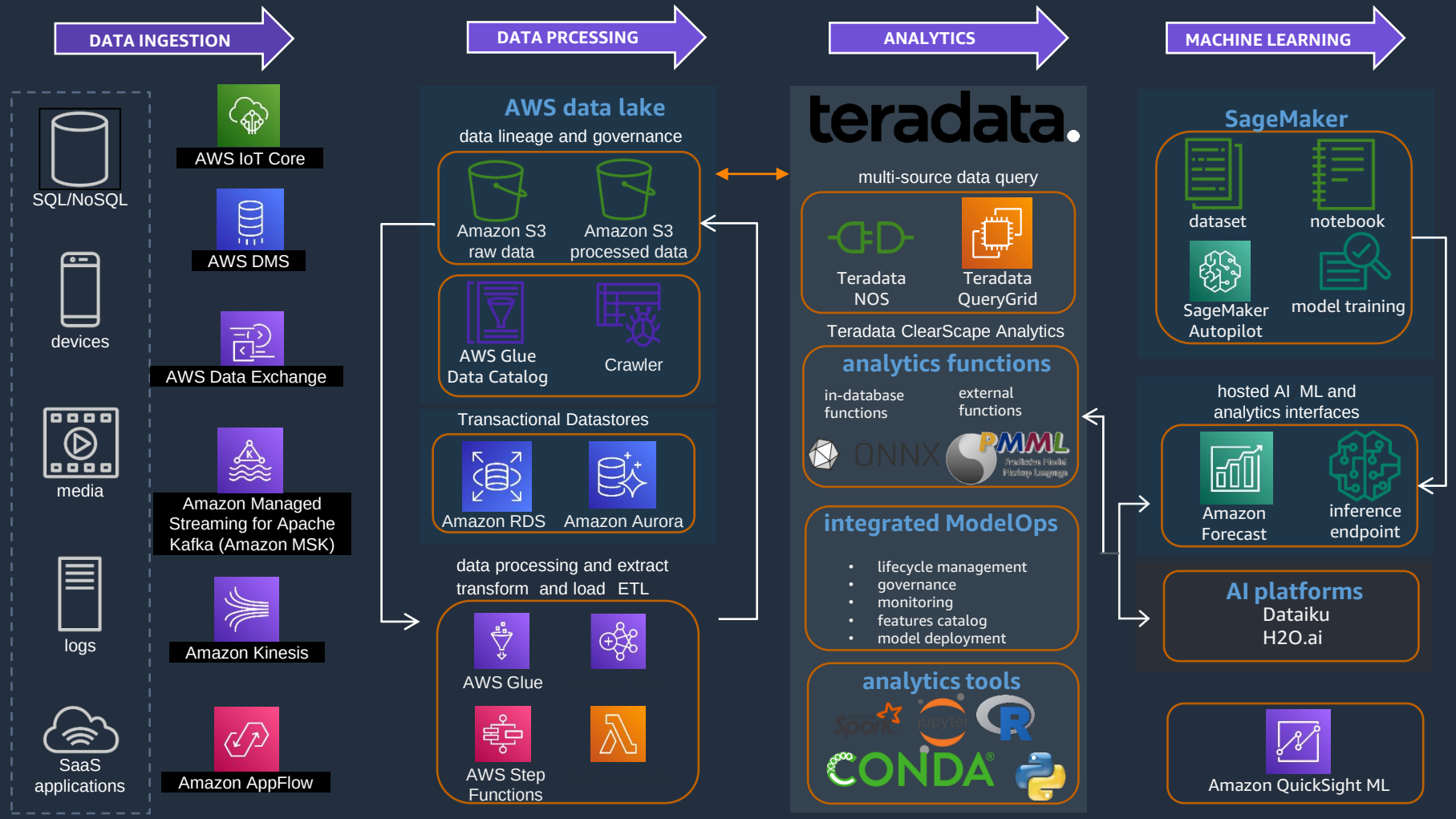
Deep Integration with Sagemaker

Use the language and tools of your choice e.g. Python, R, SQL

All built into the Vantage platform

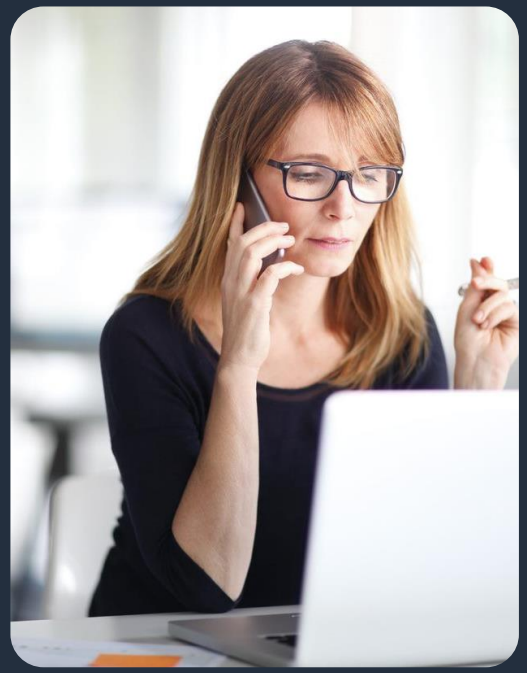
Modern Data Pipeline using AWS and Teradata

VantageCloud Enterprise is part of the Teradata VantageCloud offering, the complete cloud analytics and data platform that includes Teradata's significantly expanded ClearScape Analytics.



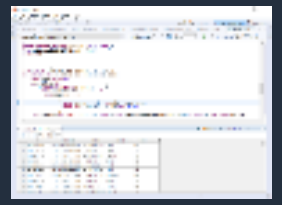
Vantage API integration with AWS analytic services

Operationalize Analytic Models with real-time access by Vantage Business Users



SQL queries

Vantage

SQL and analytic functions

Enterprise data

API

AWS AutoML services

- Amazon Forecast
- Amazon Fraud Detector
- Amazon Augmented AI
- Amazon Comprehend

AWS data science

Amazon SageMaker

Analytics 1-2-3

50-80% of time is taken preparing raw data:

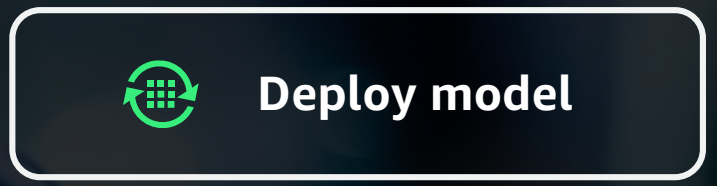
- Data integration
- Data access and exploration
- Data cleansing
- Feature engineering
- Feature selection

Leverage **Data Labs** to support rapid experimentation and build a **Feature Store** of variables with known predictive value.

Fit ML algorithm to the training data:

- Algorithm selection
- Test and training data-set split
- Model training and evaluation
- Model optimization
- Model export

Be prepared to use multiple analytic tools—but ensure that they are trained on data pulled from the Feature Store and that **models are consumable.**



Operationalize model to predict outcomes:

- Write-back new features
- Import model to model repository
- Operational scoring
- Business process integration
- Model monitoring

Bring models to the data in the Feature Store wherever possible; instrument models to capture meta-data and predictions.

AI is a security problem

89%: Data Transformation is part of all business growth strategies

50-500 data-sources to train models

Wide access to data for decision making

66%: Security is biggest threat in next two years

73%: Developers forced to compromise on security

82%: At least one data breach

Comprehensive AI strategy: People, Processes, not just Technology

Frameworks: NIST Cyber Security Framework, MITRE ATT&CK, Center for Internet Security (CIS), SOC2, Amazon Well-architected: Security pillar

Security considerations

Access control

- Limit, and monitor access to training data, particularly personal or sensitive data.
- Both in scope and time!

Authentication

- Require strong authentication mechanisms to access training data.
- Monitor for suspicious or abnormal access patterns.
- Defined escalation paths.

Lifecycle policies

- Clear, up to date, governed and enforced policies on what to do with datasets after training is complete.

3rd parties

- Verify that 3rd data providers and consumers follow reasonable security and compliance standards.

AI is a problem of trust

85%: customers prefer products transparent on how ML is trained, used and monitored

56%: trusted AI is important to maintain brand

50%: to satisfy regulatory requirements

63%: lack of skills to maintain trusted AI

59%: lack of AI strategy

57%: AI not explainable, inherent bias

Responsible AI reviews

Privacy-preserving ML, MLOPS

Frameworks: MITRE Atlas, NIST AI Risk Assessment, Amazon Well-architected: ML lens

Privacy considerations

- Most current privacy laws require the consumer's consent for data usage (e.g. training ML models).
- Opt-out support.

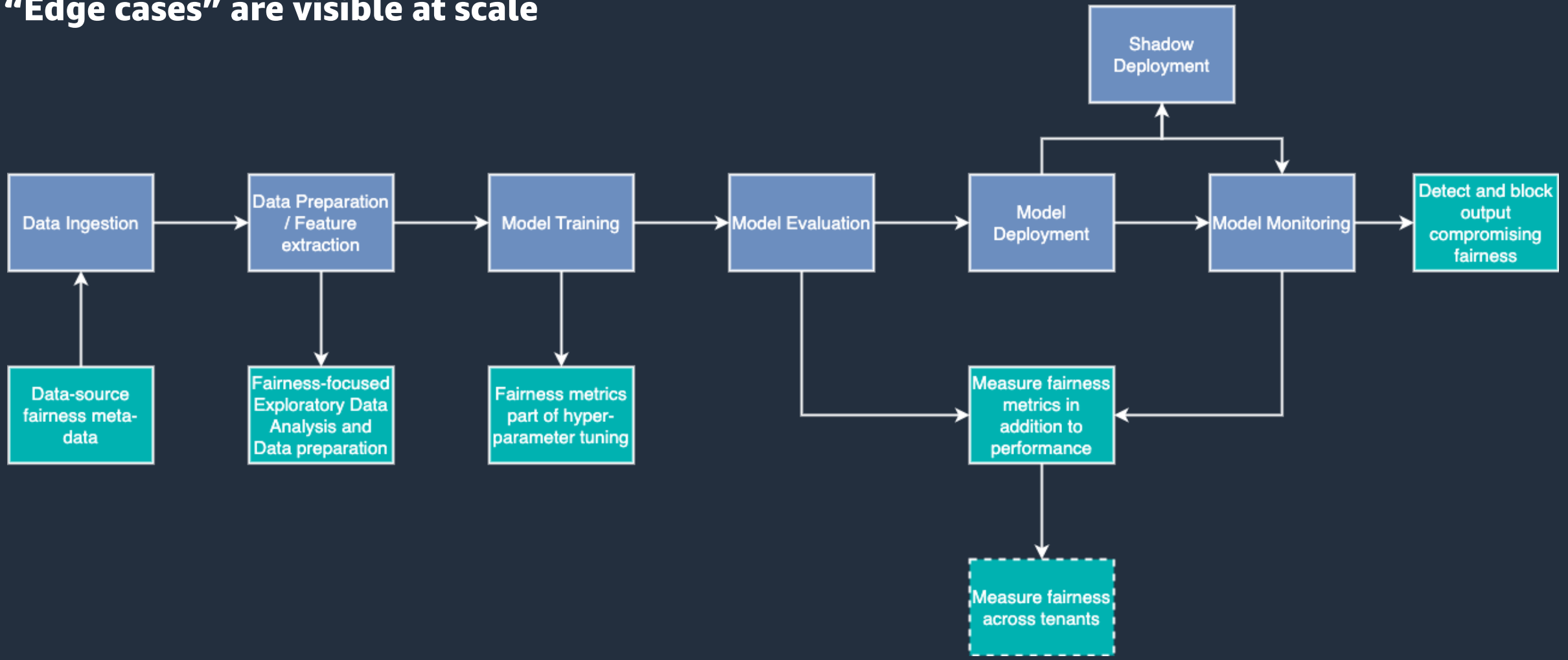
- GDPR lays out 6 valid reasons that personal data can be collected and used: consumer consent, contractual obligation, legal obligation, public interest, vital interest, or legitimate business interests.

- Laws such as HIPAA and GDPR include data anonymization requirements for ML training data.
- If you're working with personal or sensitive data, it should be anonymized.
- Consider using **synthetic data**.

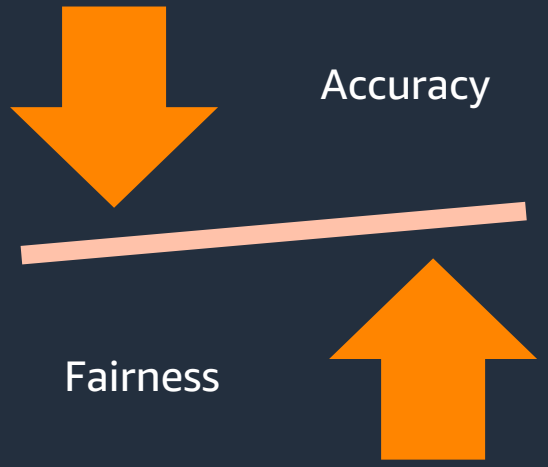
- Sensitive consumer data often comes with conditions on how long it must be stored or when it must be destroyed.

Fairness considerations

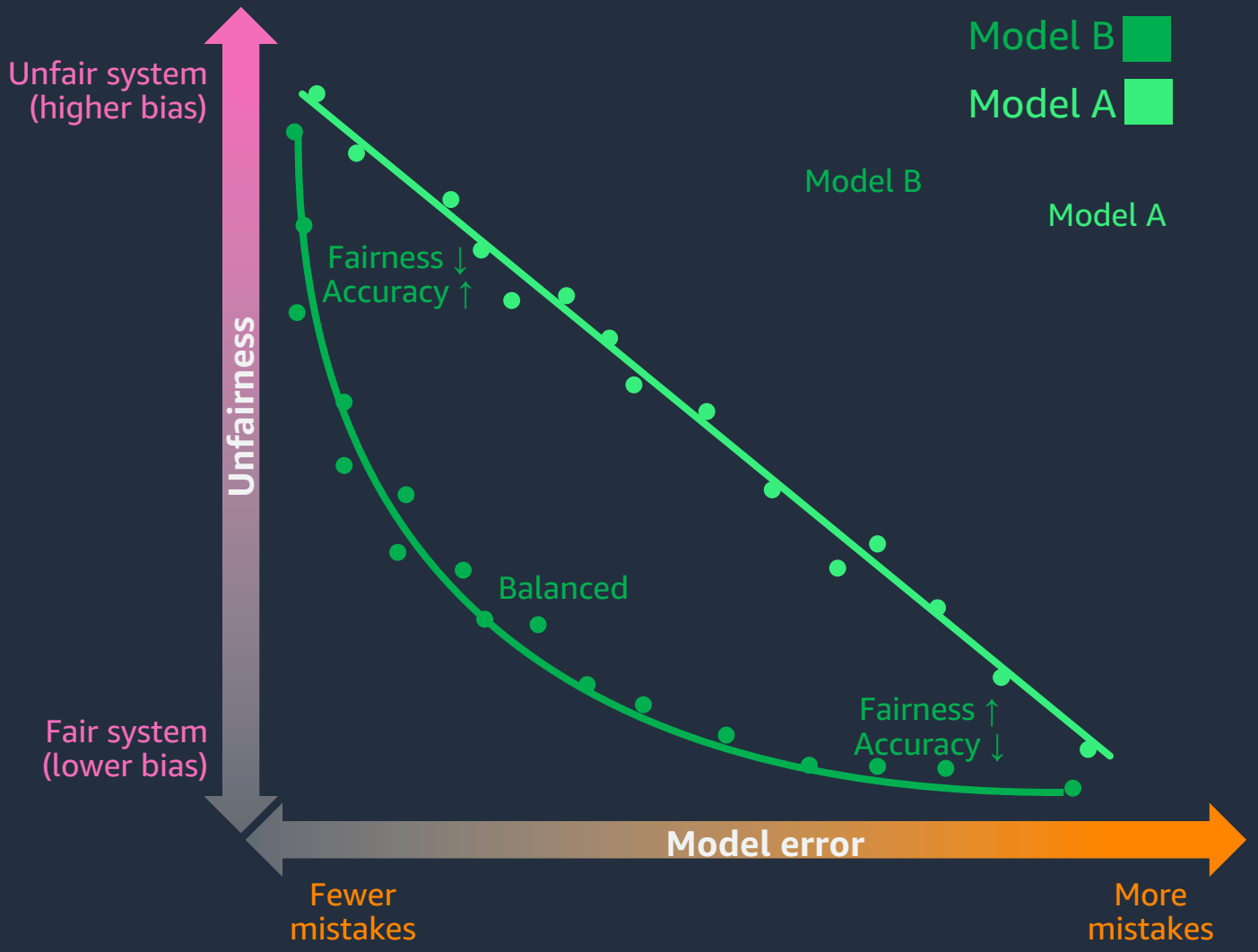
“Edge cases” are visible at scale



Fairness vs accuracy



- In many cases fairness forms a **trade-off** with accuracy
- The decision on where to operate on this trade-off is a **strategic choice**
- Pareto frontier graphs (between model error and some measure of unfairness), can help identify better ML models and trade-off operating points



Pareto frontiers of model error vs unfairness

An Empirical Study of Rich SubgroupFairness for Machine Learning: https://arxiv.org/pdf/1808.08166.pdf

Training data challenges

Difficulty to source sensitive data

Difficulty to share sensitive data

Data for specific use-cases hard or uncomfortable to get

Datasets are often unbalanced

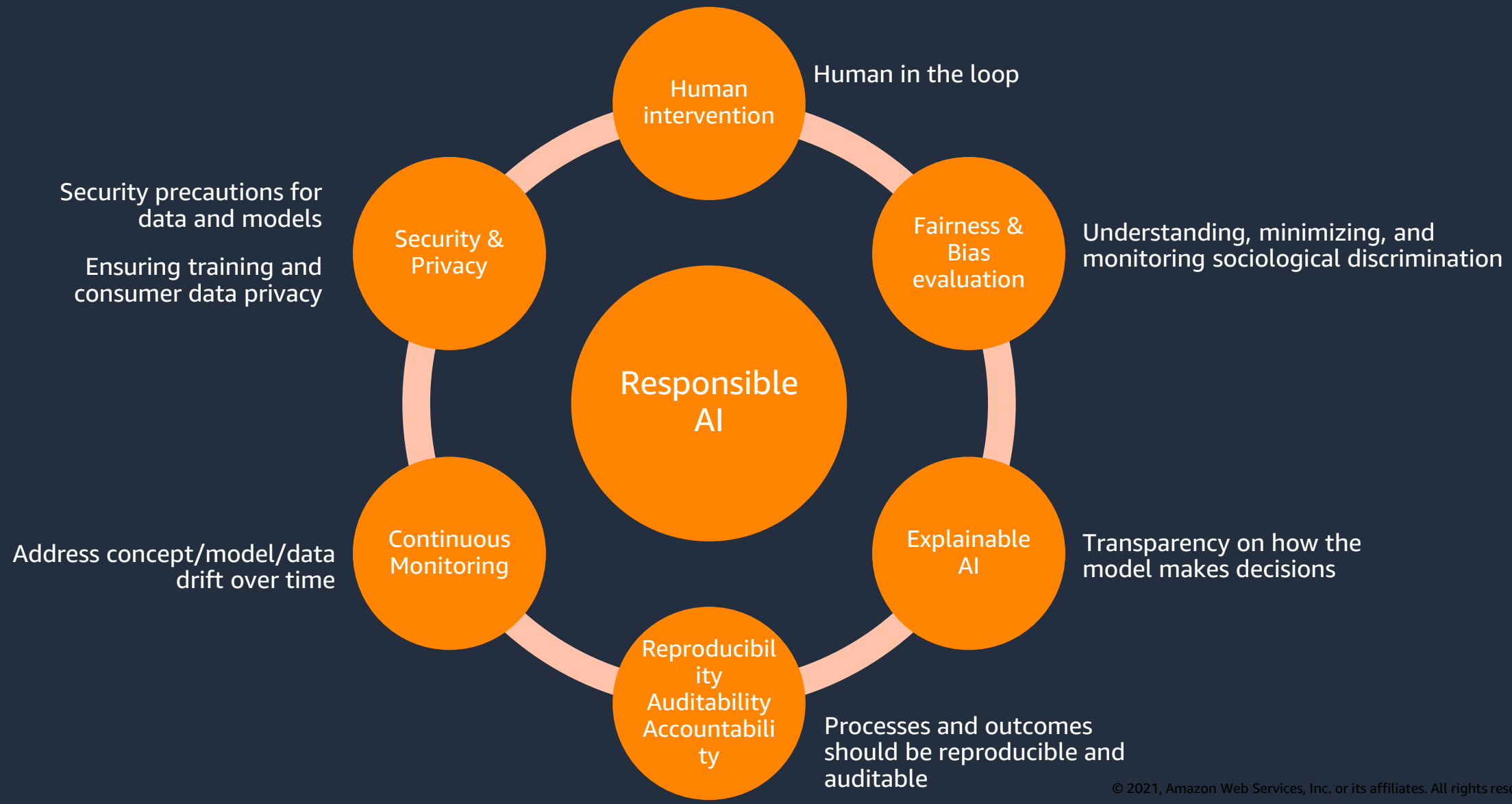
Source: classify, filter, encrypt, anonymize / tokenize / coarsen data and be aware of indirect identifiers

Secure collaboration: Amazon DataZone, AWS Clean Rooms

Marketplaces: AWS Data Exchange

Synthetic data: Generative AI, Amazon Bedrock, Amazon Titan

Components of Responsible AI



Responsible AI in AWS