Managing AI risks: Challenges & Solutions (DASF)

Omar Khawaja

September 2024

Traditional Programs vs. ML models

Traditional Programs

- Conditions and attributes
- Code and builds
- Deterministic
- Bugs
- Black/white
- Decisions system
- Human instructions
- Yes/no

Machine Learning

- Features & labels
- Probabilistic
- Biases & hallucinations or just bugs?
- Prediction systems
- Data as instructions
- Spectrum
- Hyper parameters
- Models
- Some human instructions
- Human in the loop
- Open ended

AI ≠ traditional computer applications

Don’t overestimate AI

“The illiterate of the twenty-first century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn.”

—Alvin Toffler

Fully autonomous vehicles could reduce traffic fatalities by up to 94%..

[US Dept of Transportation]

AI doesn't stop learning!

Generative AI is taking the world by storm

91%

of organizations are experimenting with or investing in GenAI[1]

75%

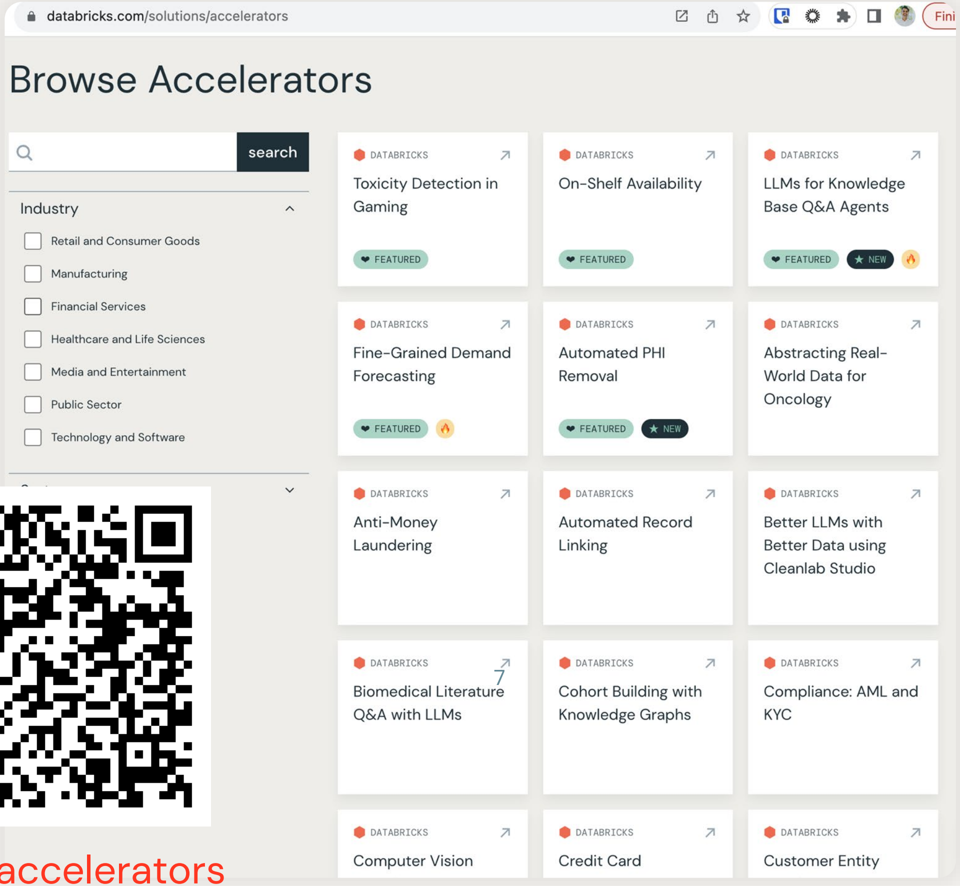
of CEOs say companies with advanced GenAI will have a competitive advantage[2]

40%

increase in performance of employees who used GenAI[3]

1. Laying the foundation for data and AI-led growth, MIT Technology Review
2. CEO decision-making in the age of AI, IBM Institute for Business Value
3. How generative AI can boost highly skilled workers' productivity, MIT Management Sloan School

82 ways organizations across 7 industries are using Data+AI



https://www.databricks.com/solutions/accelerators

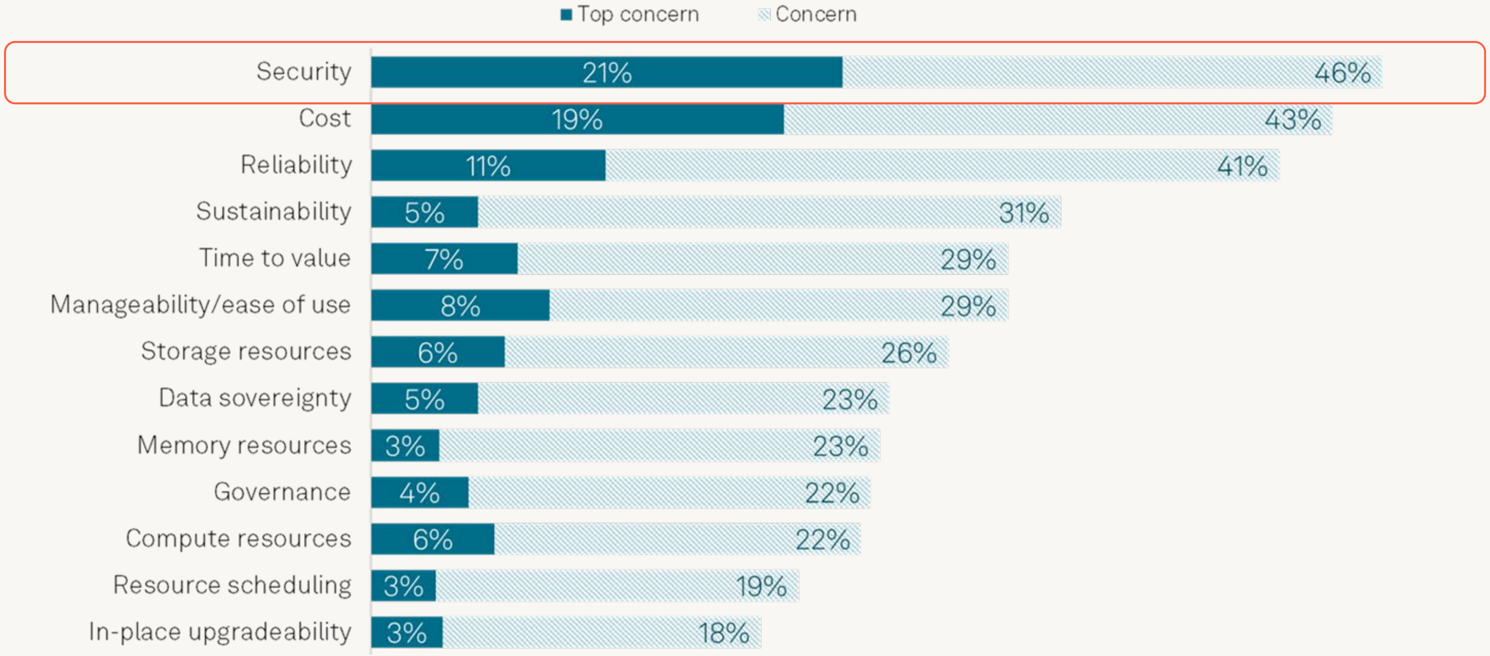
Challenge:

Building and deploying production-quality Gen AI solutions

90%

of enterprises *not* confident going to production

Security is the top concern for AI adoption



Q. What are your organization's main concerns about the infrastructure that [hosts/will host] its AI/ML workloads? Please select all that apply; Base: All respondents (n=712).

Q. And which is your organization's top concern about the infrastructure that [hosts/will host] its AI/ML workloads? Base: Organization has concerns about the infrastructure that [hosts/will host] its AI/ML workloads (n=683).

Source: 451 Research's Voice of the Enterprise: AI & Machine Learning, Infrastructure 2023.

Today, GenAI in production is difficult and expensive

No control over the data or the models

- Concern over data leakage
- Lack of control and ownership

Bringing GenAI to production is difficult

- Unpredictable performance
- Need automation and scale

Too expensive at scale

- Foundation models are expensive at scale
- Expensive to build LLMs

How do we manage risks w/ traditional tech?

As risk leaders, we have honed various risk management skills over decades

1. **Tech**: mental model of components and data flows
2. **People & Process**: defined roles and operating model
3. **Risks (all)**: knowledge of harms that can be caused
4. **Architecture**: proficiency in various deployment models and their risk implications
5. **Threats**: known classes of threats to be considered
6. **Risks (contextual)**: for specific use case, conduct risk analysis to identify specific risks worth mitigating
7. **Controls**: well known set of controls, where to implement them and their efficacy in mitigating risks

A: Leverage ***risk instincts*** to identify appropriate controls

Why is it *hard* to manage AI risks?

As risk leaders, we have not yet built confidence in our ability to manage AI risks

1. **Tech**: missing mental model of complete AI components

2. **People & Process**: unsure of roles and operating model

3. **AI Risks (all)**: missing comprehensive AI risks catalog

4. **Architecture**: unaware of security implications of various AI deployment models

5. **Threats**: unclear which AI threats to be concerned with

6. **AI Risks (contextual)**: unsure which particular risks to focus on mitigating

7. **Controls**: unsure which controls to apply and where to apply them

A: Because AI still feels novel and our typical ***risk instincts*** haven't been activated yet

How do we make it *easy* to manage AI risks?

As risk leaders, we have not yet built confidence in our ability to manage AI risks

1. **Tech**: define mental model of AI components
2. **People & Process**: define roles and operating model
3. **AI Risks (all)**: enumerate comprehensive AI risks
4. **Architecture**: define AI deployment models
5. **Threats**: map AI risks to AI threats
6. **AI Risks (contextual)**: filter AI risks based on use case and threat model
7. **Controls**: map each AI risk to mitigating controls and AI component

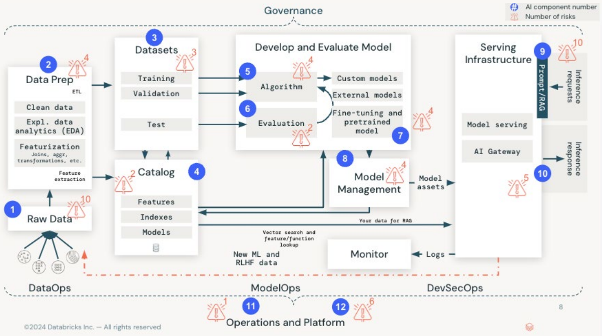
A: Activate ***instincts*** to manage AI risks!

How do we make it *easy* to manage AI risks?

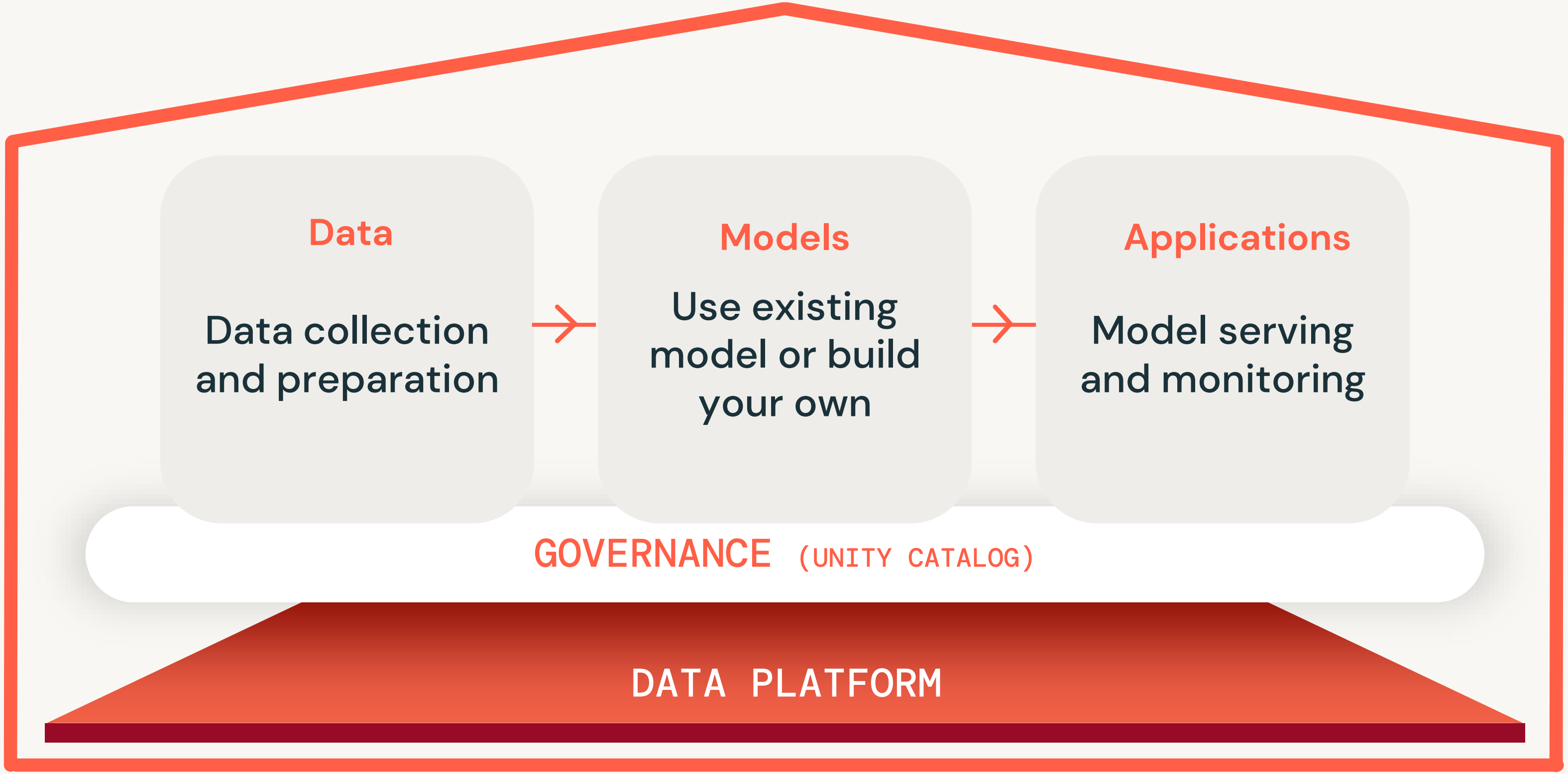
As risk leaders, we have not yet built confidence in our abilities manage AI risks

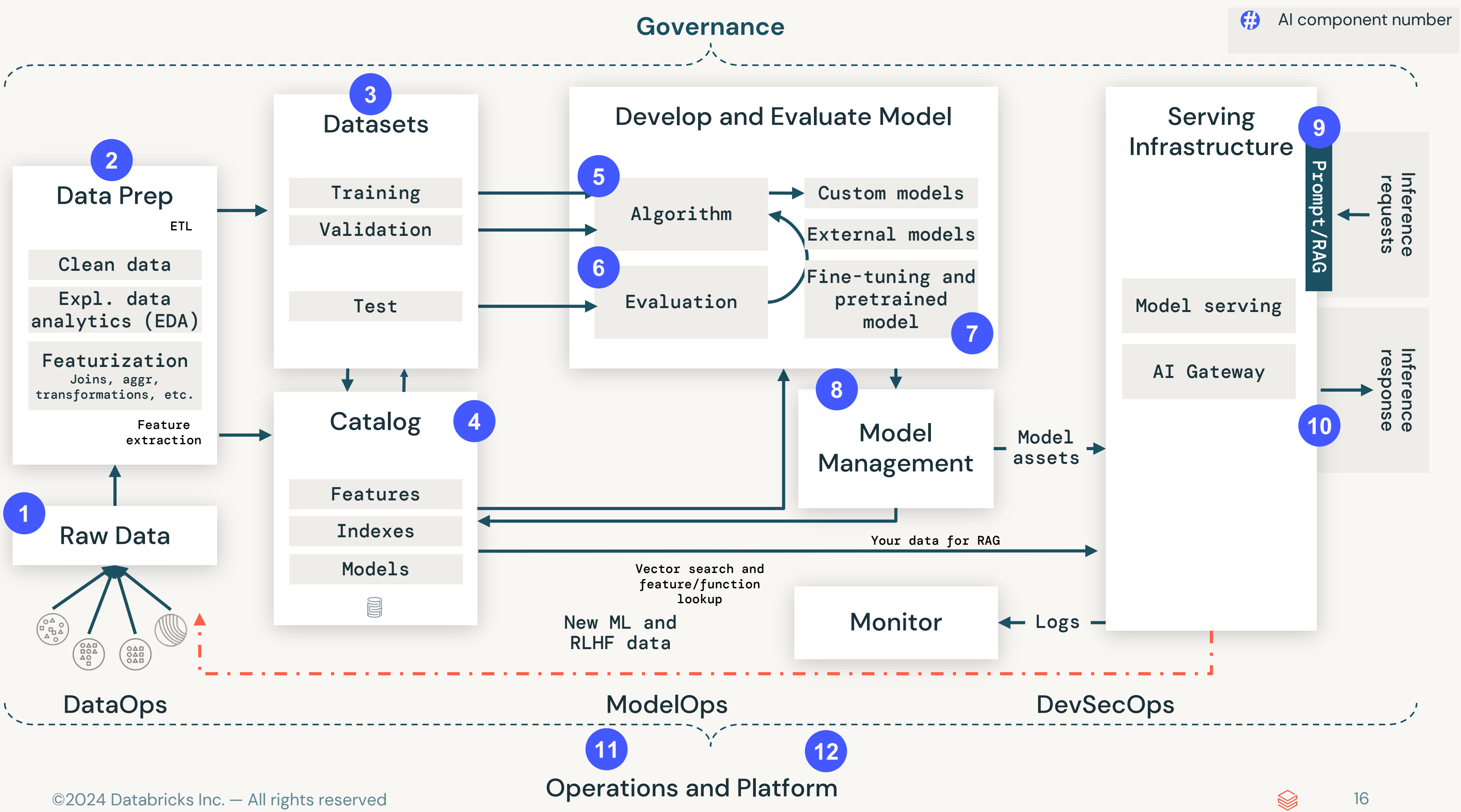
1. **Tech**: define mental model of AI components
2. **People & Process**: define roles and operating model
3. **AI Risks (all)**: enumerate comprehensive AI risks
4. **Architecture**: define AI deployment models
5. **Threats**: map AI risks to AI threats
6. **AI Risks (contextual)**: filter AI risks based on use case and threat model
7. **Controls**: map each AI risk to mitigating controls and AI component

12x components of end-end AI system



What subsystems make up an AI system?





AI component number
Governance
Data Prep
ETL
Clean data
Expl. data analytics (EDA)
Featurization
Joins, aggr, transformations, etc.
Feature extraction
Raw Data
Datasets
Training
Validation
Test
Catalog
Features
Indexes
Models
Develop and Evaluate Model
Algorithm
Evaluation
Custom models
External models
Fine-tuning and pretrained model
Model Management
Model assets
Serving Infrastructure
Prompt/RAG
Model serving
AI Gateway
Inference requests
Inference response
Your data for RAG
Vector search and feature/function lookup
New ML and RLHF data
Monitor
Logs
DataOps
ModelOps
DevSecOps
Operations and Platform
1
2
3
4
5
6
7
8
9
10
11
12

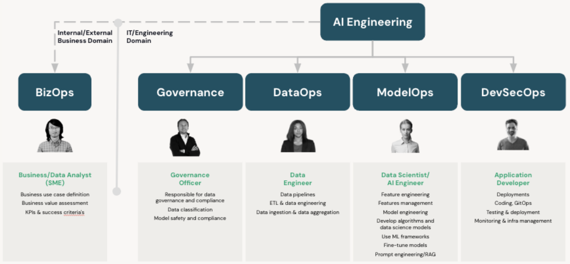
How do we make it *easy* to manage AI risks?

As risk leaders, we have not yet built confidence in our abilities manage AI risks

1. **Tech**: define mental model of AI components
2. **People & Process**: define roles and operating model
3. **AI Risks (all)**: enumerate comprehensive AI risks
4. **Architecture**: define AI deployment models
5. **Threats**: map AI risks to AI threats
6. **AI Risks (contextual)**: filter AI risks based on use case and threat model
7. **Controls**: map each AI risk to mitigating controls and AI component

Define roles across 3 subsystems of AI

People and process



How do we make it *easy* to manage AI risks?

As risk leaders, we have not yet built confidence in our abilities manage AI risks

1. **Tech**: define mental model of AI components
2. **People & Process**: define roles and operating model
3. **AI Risks (all)**: enumerate comprehensive AI risks
4. **Architecture**: define AI deployment models
5. **Threats**: map AI risks to AI threats
6. **AI Risks (contextual)**: filter AI risks based on use case and threat model
7. **Controls**: map each AI risk to mitigating controls and AI component

Catalog of 55x AI System risks across 12x components

Raw data
- 1.1: Insufficient access controls
- 1.2: Missing data classification
- 1.3: Poor data quality
- 1.4: In effective storage and encryption
- 1.5: Lack of data versioning
- 1.6: Insufficient data lineage
- 1.7: Lack of data trustworthiness
- 1.8: Data legal
- 1.9: Stale data
- 1.10: Lack of data access logs

Algorithms
- 5.1: Lack of tracking and reproducibility of experiments
- 5.2: Model drift
- 5.3: Hyperparameters stealing
- 5.4: Malicious Libraries

Data Prep
- 2.1: Preprocessing Integrity
- 2.2: Feature manipulation
- 2.3: Raw data criteria
- 2.4: Adversarial partitions

Datasets
- 3.1: Data poisoning
- 3.2: In effective storage and encryption
- 3.3: Label Flipping

Evaluation
- 6.1: Evaluation data poisoning
- 6.2: Insufficient evaluation data

Model
- 7.1: Backdoor Machine Learning / Trojaned model
- 7.2: Model assets leak
- 7.3: ML Supply chain vulnerabilities
- 7.4: Source code control attack

Governance
- 4.1: Lack of traceability and transparency of model assets
- 4.2: Lack of end-to-end ML lifecycle

Model Management
- 8.1: Model attribution
- 8.2: Model theft
- 8.3: Model lifecycle without HITL
- 8.4: Model inversion

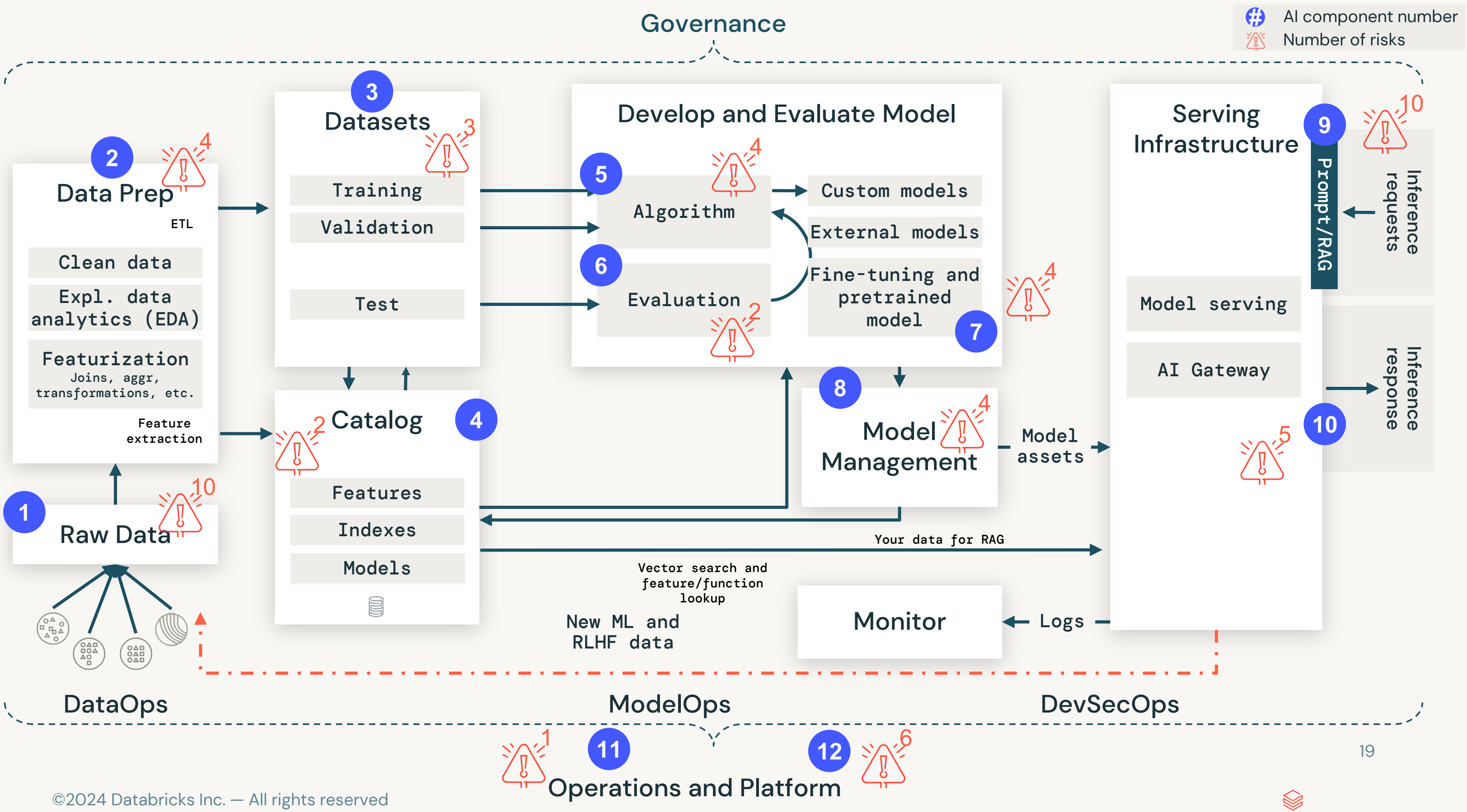
Model Serving – Inf response
- 10.1: Lack of audit and monitoring inference quality
- 10.2: Output manipulation
- 10.3: Discover ML Model Ontology
- 10.4: Discover ML Model Family
- 10.5: Black box attacks

Operations
- 11.1: Lack of MLOps – repeatable enforced standards

Model Serving – Inf requests
- 9.1: Prompt inject
- 9.2: Model inversion
- 9.3: Model breakout
- 9.4: Looped input
- 9.5: Infer training data membership
- 9.6: Discover ML Model Ontology
- 9.7: Denial of Service
- 9.8: LLM hallucinations
- 9.9: Input Resource Control

Platform
- 12.1: Lack of vulnerability management
- 12.2: Lack of penetration testing and bug bounty
- 12.3: Lack of Incident response
- 12.4: Unauthorized privileged access
- 12.5: Poor SDLC
- 12.6: Lack of compliance

Risks is red indicate novel risks for AI


Governance
AI component number
Number of risks
1
Raw Data
10
2
Data Prep
4
ETL
Clean data
Expl. data analytics (EDA)
Featurization
Joins, aggr., transformations, etc.
Feature extraction
3
Datasets
3
Training
Validation
Test
4
Catalog
2
Features
Indexes
Models
Develop and Evaluate Model
5
Algorithm
4
6
Evaluation
2
Custom models
External models
Fine-tuning and pretrained model
7
4
8
Model Management
4
Model assets
Serving Infrastructure
9
Prompt/RAG
10
Inference requests
Model serving
AI Gateway
10
5
Inference response
Your data for RAG
Vector search and feature/function lookup
New ML and RLHF data
Monitor
Logs
DataOps
ModelOps
DevSecOps
1
11
12
6
Operations and Platform

55 risks across 12 components of AI (20 traditional , 35 novel)

databricks

1 Raw data
- 1.1: Insufficient access controls
- 1.2: Missing data classification
- 1.3: Poor data quality
- 1.4: In effective storage and encryption
- 1.5: Lack of data versioning
- 1.6: Insufficient data lineage
- 1.7: Lack of data trustworthiness
- 1.8: Data legal
- 1.9: Stale data
- 1.10: Lack of data access

2 Data Prep
- 2.1: Preprocessing Integrity
- 2.2: Feature manipulation
- 2.3: Raw data criteria
- 2.4: Adversarial partitions

3 Datasets
- 3.1: Data poisoning
- 3.2: Ineffective storage and encryption
- 3.3: Label Flipping

4 Governance
- 4.1: Lack of traceability and transparency of model assets
- 4.2: Lack of end-to-end ML lifecycle

5 Algorithms
- 5.1: Lack of tracking and reproducibility of experiments
- 5.2: Model drift
- 5.3: Hyperparameters stealing
- 5.4: Malicious Libraries

6 Evaluation
- 6.1: Evaluation data poisoning
- 6.2: Insufficient evaluation data

7 Model
- 7.1: Backdoor Machine Learning / Trojaned model
- 7.2: Model assets leak
- 7.3: ML Supply chain vulnerabilities
- 7.4: Source code control attack

8 Model Management
- 8.1: Model attribution
- 8.2: Model theft
- 8.3: Model lifecycle without HITL
- 8.4: Model inversion

9 Model Serving – Inf requests
- 9.1: Prompt inject
- 9.2: Model inversion
- 9.3: Model breakout
- 9.4: Looped input
- 9.5: Infer training data membership
- 9.6: Discover ML Model Ontology
- 9.7: Denial of Service
- 9.8: LLM hallucinations
- 9.9: Input Resource Control
- 9.10: Accidental exposure of unauthorized data to models

10 Model Serving – Inf response
- 10.1: Lack of audit and monitoring inference quality
- 10.2: Output manipulation
- 10.3: Discover ML Model Ontology
- 10.4: Discover ML Model Family
- 10.5: Black box attacks

11 Operations
- 11.1: Lack of MLOps – repeatable enforced standards

12 Platform
- 12.1: Lack of vulnerability management
- 12.2: Lack of penetration testing and bug bounty
- 12.3: Lack of Incident response
- 12.4: Unauthorized privileged access
- 12.5: Poor SDLC
- 12.6: Lack of compliance

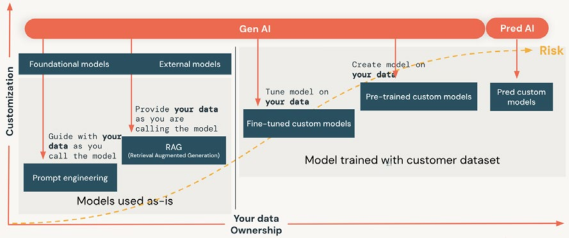
Red = Novel Risks

How do we make it *easy* to manage AI risks?

As risk leaders, we have not yet built confidence in our ability to manage AI risks

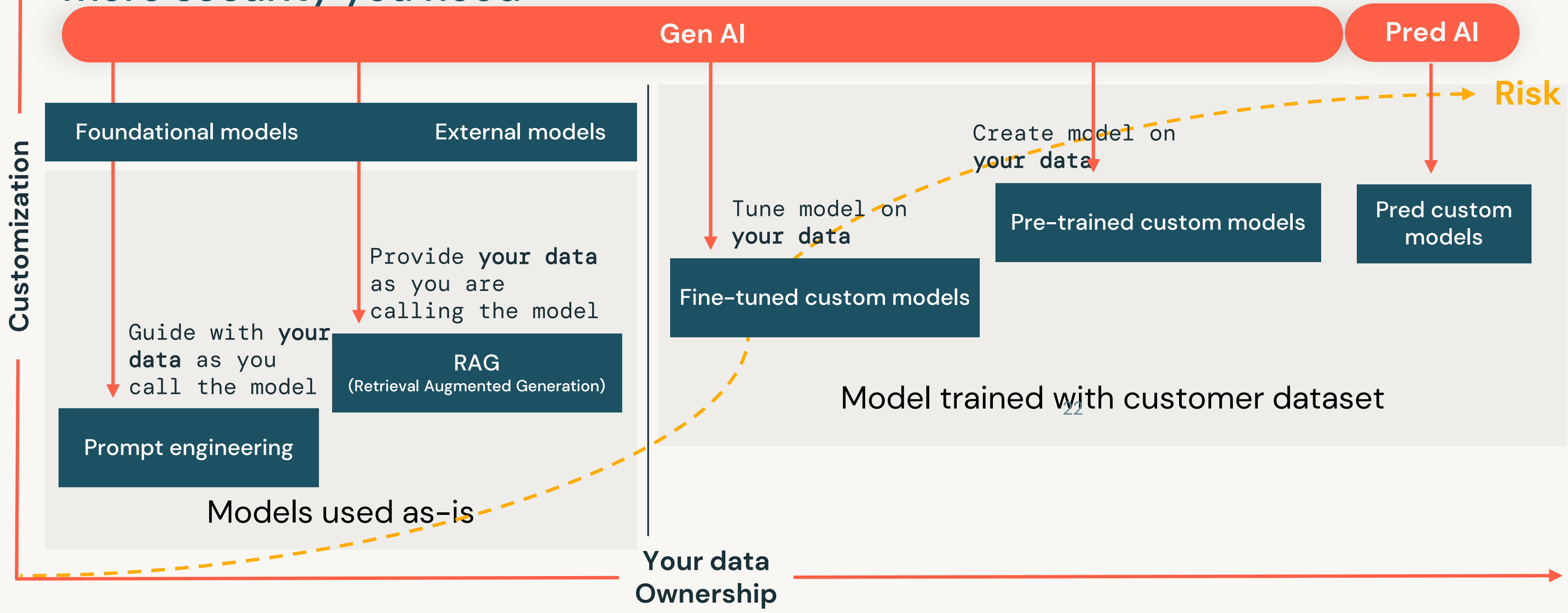
1. **Tech**: define mental model of AI components
2. **People & Process**: define roles and operating model
3. **AI Risks (all)**: enumerate comprehensive AI risks
4. **Architecture**: define AI deployment models
5. **Threats**: map AI risks to AI threats
6. **AI Risks (contextual)**: filter AI risks based on use case and threat model
7. **Controls**: map each AI risk to mitigating controls and AI component

Implications of shared responsibility across 6 AI deployment models



Customization of AI with your data

The more you customize models with your data, the more accurate and more security you need



How do we make it *easy* to manage AI risks?

As risk leaders, we have not yet built confidence in our ability to manage AI risks

1. **Tech**: define mental model of AI components

2. **People & Process**: define roles and operating model

3. **AI Risks (all)**: enumerate comprehensive AI risks

4. **Architecture**: define AI deployment models

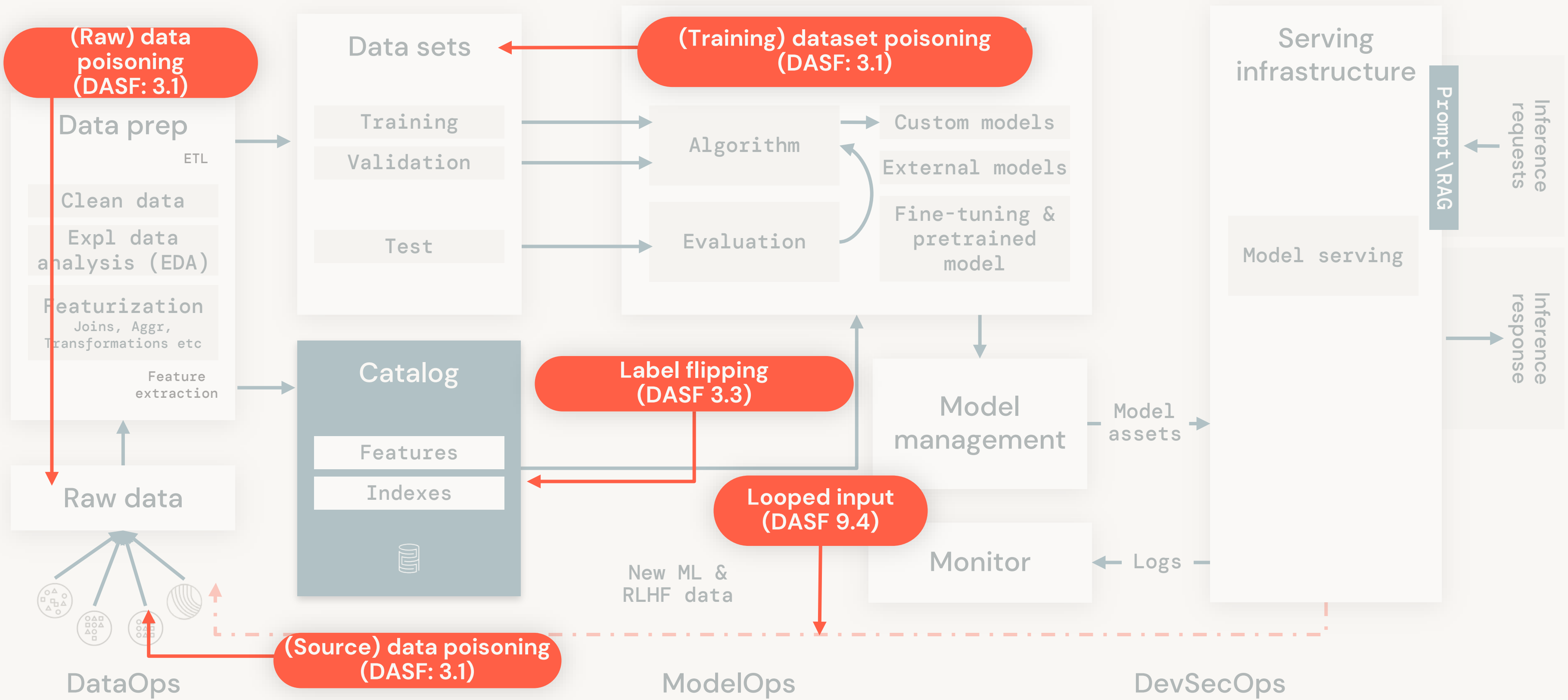
5. **Threats**: map AI risks to AI threats

6. **AI Risks (contextual)**: filter AI risks based on use case and threat model

7. **Controls**: map each AI risk to mitigating controls and AI component

Map 55 AI risks to Mitre ATLAS attack techniques

Ex.: Training Data Poisoning: *threats*



How do we make it *easy* to manage AI risks?

As risk leaders, we have not yet built confidence in our ability to manage AI risks

1. **Tech**: define mental model of AI components

2. **People & Process**: define roles and operating model

3. **AI Risks (all)**: enumerate comprehensive AI risks

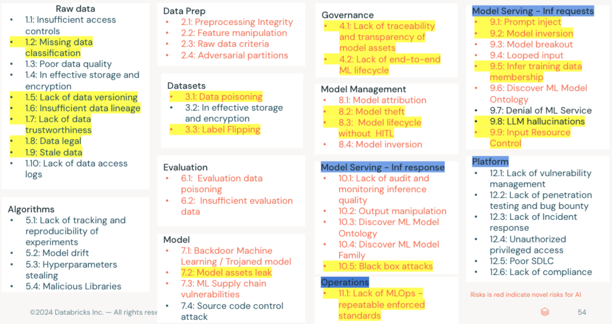
4. **Architecture**: define AI deployment models

5. **Threats**: map AI risks to AI threats

6. **AI Risks (contextual)**: filter AI risks based on use case and threat model

7. **Controls**: map each AI risk to mitigating controls and AI component

Select subset of 55 AI risks that are most pertinent



How do we make it *easy* to manage AI risks?

As risk leaders, we have not yet built confidence in our ability to manage AI risks

1. **Tech**: define mental model of AI components
2. **People & Process**: define roles and operating model
3. **AI Risks (all)**: enumerate comprehensive AI risks
4. **Architecture**: define AI deployment models
5. **Threats**: map AI risks to AI threats
6. **AI Risks (contextual)**: filter AI risks based on use case and threat model
7. **Controls**: map each AI risk to mitigating controls and AI component

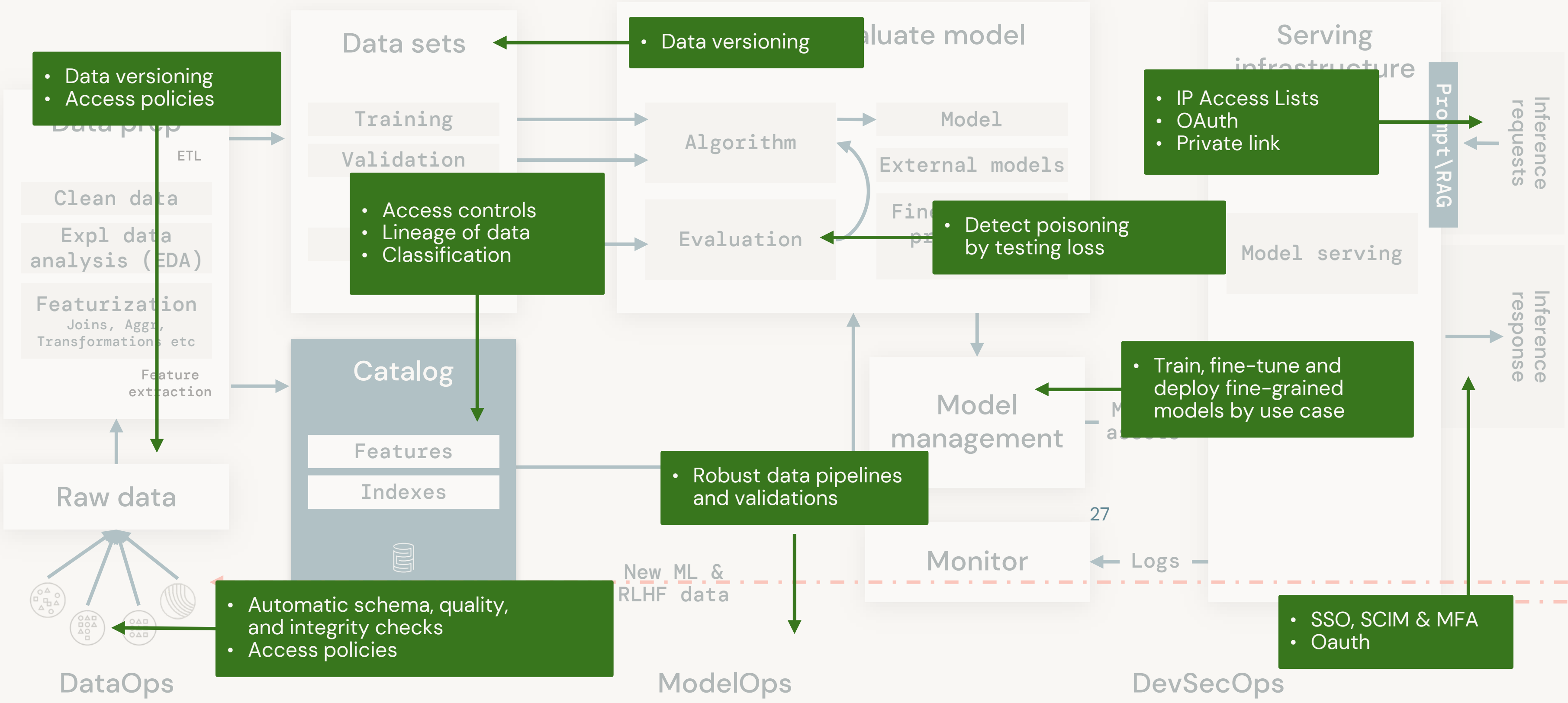
59 controls mapped to AI risks

Top 12 controls for mitigating AI risks

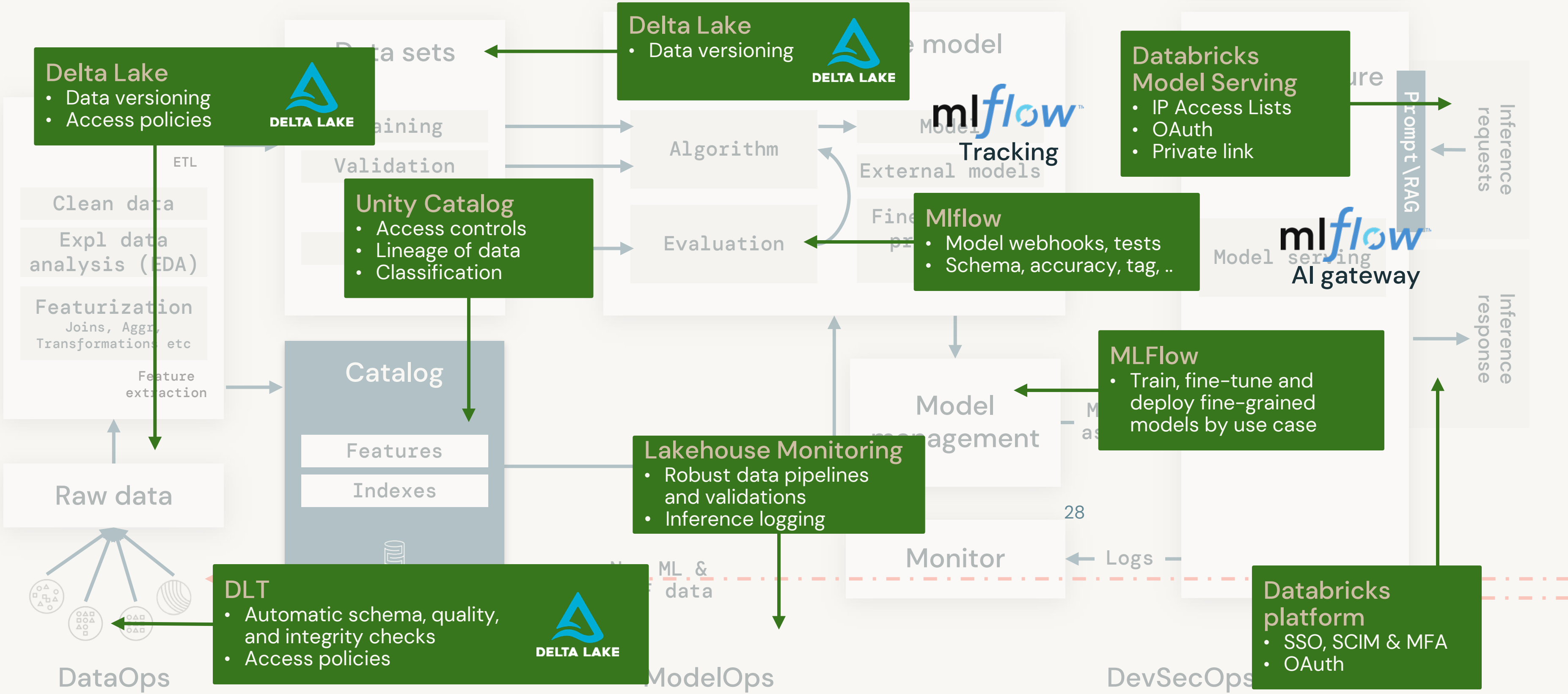
AI Novelty

Controls	Data poisoning	Prompt injection	Model theft	Reproducibility	Trustworthiness
Audit & monitor	●	●	●	◐	◐
Authentication and authorization	●	○	○	●	●
Data quality checks	●	○	○	●	●
Data governance	●	○	○	●	●
Data encryption	●	○	○	●	●
Secure MLOps	●	○	○	●	●
Track model artifacts	○	◐	◐	●	●
Testing and detect loss after (re)training	◐	○	◕	●	●
Encrypt models and auth endpoints	○	●	●	○	◐
Zero trust/ML segregation	●	●	●	◐	●
MLOps with HITL	◐	◐	●	●	●
Secure with Model Gateway	○	●	●	◐	◐

Ex.: Training data poisoning: *mitigating controls*



Ex.: Training data poisoning: *Databricks controls*



Top 10 controls for mitigating AI risks

AI Novelty

Controls	Data poisoning	Prompt injection	Model theft	Trojaned model	Trustworthiness
Authentication and authorization	●	◐	●	◕	◐
Data and model encryption	●	○	●	○	●
Data governance	●	○	○	○	●
Model governance	○	◐	◐	◕	●
Secure MLOps	●	◐	◐	◕	●
Testing and detect loss after (re)training	◐	○	◕	●	●
Securely serve models	○	●	●	○	◐
Zero Trust/Model Segregation	○	○	◐	●	●
Secure with Model Gateway	○	●	●	◐	◐
Audit & monitor	●	●	●	◐	◐

Databricks AI Security Framework (DASF)

AI Business Use Case

- Datasets
- Stakeholders
- Compliance
- Applications

AI Deployment Models

- Predictive ML models
- Foundational APIs
- Fine-tuned LLMs
- Pre-trained LLMs
- RAG with LLMs
- External Models

1 Use case

2 Deployment Model

Select subset of DASF **risks**

3 <55 Risks

Select subset of DASF **controls**

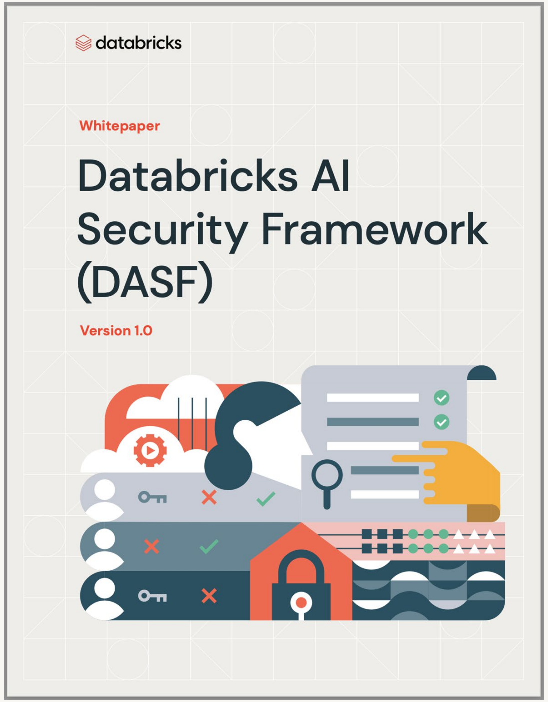
4 <59 Controls

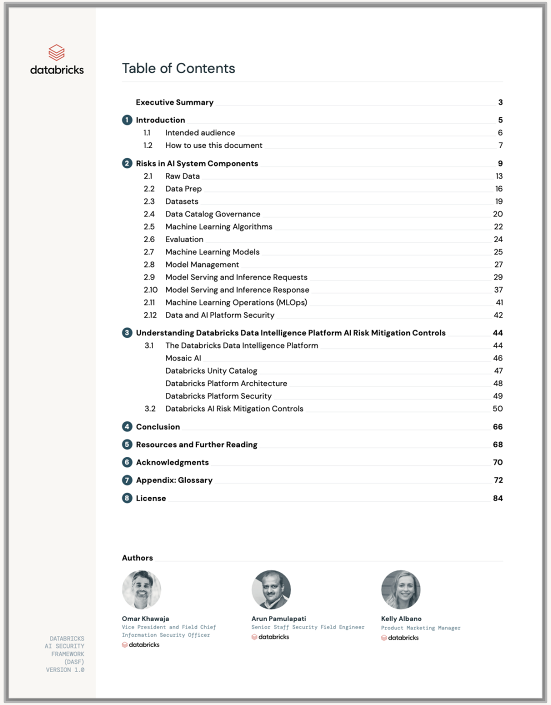
Implement controls on data platform across 12 AI components

Introducing the Databricks AI Security Framework!

- Securing AI will become easier as we better understand AI
- Each AI use case may have a distinct risk profile
- Be prepared to be wrong… adapt your process
- Adopt an open framework to hasten AI security, e.g.: **DASF**

How to get it?





databricks

Table of Contents

Authors

Omar Khawaja
Vice President and Field Chief Information Security Officer
databricks

Arun Pamulapati
Senior Staff Security Field Engineer
databricks

Kelly Albano
Product Marketing Manager
databricks

DATABRICKS AI SECURITY FRAMEWORK (DASF) VERSION 1.0

Strengthening AI security with industry luminaries, partners, and customers: Analyzed 12 authoritative papers, 16 contributors, and executed 15 external peer reviews!

Additional Slides

Top AI risk areas

ORGANIZATIONAL

- Talent
- Operating model
- Change management
- Decision support

- Did these risk areas exist pre-AI?
- Is there a risk bigger than the business not realizing outcomes from Data+AI?

Are you afraid of snakes?

We are wired to fear the unfamiliar