Challenges and Opportunities for AI Industrialization

Hui Lei, PhD, FIEEE
VP and CTO, Cloud and Big Data

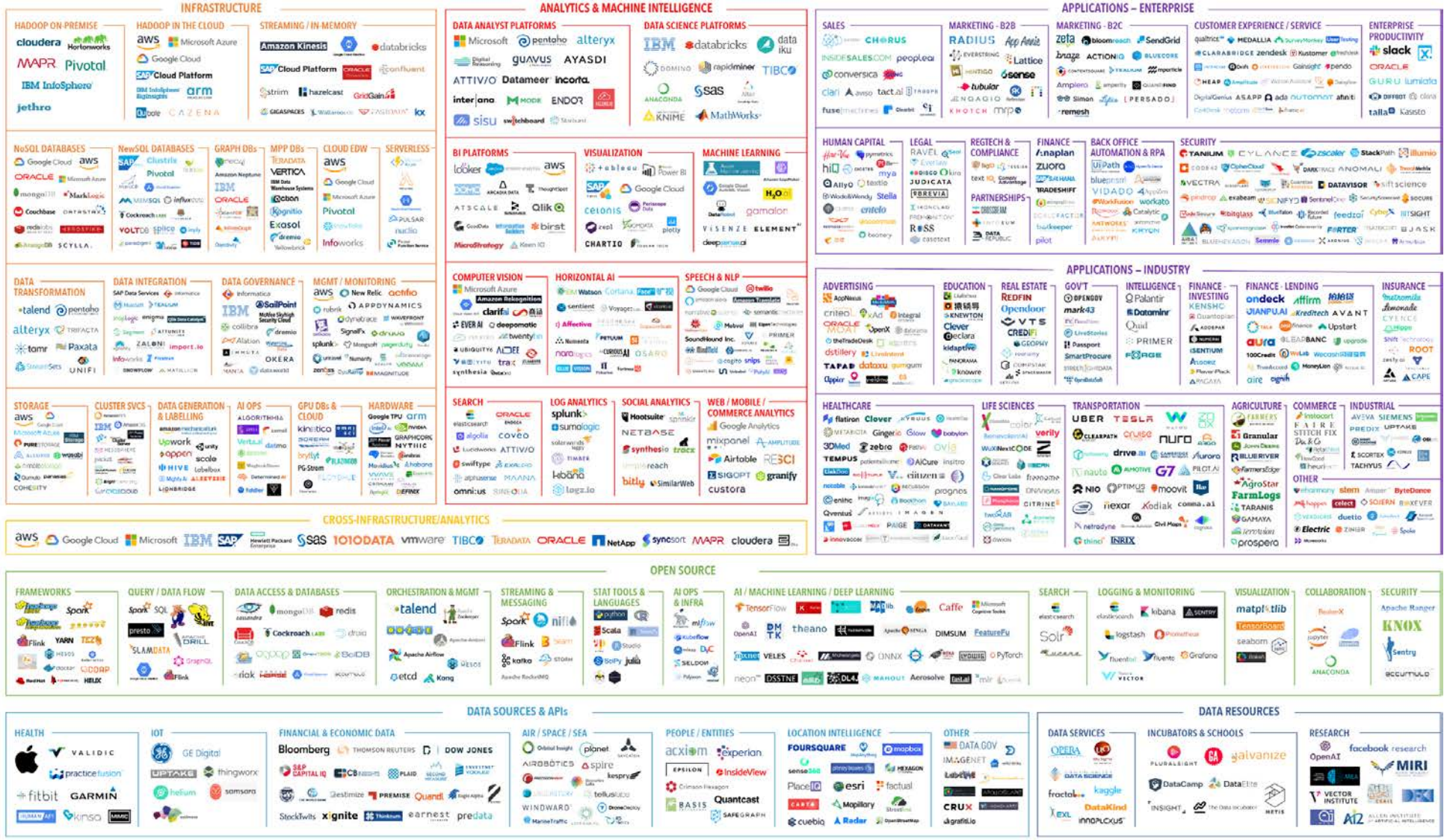
Futurewei Technologies

The Reality of AI Industrialization

- Less than 10% of AI pilot projects have reached full-scale production (IIA)
- Only 25% of companies have revenue-bearing AI projects in production (O’Reilly)
- By 2022, just 15% of projects for AI and IoT will be successful (Gartner)

Don’t We Have Many, Many Tools Already?

- AI tooling landscape is crowded and confusing
- Existing tools require deep expertise and skills
- Siloed tools result in overhead and technical debt



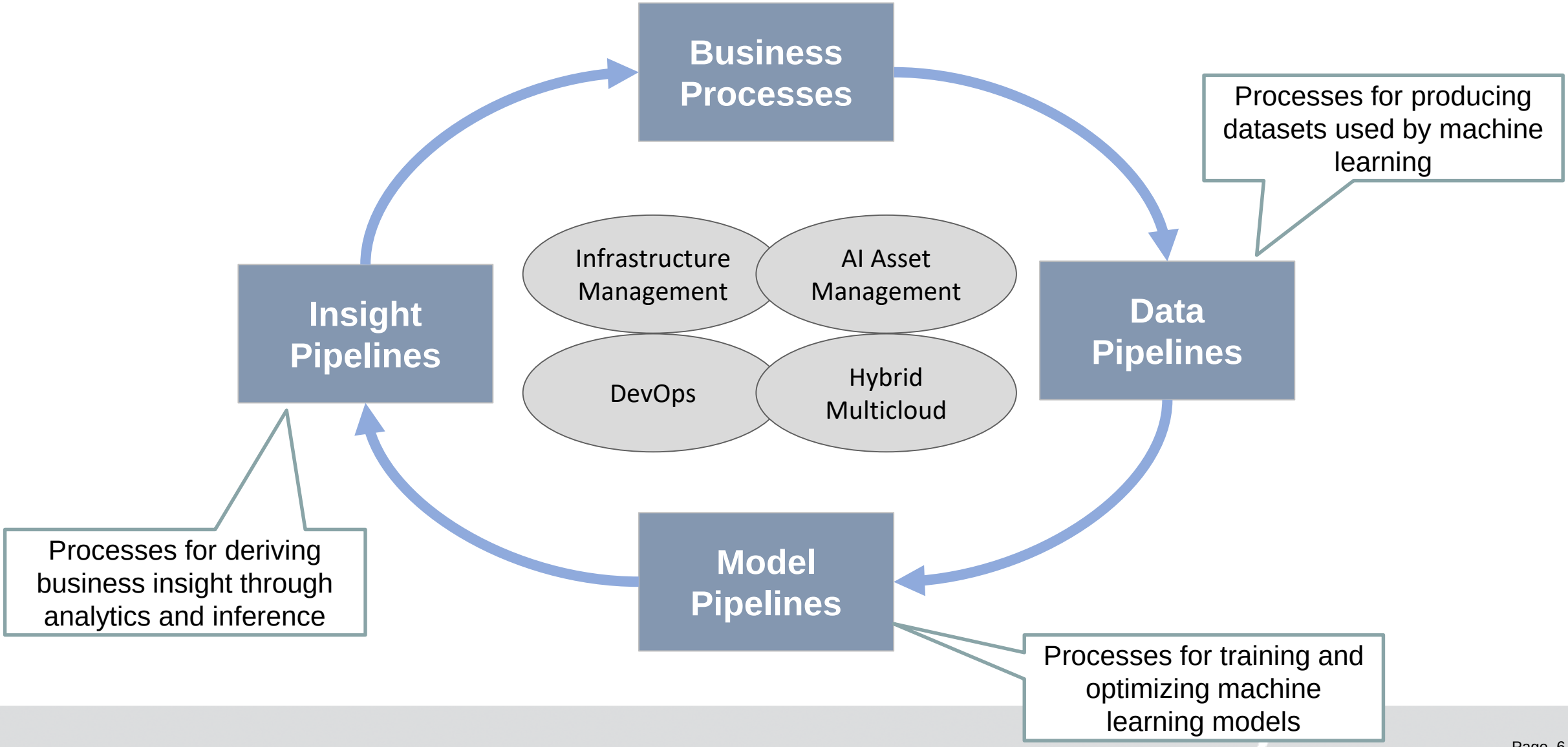
Source: https://mattturck.com/data2019

Existing ML Platforms Offer Partial Relief

- Commercial: Microsoft Azure ML Studo, Amazon SageMaker, IBM Watson Studio, Google TensorFlow Extended
- Open source: Kubeflow, MLflow
- Proprietary: Facebook FBLearner, Uber Michelangelo, AirBnb BigHead, LinkedIn Pro-ML

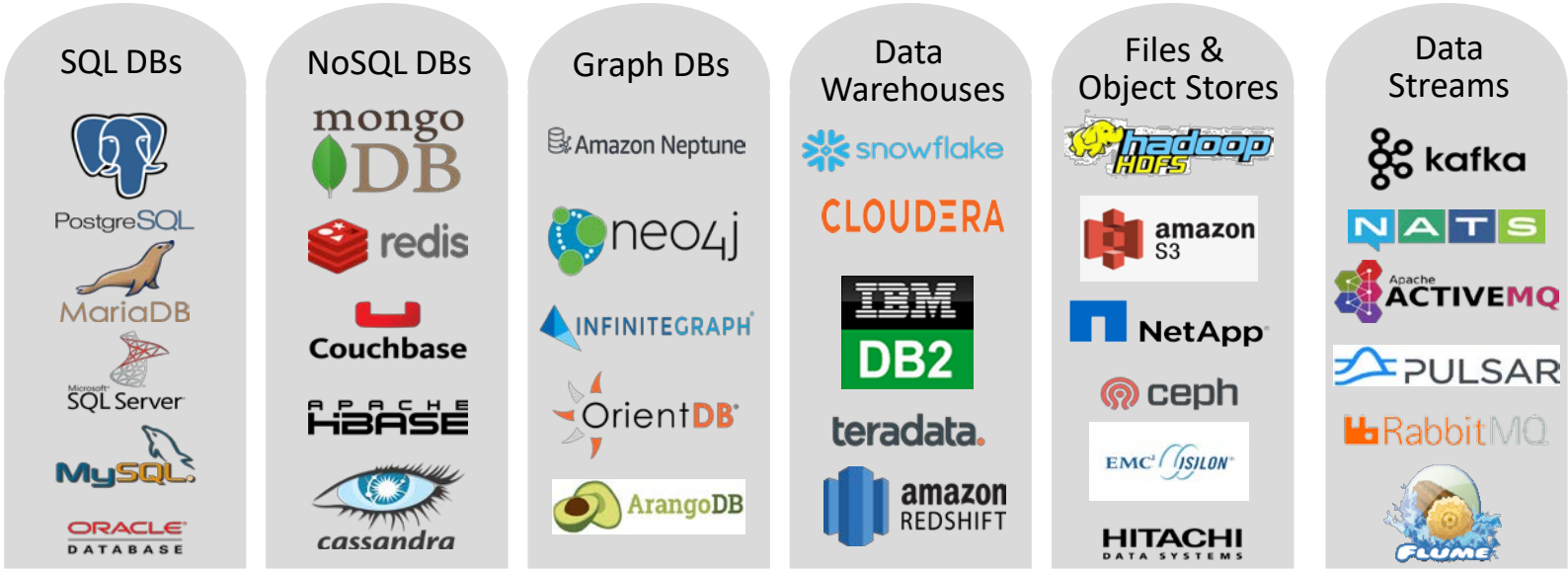
Challenges

Production AI Is More Than ML



Siloed, Distributed and Heterogeneous Data

Nearly 90% of companies report high or moderate degrees of data silos (CompTIA)



Existing Approaches

Technology	Benefits	Limitations
Data Lakes	Centralized storage of disparate data Computation close to data	Incomplete and disintegrated data Data duplication and staleness
Virtual distributed File Systems	Global namespace Local or memory caching for performance	File abstraction too primitive Unaware of higher-level semantics
Distributed SQL Engines	Declarative interface Masking data heterogeneity	Structured content only Pre-defined data schemas
Dataflow Engines	Flexible data analysis and processing Pushdown processing	Requirements on expertise and skills Data orchestration left with applications

Complexity in Insight Generation

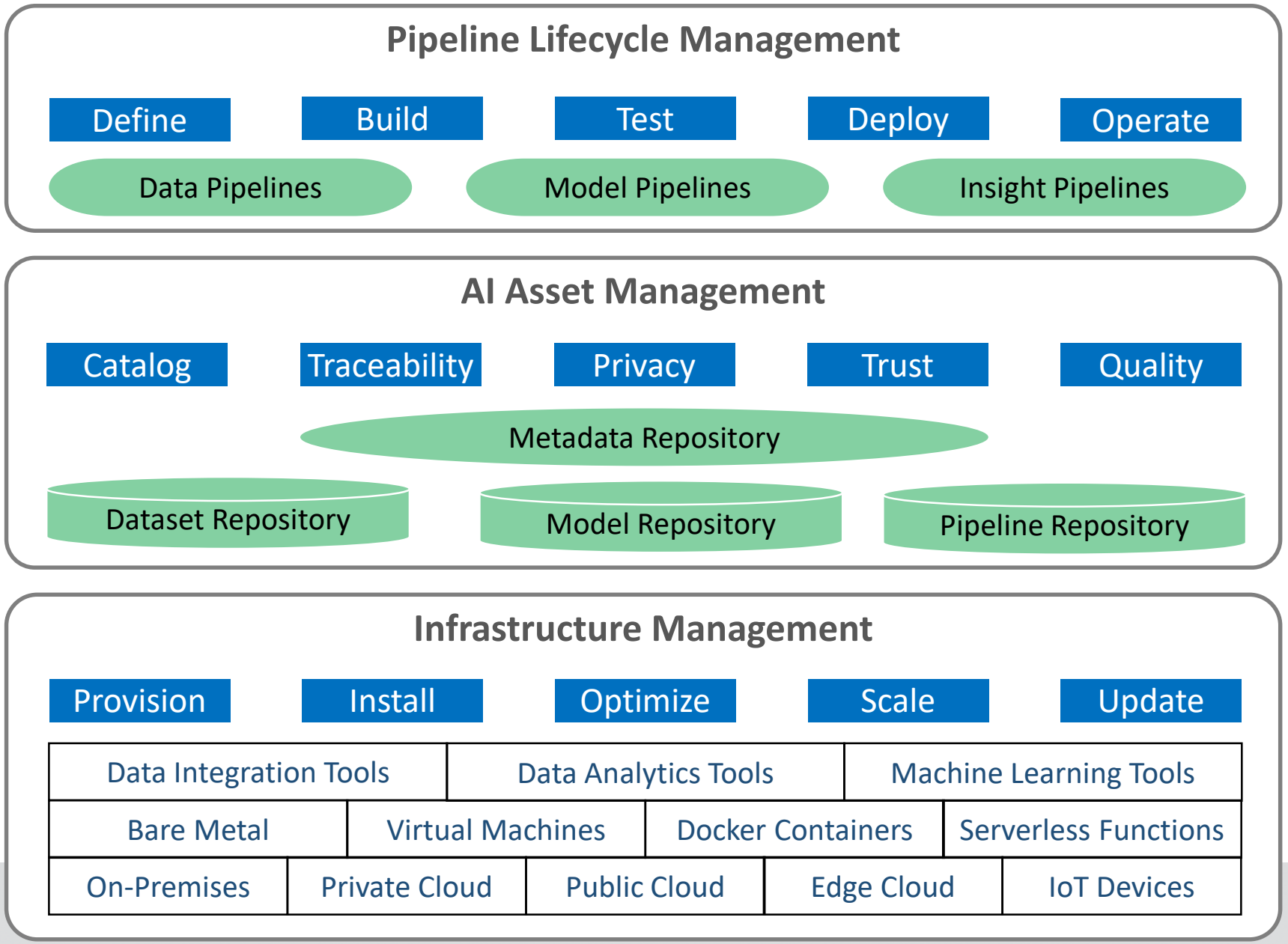
Data Collection | Data Processing | Model Composition | Model Hardening | Backend Integration | Data/Model Monitoring | Reporting & Dashboarding

Execution container selection
Resource selection and scheduling
Component replication and scaling
Data caching and prefetching
Request batching
Pipeline orchestration
Messaging and queuing
Straggler mitigation
Latency and throughput monitoring
Operational health monitoring

Data/model lineage and provenance
Data quality and outlier management
Data drift detection and mitigation
Concept drift detection and mitigation
Adversarial attack detection and mitigation
Prediction bias detection and mitigation
Data and model explanation
Data subject rights management
Metadata management
Model re-training

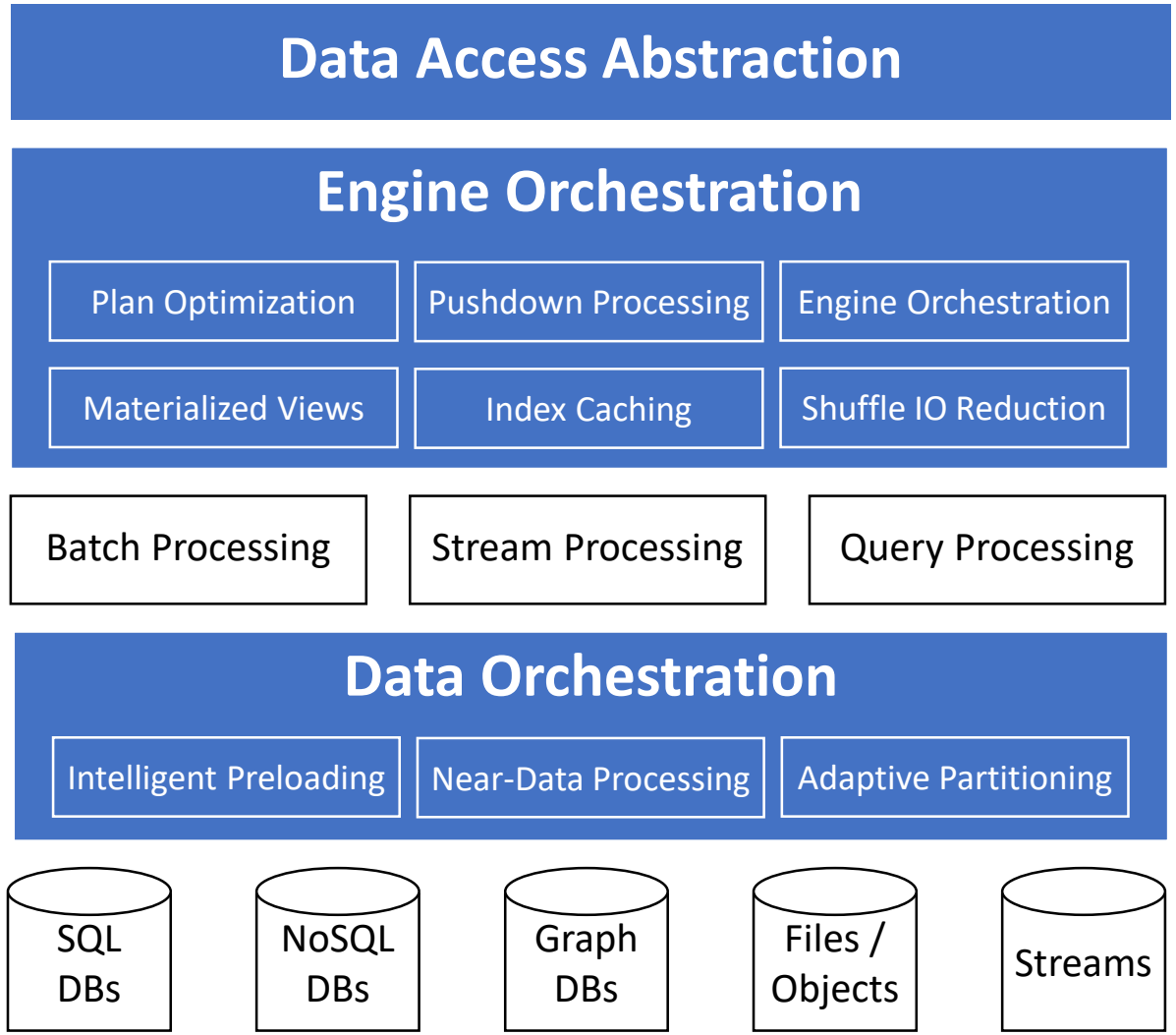
Innovation Opportunities

Integral AI Development



- Integrated and simplified development of all AI pipelines
- Unified catalog and governance for all AI assets
- Automated management of hardware/software infrastructure

Data Virtualization



Unified declarative and procedural interfaces

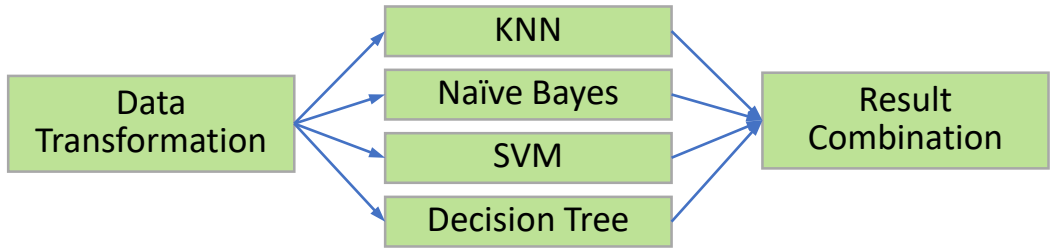
Optimization across data processing engines

Geographically distributed analytics capabilities

Moving data and compute closer together

Siloed and heterogeneous data storage

Composable Insight



Classification

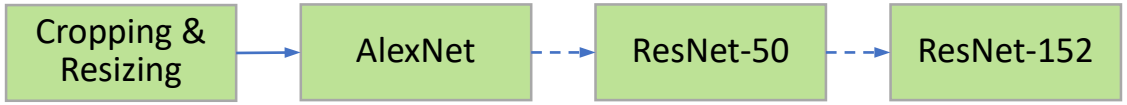
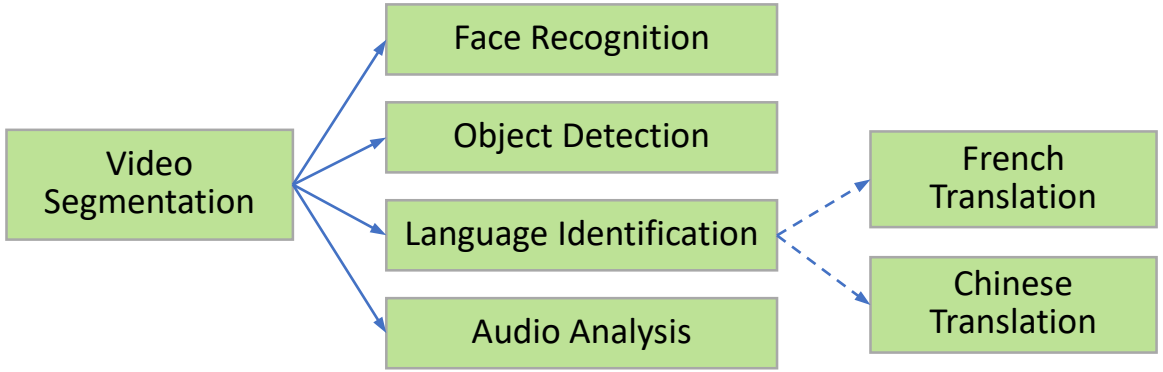
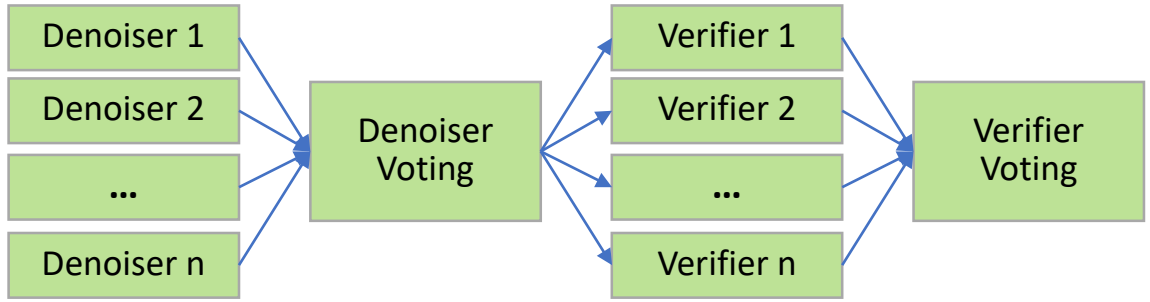


Image Processing



Video Analysis



Defense Against Adversarial Attacks

De-biasing

Assembling insight pipelines from reusable components

- Accuracy
- Throughput
- Simplicity
- Cost efficiency
- Robustness
- Fairness

Summary

Challenge	Innovation Opportunity
Production AI is more than ML	Integral AI development
Siloed, distributed and heterogeneous data	Data virtualization
Complexity in insight generation	Composable insight