AI-RAN and Open Source: Building Scalable Dataset Pipelines for AI Training

Alex Jinsung Choi
AI-RAN Alliance
SoftBank Corp.

Mar 31st 2025

Shaping the Future of Generative AI

84% of organizations have moderate, high, or very high **adoption of GenAI.**

For **92% of surveyed companies, GenAI is important**, and 51% consider it extremely important.

41% of GenAI infrastructure code **is open source.**

For 71% of organizations, the **open source nature of a model / tool has a positive influence** on its adoption, due to transparency and cost efficiency.

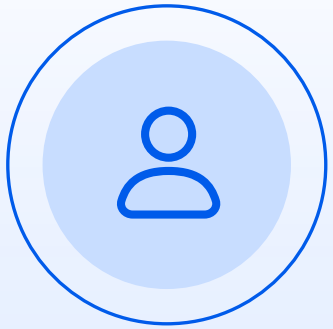
78% of organizations believe it is important to use open source tools hosted by a **neutral party**, primarily due to standards & regulations compliance and trust.

82% of respondents agree that **open source AI is critical for a positive AI future.**

What is AI-RAN?

AI-RAN: Accelerating AI Development for RAN Optimization and Automation

AI-RAN integrates AI into the Radio Access Network to optimize performance.

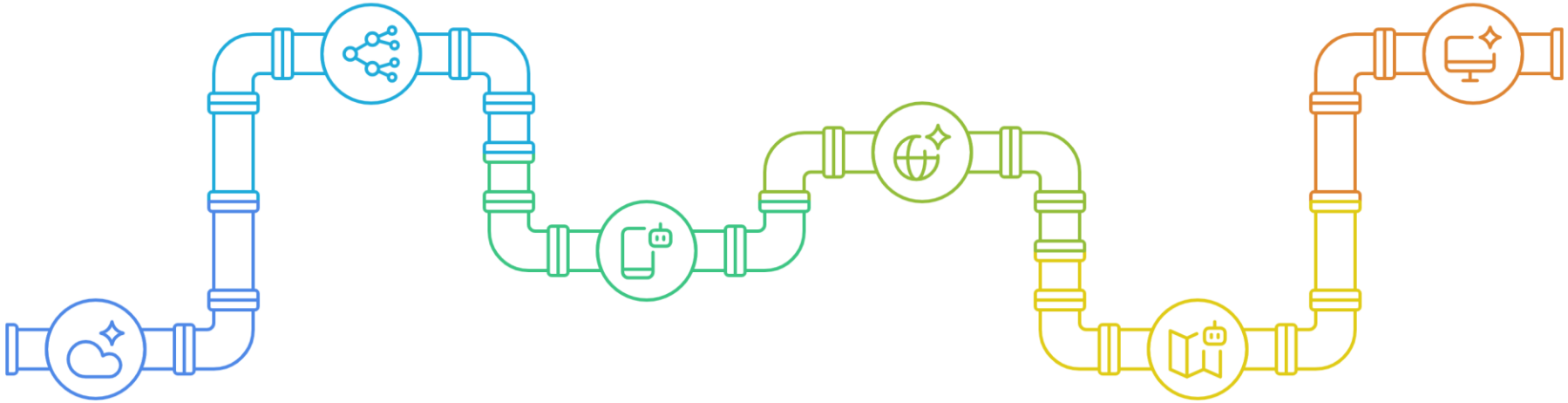


AI-RAN enhances network optimization and automation, reducing manual intervention.

It enables AI-driven traffic management and anomaly detection.

AI-RAN utilizes real and synthetic data to improve AI model training.

AI-RAN Goal: The Evolution of Telecom Networks into AI Platform



01

Demand for Low-Latency AI

The surge in need for quick AI processing.

02

Deployment of Edge AI Clouds

Setting up distributed AI processing nodes.

03

Transformation of RAN

Upgrading RANs into AI inferencing hubs.

04

Integration of AI in 6G

Embedding AI deeply into 6G networks.

05

Development of AI-Native Networks

Creating networks that are inherently AI-driven.

06

Federated AI Learning

Spreading AI training across multiple nodes.

AI-ML in 5G

01 AI/ML in 5G RAN and Air Interface

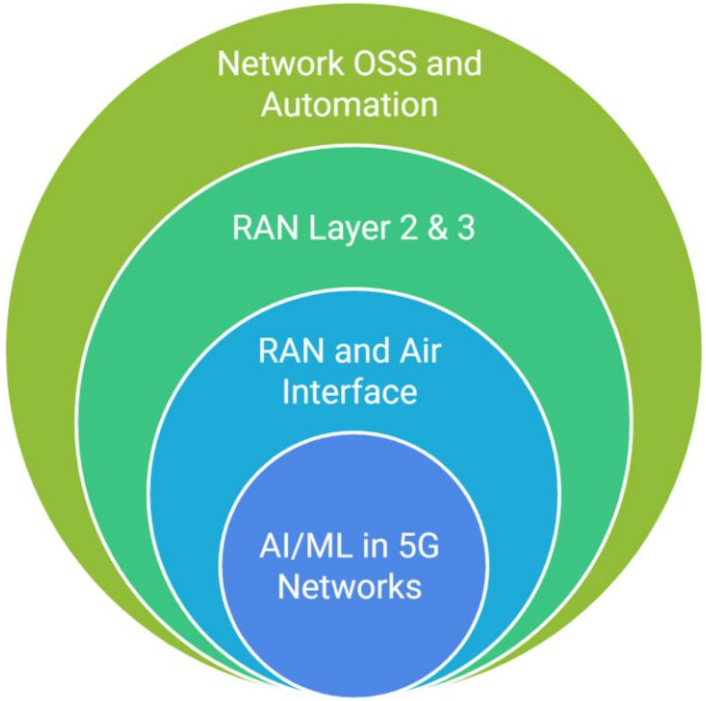
- **Channel State Prediction**: Neural networks improve link performance by predicting Channel State Information (CSI).
- **Beamforming Optimization**: AI models enhance directional transmissions, especially in mmWave communications.
- **Scheduling & Handovers**: Reinforcement learning optimizes resource allocation and ensures seamless mobility management.
- **Energy Efficiency**: Dynamic power adjustments reduce energy consumption without compromising Quality of Service .

02 AI/ML in RAN Layer 2 & Layer 3

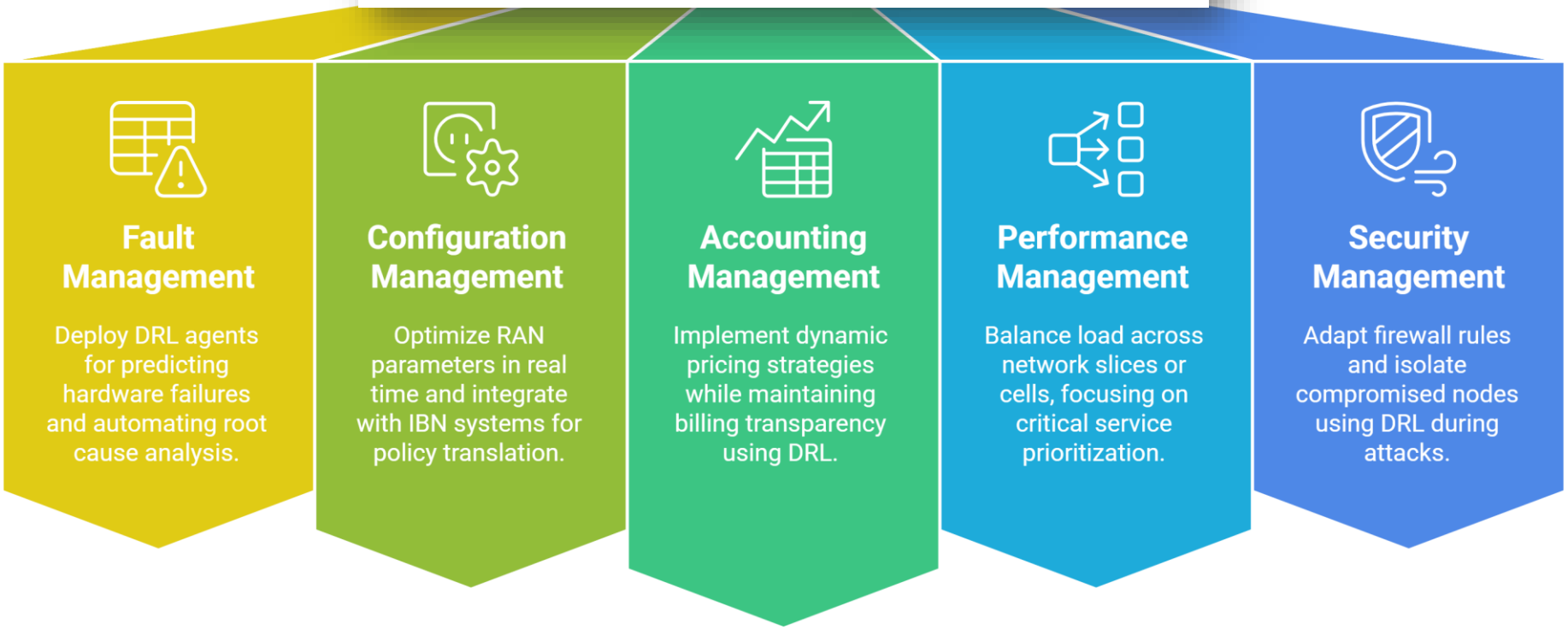
- **Handover Prediction**: Models like RNNs and LSTMs forecast UE movement for proactive handover decisions.
- **Intercell Coordination**: Algorithms minimize interference and balance traffic across cells dynamically.
- **RAN Intelligent Controllers**: Near-real-time and non-real-time RICs deliver actionable insights for smarter network operations.

03 AI/ML in Network OSS and Automation

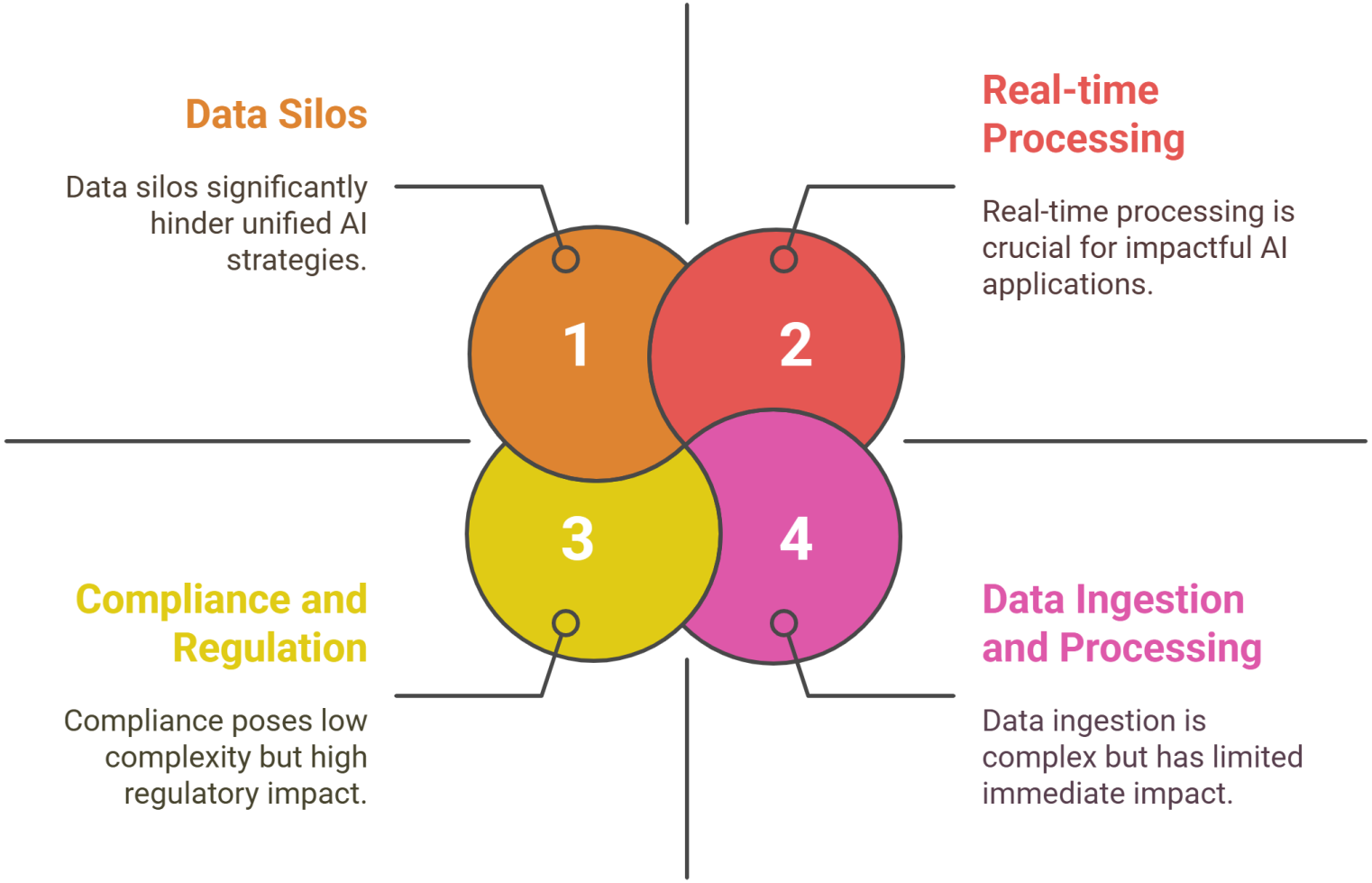
- **Predictive Analytics**: Supervised learning predicts faults and forecasts demand for capacity planning.
- **Self-Healing Networks**: Unsupervised learning detects anomalies and automates fault resolution.
- **Service Assurance**: Real-time analytics ensure optimal performance and faster issue resolution.
- **Security Management**: Behavioral analysis and threat detection enhance network security.
- **Customer Experience**: Virtual assistants and churn prediction improve user satisfaction and retention.



Reinforcement Learning for Telecom RAN Operation Management (FCAPS)



AI Data Management Challenges in Telecom Networks



Solutions

01

Data Silos Solution

Implement federated data platform, using AI-driven data lakes, to consolidate and normalize network data across departments.

02

Real-Time AI Processing

Deploy edge AI processing at distributed locations, utilize low-latency AI inferencing models for real-time network adaptability.

03

Data Ingestion & Preprocessing

Use automated data pipelines, AI-driven cleaning and augmentation techniques to ensure quality datasets and leverage synthetic data.

04

Compliance and Regulatory Adherence

Embed privacy-preserving AI techniques and governance frameworks to comply with GDPR, CCPA, and telecom regulations.

The Challenge of Data Quality

01

AI models require large, high-quality datasets, which are often scarce.

02

Data privacy laws (e.g., GDPR, CCPA) restrict data sharing and utilization.

03

The cost of data collection and annotation is significant, slowing down AI adoption.

04

Raw data is often unstructured, vendor specific and requires extensive preprocessing before AI training.

Solutions

Synthetic Data Generation

Supplements real-world data, ensuring privacy (GDPR, CCPA) by simulating network conditions, training AI models effectively without accessing sensitive telecom data.

Federated Learning

AI training decentralized, compliant, improves models, preserving privacy and mitigating regulatory risk by training across datasets

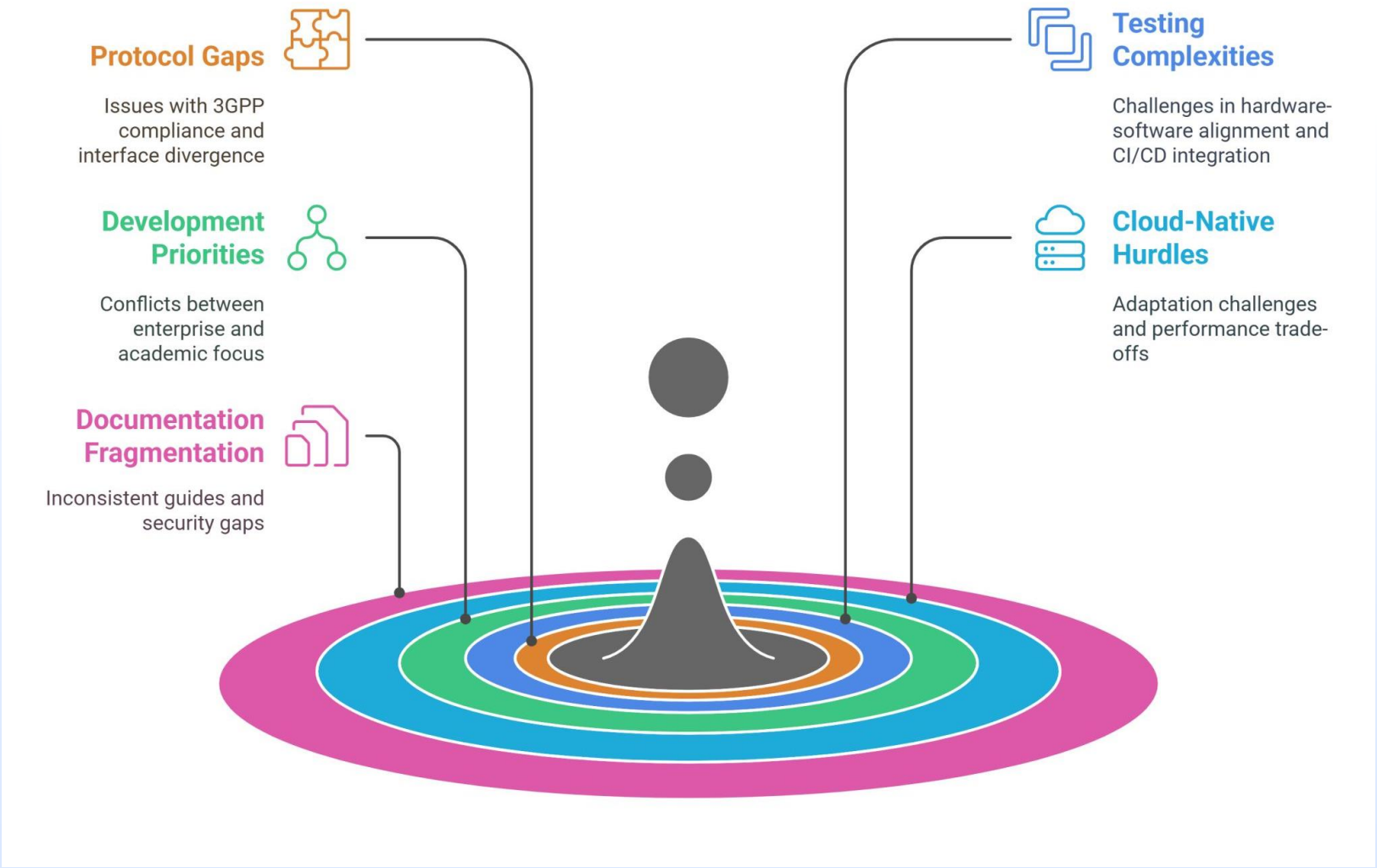
Collaborative Open Data

Establishes industry-wide standards for dataset curation, annotation, and sharing.

Automated Data Pipelines

Streamlines preprocessing of unstructured, vendor-specific data, improving efficiency and quality through AI-driven cleaning, labeling, and augmentation techniques.

AI/ML System Integration Challenges in Telecom Networks



Solutions

Standardized APIs and frameworks

Implement standardized APIs aligned with 3GPP to ensure interoperability between different ...

Industry-academia collaboration

Foster collaboration to ensure AI research aligns with practical telecom needs and streamline ...

Unified Technical Guides

Create unified, well-structured, and regularly updated technical guides, emphasizing security best practic ...

Streamlining AI/ML Integration in AI-RAN

Automated Testing

Deploy automated CI/CD pipelines and AI-driven validation tools to overcome testing complexities ...

Cloud-Native Optimization

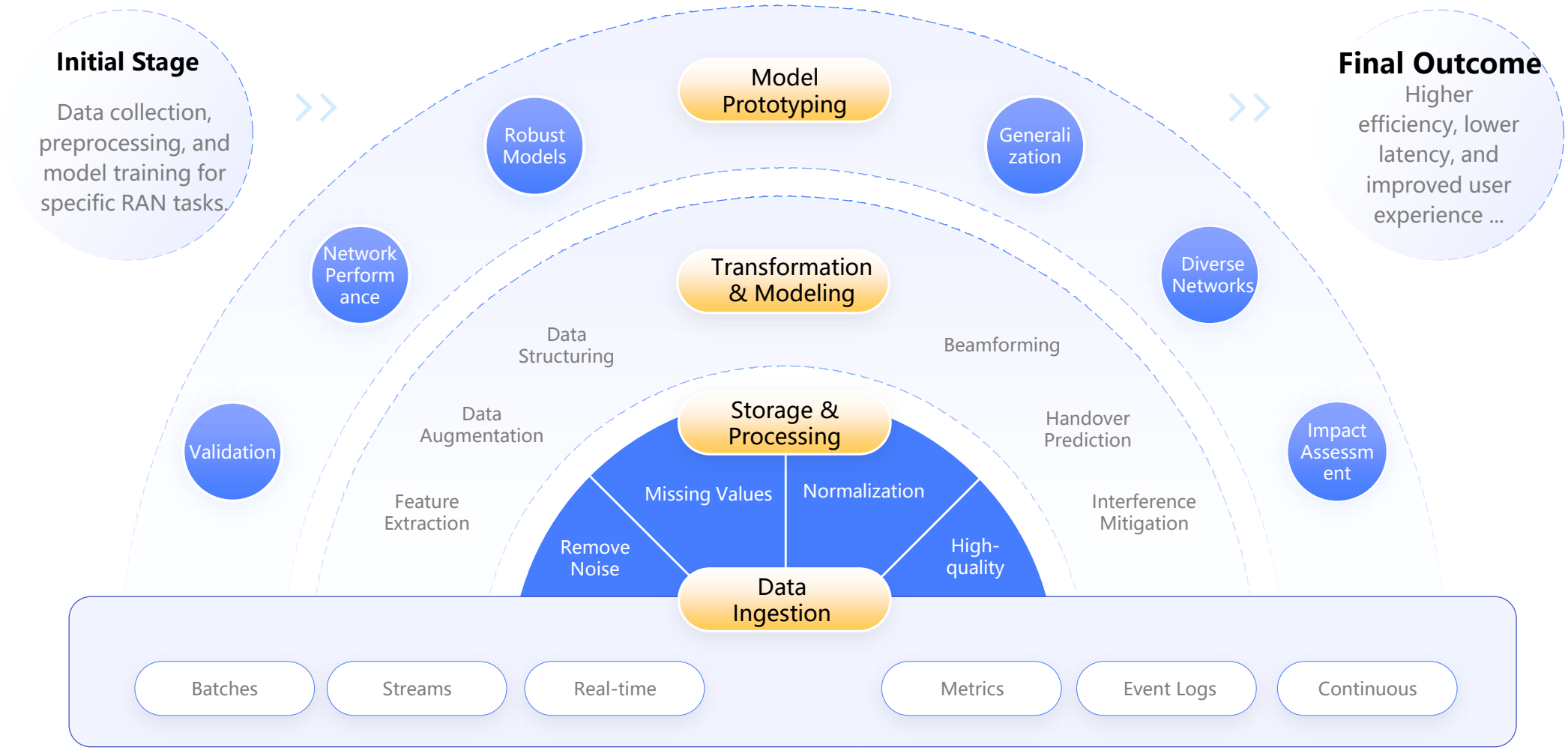
Optimize AI workloads for distributed environments using containerization and orchestration ...

Hybrid and Edge Deployment

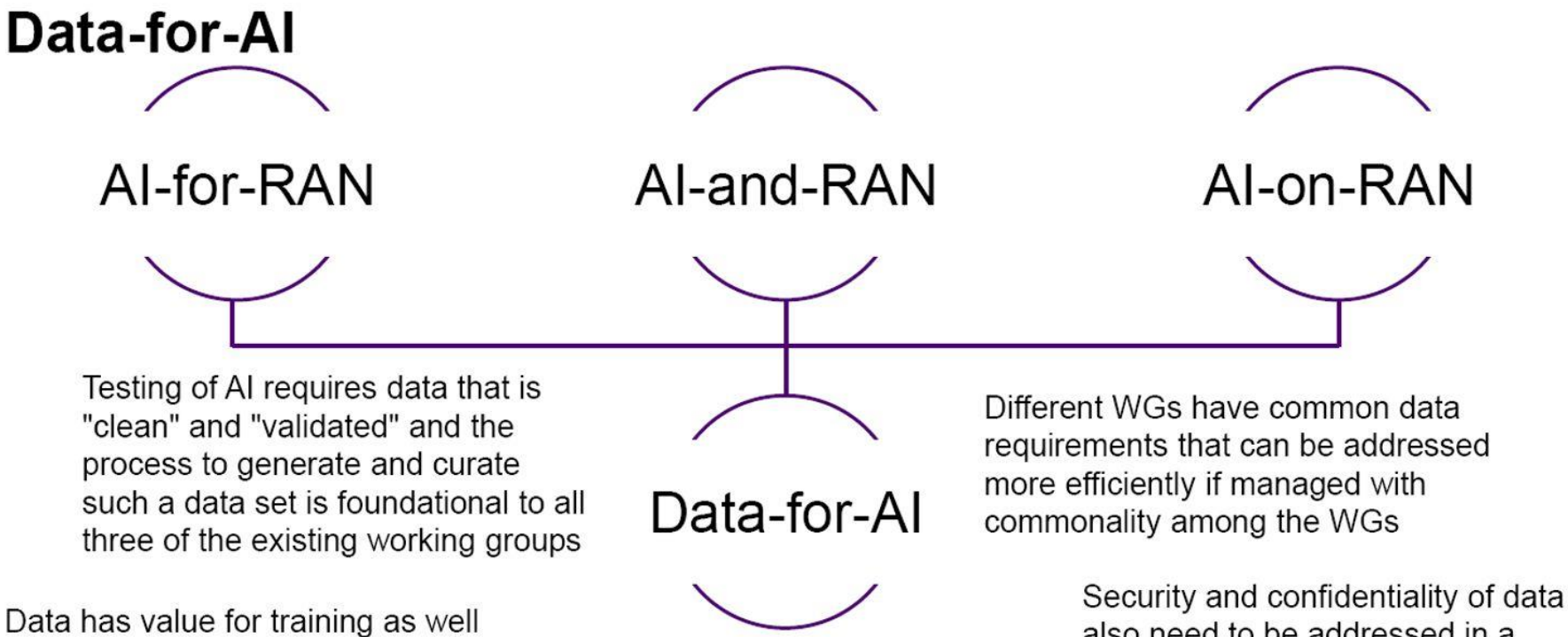
Consider hybrid and edge cloud deployments for performance, cost, and scalability balance in AI-RAN ...

AI ML Model Development Framework in AI-RAN

To streamline data-driven AI integration in Radio Access Networks for enhanced automation and optimized resource management.



Data-for-AI Task Group in AI-RAN Alliance



A **unified approach** to address these aspects across all the AI-RAN working groups is best addressed by a **dedicated working group**

The Collaborative MLOps between Partners

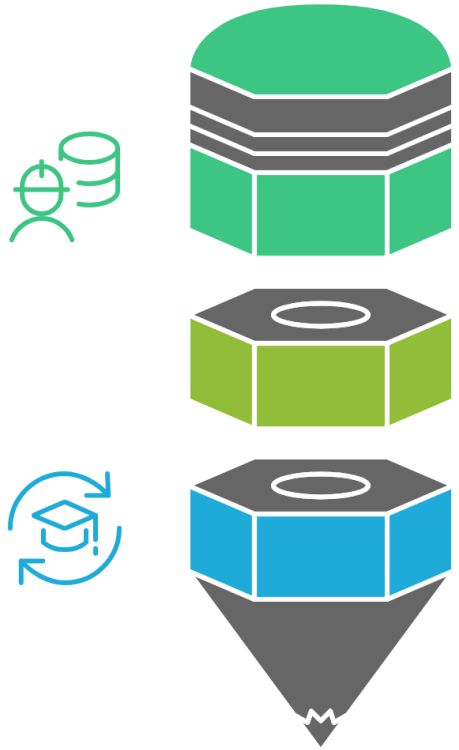
MLOps Framework for VR-DRL

Pre-Training by Vendors

Vendors develop baseline models using synthetic datasets and verify safety constraints.

Continuous Online Learning

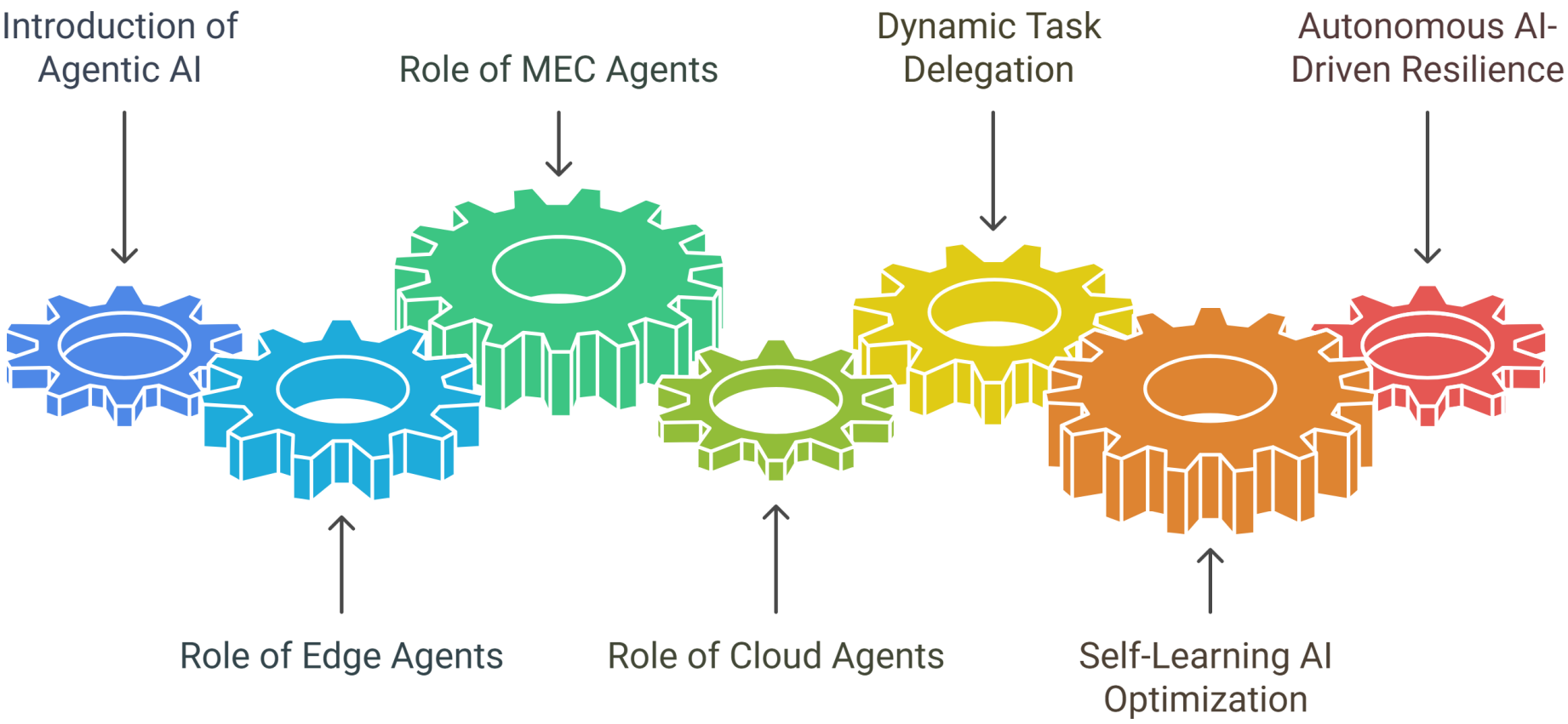
Models are deployed with online learning, using digital twins for safe updates.



Operator-Specific Pre-Training

Telecoms fine-tune models with operational data using federated or transfer learning.

Agentic AI-Driven Telecom Network Optimization



The Promise of Open Source

“ Leveraging Open Source for AI

Open-source AI **democratizes** technology and encourages widespread adoption.

Community-driven innovation **accelerates** AI advancements and knowledge sharing.

Open-source tools **reduce** costs compared to proprietary solutions.

Transparency **ensures** accountability and ethical AI development.

Thanks