Data Governance in DS and MLOps cycle

Clarify what the Data Science and Machine Learning difference in Lifecycle.

Data Science Lifecycle

The Data Science lifecycle is an iterative process aimed at extracting valuable insights from data to solve business problems. It encompasses a broad range of activities, from understanding business needs to communicating results. Key stages typically include:

- **Discovery/Business Understanding**: Identifying the problem, objectives, and requirements.
- **Data Preparation**: Acquiring, cleaning, and exploring data.
- **Model Planning and Building**: Selecting techniques (which may include machine learning) and developing models or analyses.
- **Evaluation and Communication**: Assessing results and presenting insights.
- **Operationalize/Deployment**: Implementing solutions and monitoring for ongoing value.

This lifecycle emphasizes exploratory data analysis, statistical methods, and decision-making support, often beyond just predictive modeling.

MLOps Lifecycle

MLOps (Machine Learning Operations) lifecycle focuses on the end-to-end process of building, deploying, and maintaining machine learning model pipeline in production environments. It applies DevOps principles to ML, emphasizing automation, scalability, and reliability. Common stages include:

- **Scoping/Problem Definition**: Defining the ML task and requirements.
- **Data Management**: Collecting, preparing, and versioning data.
- **Modeling**: Experimenting, training, and validating models.
- **Deployment**: Integrating models into production via CI/CD pipelines.
- **Monitoring and Maintenance**: Tracking performance, handling model drift, and retraining as needed.

MLOps is designed to streamline workflows, reduce deployment risks, and ensure models operate efficiently at scale.

Key Differences Between Data Science and MLOps Lifecycles

- While both lifecycles involve data handling and modeling, they differ in focus, scope, and application:
- **Focus**: The Data Science lifecycle is centered on insight generation, hypothesis testing, and business value through various analytical methods (e.g., statistics, visualization). In contrast, MLOps is engineering-oriented, prioritizing the operationalization of ML models, such as automation, deployment pipelines, and continuous integration.
- **Scope**: Data Science is broader and more exploratory, often including non-ML elements like descriptive analytics and ad-hoc projects. MLOps is narrower but deeper in production aspects, treating ML models as software that requires versioning, scaling, and lifecycle management similar to application development.
- **Roles and Integration**: Data scientists typically drive the Data Science lifecycle, focusing on research and prototyping. MLOps involves collaboration between data scientists, engineers, and operations teams to bridge the gap from prototype to production, addressing challenges like model reproducibility and infrastructure.
- In summary, Data Science lifecycle is about discovering and analyzing data for insights, while MLOps extends this by operationalizing ML models for reliable, scalable use in real-world applications.
-

Summary in One Sentence

Data Science Models → Built for understanding and decision support (may stop at insights).

ML Models → Built for automation and prediction in production, requiring ongoing monitoring and retraining.

How Data Governance Affects These Lifecycles

Data governance involves establishing policies, standards, and processes to manage data quality, security, compliance, and ethical use throughout its lifecycle. It acts as a foundational framework that enhances reliability and mitigates risks in both Data Science and MLOps, but its impact varies based on the lifecycle's emphasis.

- **In the Data Science Lifecycle**: Governance ensures data integrity and compliance at every stage, from acquisition to analysis. It prevents manipulation, enforces quality checks, and promotes a single source of truth, reducing errors in insights and enabling better decision-making. For example, it impacts data preparation by requiring metadata management and bias detection, and in evaluation by mandating audits for fairness. Without it, projects risk poor data quality, leading to unreliable results; with it, governance boosts efficiency, security, and regulatory adherence.

- **In the MLOps Lifecycle**: Governance extends to model governance, covering access controls, versioning, tracking, and ethical sourcing of data and models. It smooths deployment and monitoring by ensuring data quality for training, minimizing risks like bias or drift, and facilitating scalability in multi-account environments. In regulated industries, it enforces accountability and traceability throughout the ML lifecycle, from data collection to production. Strong governance prevents issues like security vulnerabilities or non-compliance, while enabling automation and collaboration; weak governance can lead to inefficient workflows and higher operational risks.

Overall, data governance is a critical enabler for both lifecycles, fostering trust, efficiency, and compliance, but it is especially integral to MLOps for handling the complexities of production ML systems.

What we should focus on:

1. **Strategy & Planning**
    - Define data governance goals, aligning with business and compliance needs.
    - Work with cross-functional team (data stewards, DS/ML engineers, IT).
    - Create a framework for data policies, DS lifecycle, and MLOps integration.
2. **Data Management**
    - Catalog data with metadata tools for lineage and quality.
    - Set up secure ingestion, access controls, dataset versioning, and data cleaning (e.g., handling missing values, outliers, normalization).
    - Audit data for biases, quality, and compliance before modeling.
    - Trustworthy analysis.
    -
3. **Model Development & Evaluation**
    - Conduct feature selection (e.g., correlation analysis, feature importance) to identify relevant predictors base on the target.
    - Perform correct model selection (e.g., comparing algorithms like SVM, random forest) based on the best practice of performance metrics.
    - Enforce ethical AI reviews and bias checks, using tools (Jupyter, MLflow) for experiment tracking and evaluation (e.g., accuracy, fairness).
4. **Deployment**

- Build CI/CD pipelines for automated model testing and deployment.
- Implement model registries and secure deployment with A/B testing.
- Document processes for governance and rollback plans.
-

5. **Monitoring & Improvement or Continuous Evaluation**

    - Monitor data drift and model performance with real-time tools like dashboards.
    - Schedule retraining and conduct regular audits.
    - Refine governance and scale practices with team training.

==========

Data Science and **MLOps** are distinct disciplines—but platforms like **Dataiku** bring them together to streamline the entire lifecycle of data projects. Here's a breakdown of the difference and how they work together in Dataiku:

🔍 Data Science vs. MLOps: Key Differences

Aspect	Data Science	MLOps (Machine Learning Operations)
Focus	Exploration, modeling, and insights	Deployment, monitoring, and scalability
Goal	Build predictive models and extract insights	Operationalize models reliably and efficiently
Tasks	Data cleaning, feature engineering, model training	CI/CD, model deployment, versioning, monitoring
Tools	Python, R, notebooks, visualization	Docker, Kubernetes, MLflow, APIs
Audience	Data scientists, analysts	ML engineers, DevOps teams

🧩 How Dataiku Combines Both

Dataiku is designed to **bridge the gap** between these two worlds:

☑ *For Data Scientists:*

- Visual recipes and notebooks for data prep and modeling
- AutoML and custom model training
- Collaboration tools for experimentation

☑ *For MLOps Teams:*

- Model versioning and deployment pipelines
- Monitoring dashboards (drift, performance)
- Automation via scenarios and APIs
- Integration with CI/CD tools and cloud platforms

🔄 Unified Workflow in Dataiku

In Dataiku, you can:

1. **Build a model** → using visual tools or code.
2. **Deploy it** → to a staging or production environment.
3. **Monitor it** → for performance and data drift.
4. **Retrain it** → automatically based on triggers or schedules.

This end-to-end capability is why Dataiku is often used by **cross-functional teams**—even though Data Science and MLOps are different, they **collaborate** within the same platform.

Governance Across the Lifecycles in Dataiku

Your document highlights how **data governance** strengthens both lifecycles. Here's how it maps into Dataiku's capabilities:

☑ *In Data Science Lifecycle (Exploration & Modeling)*

- **Metadata management**: Dataiku's cataloging tools help track lineage and ensure data integrity.
- **Bias detection & auditability**: You can use plugins and notebooks to run fairness checks and document assumptions.

- **Secure access & versioning**: Role-based permissions and dataset versioning support compliance and reproducibility.

In MLOps Lifecycle (Deployment & Monitoring)

- **Model governance**: Dataiku supports model registries, version control, and rollback strategies.
- **CI/CD pipelines**: Automated deployment scenarios ensure consistency and traceability.
- **Monitoring & retraining**: Dashboards and triggers help detect drift and schedule retraining, aligned with governance policies.

Governance Strategy in Dataiku Projects

From your document's recommendations, here's how governance can be embedded in Dataiku workflows:

Governance Area	Dataiku Implementation
Strategy & Planning	Define governance goals in project documentation and scenarios
Data Management	Use metadata, lineage tools, and secure access controls
Model Development	Track experiments with MLflow, enforce ethical reviews
Deployment	Automate with CI/CD, document deployment steps
Monitoring & Improvement	Set up drift detection, schedule audits and retraining