DataOps For The Modern Computer Vision Stack

James Le

Superb AI

DataOps For The Modern Computer Vision Stack

James Le

Presenter Profile

James Le

Now

- Data Advocate
- Data Writer
- Data Podcaster

Before

- ML Researcher
- Data Scientist
- Data Journalist

Interests

- Data/ML Infrastructure
- Venture Capital
- Community-Led Growth

Superb AI

About Me

NOW

- Data Advocate
- Data Writer
- Data Podcaster

BEFORE

- Machine Learning Researcher
- Data Scientist
- Data Journalist

INTERESTS

- Data/ML Infrastructure
- Venture Capital
- Community-Led Growth

ROCHESTER INSTITUTE OF TECHNOLOGY
RIT
1829

Data Notes

Technical Concepts + Industry Advice From The Data World

Agenda

1. What Is DataOps?
2. Why DataOps For Computer Vision?
3. DataOps Key Principles
4. DataOps Pipeline for the Computer Vision Stack
5. Data Challenges for Computer Vision Teams
6. The Future of the Modern Computer Vision Stack

Agenda

Superb AI

1 - What Is DataOps?

2 - Why DataOps for Computer Vision?

3 - DataOps Key Principles

4 - DataOps Pipeline for the Computer Vision Stack

5 - Data Challenges for Computer Vision Teams

6 - The Future of The Modern Computer Vision Stack

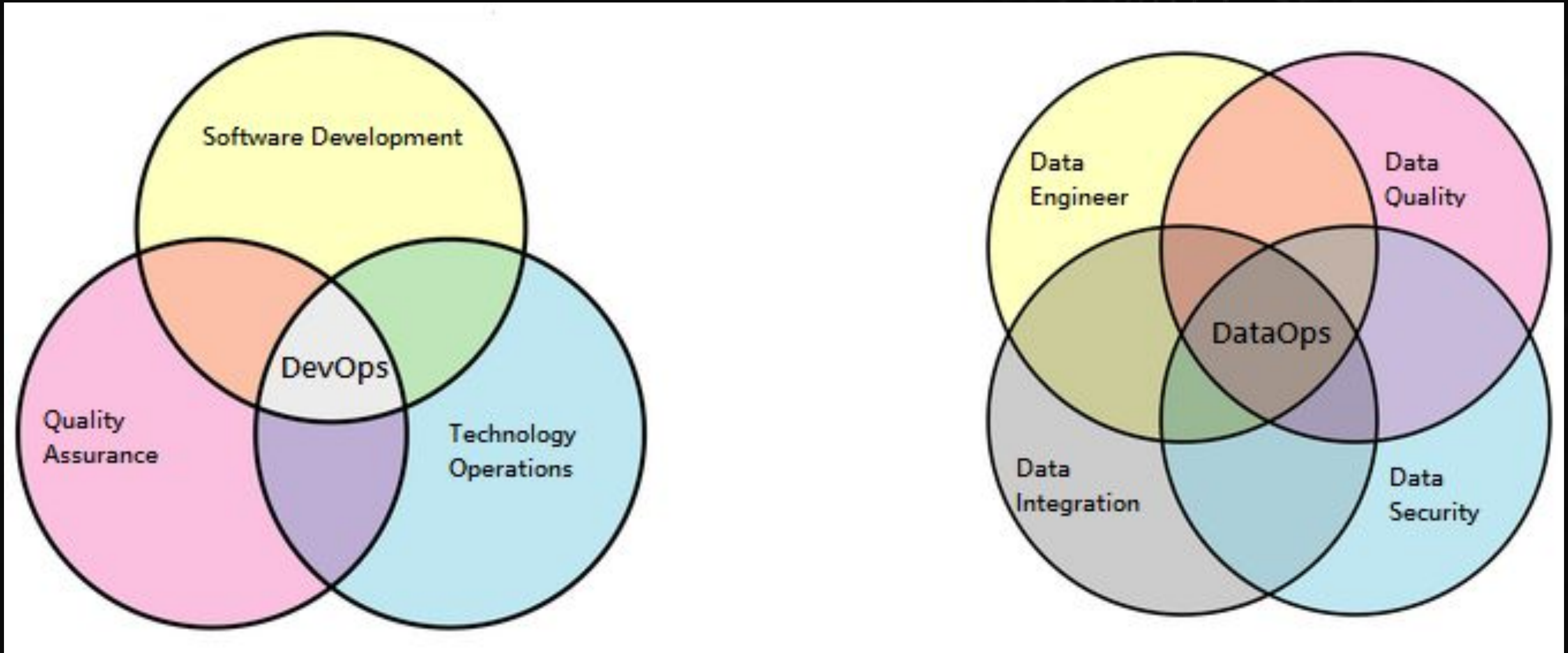
What Is

DataOps?

Superb AI

What Is DataOps?

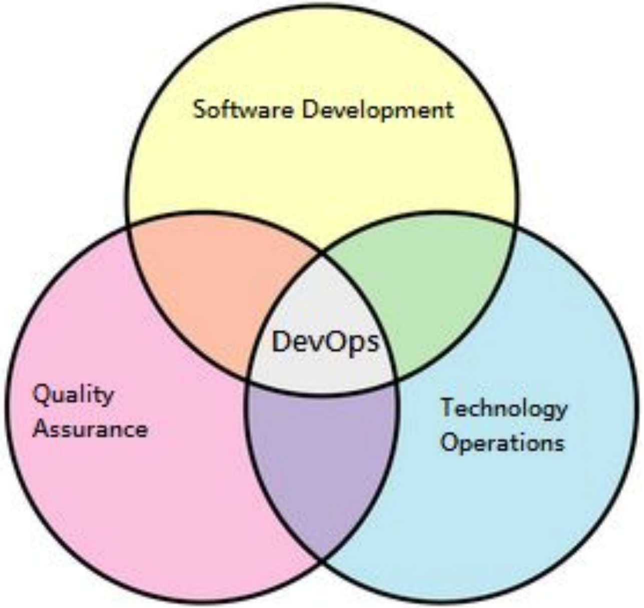
DataOps vs DevOps

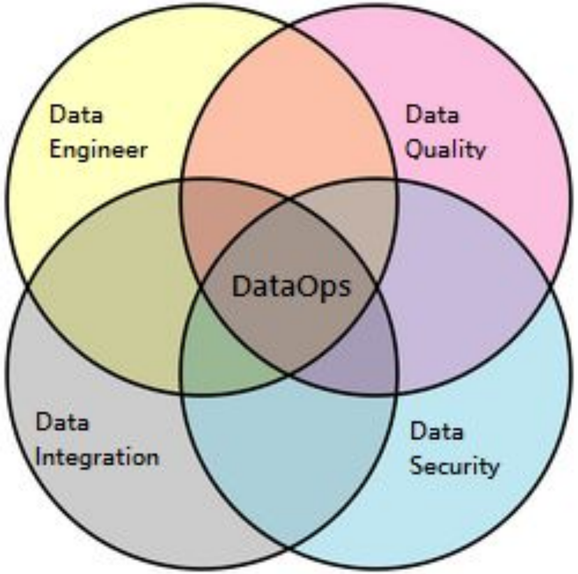


Source: DevOps vs DataOps (by Sprinkle Data)

Superb AI

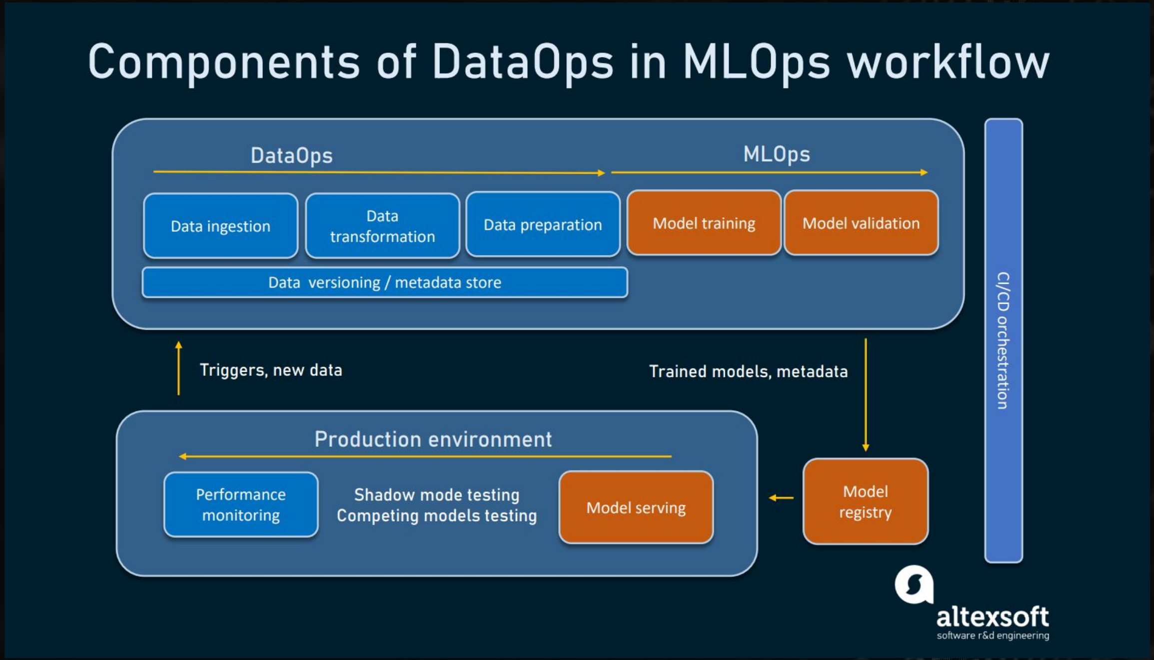
DataOps vs DevOps





Source: [DevOps vs DataOps](#) (by Sprinkle Data)

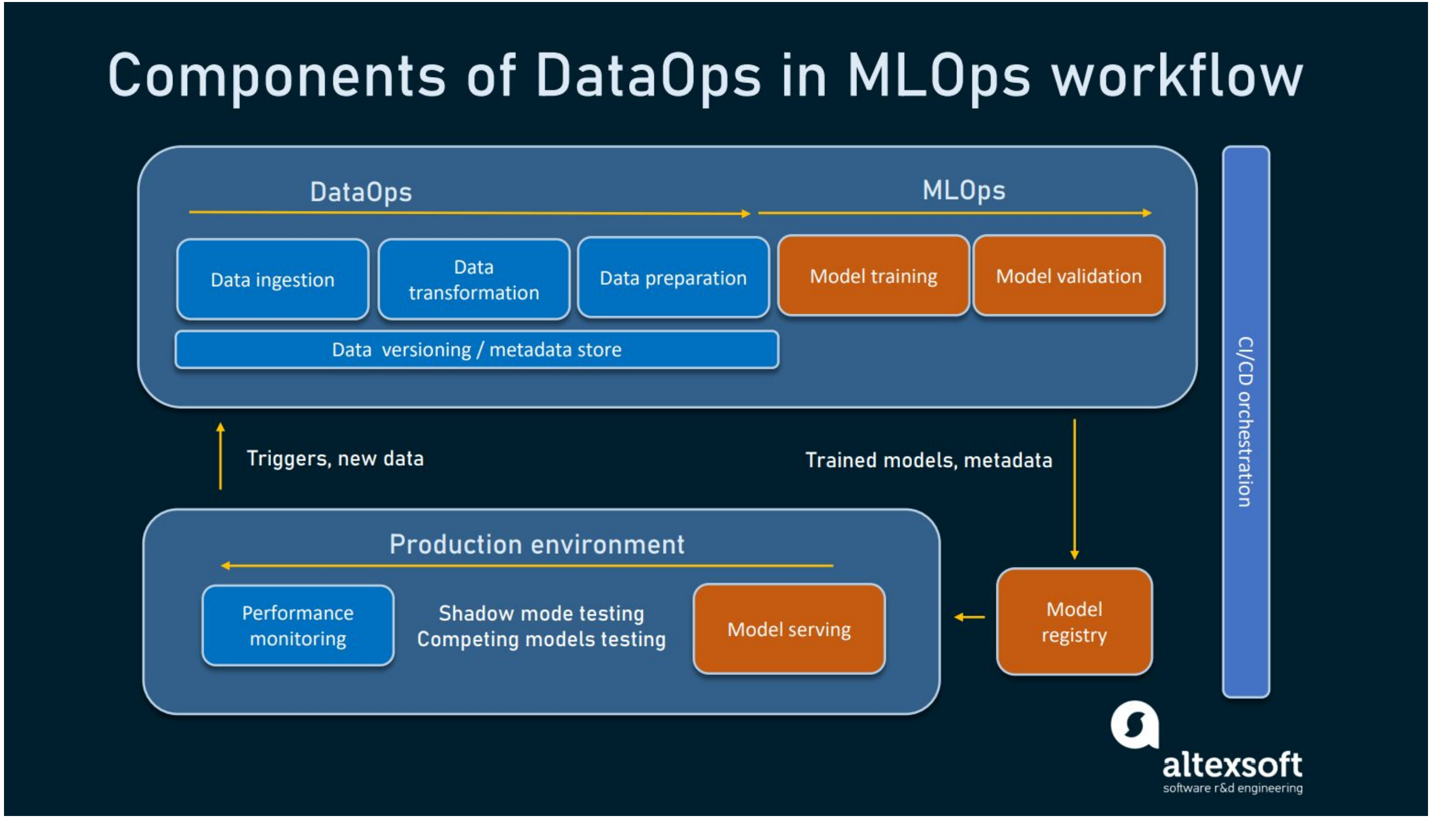
DataOps vs MLOps



Source: DataOps - Adjusting DevOps for Analytics Product Development (by Altexsoft)

DataOps vs MLOps

Superb AI

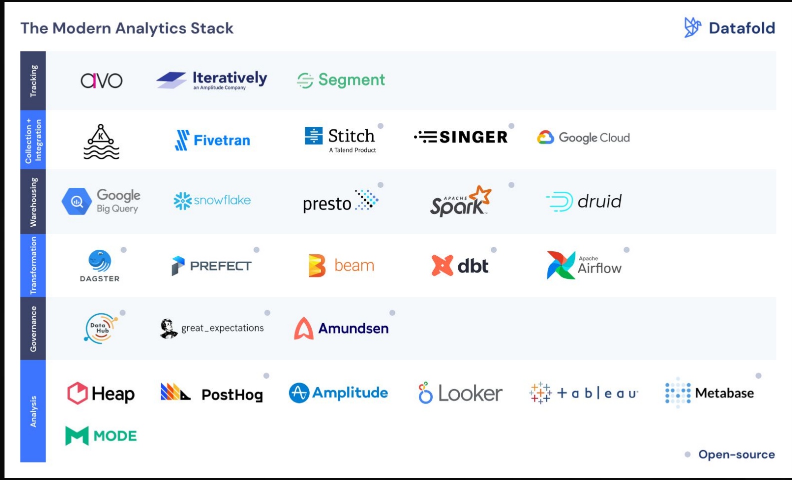


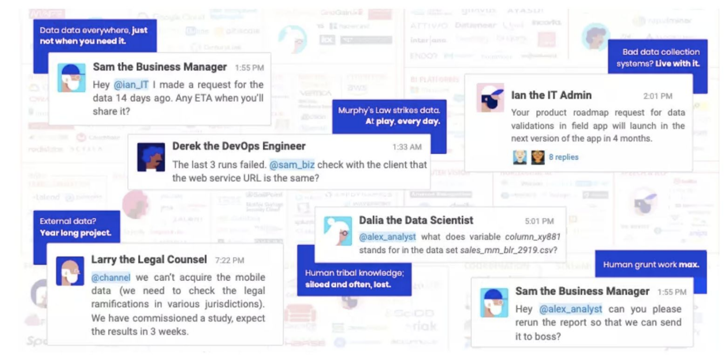
Source: [DataOps - Adjusting DevOps for Analytics Product Development](#)
(by Altexsoft)

What Led To The Rise of DataOps?

1. Massive Volumes of Complex Data
2. Technology Overload
3. Diverse Roles and Mandates

Source: [Apache Spark DataFrames for Large Scale Data Science]() (by Databricks)





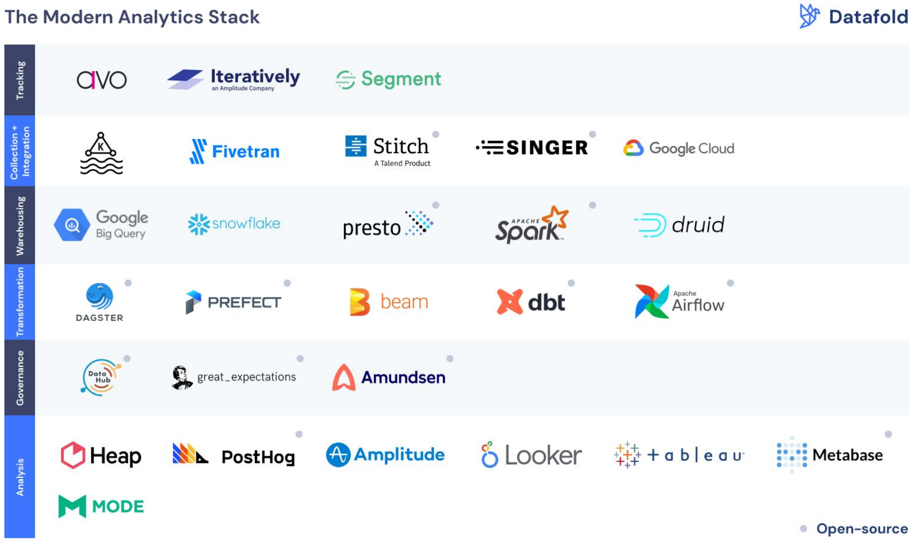
Source: [What is DataOps?]() (by Atlan)

Source: [Modern Analytics Stack]() (by Datafold)

Superb AI

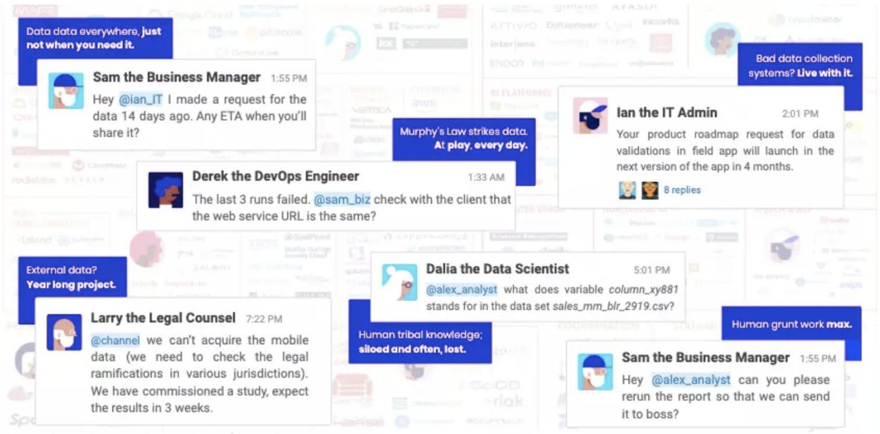
What Led To The Rise of DataOps?

1. Massive Volumes of Complex Data
2. Technology Overload
3. Diverse Roles and Mandates



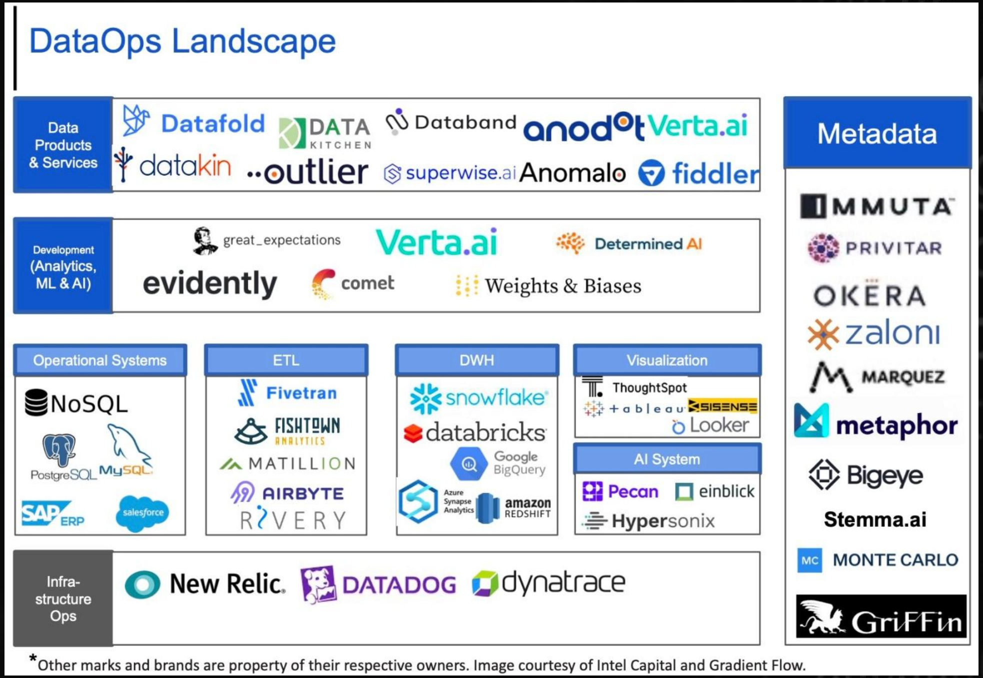
Source: [Modern Analytics Stack](#) (by Datafold)

Source: [Apache Spark DataFrames for Large Scale Data Science](#) (by Databricks)



Source: [What is DataOps?](#) (by Atlan)

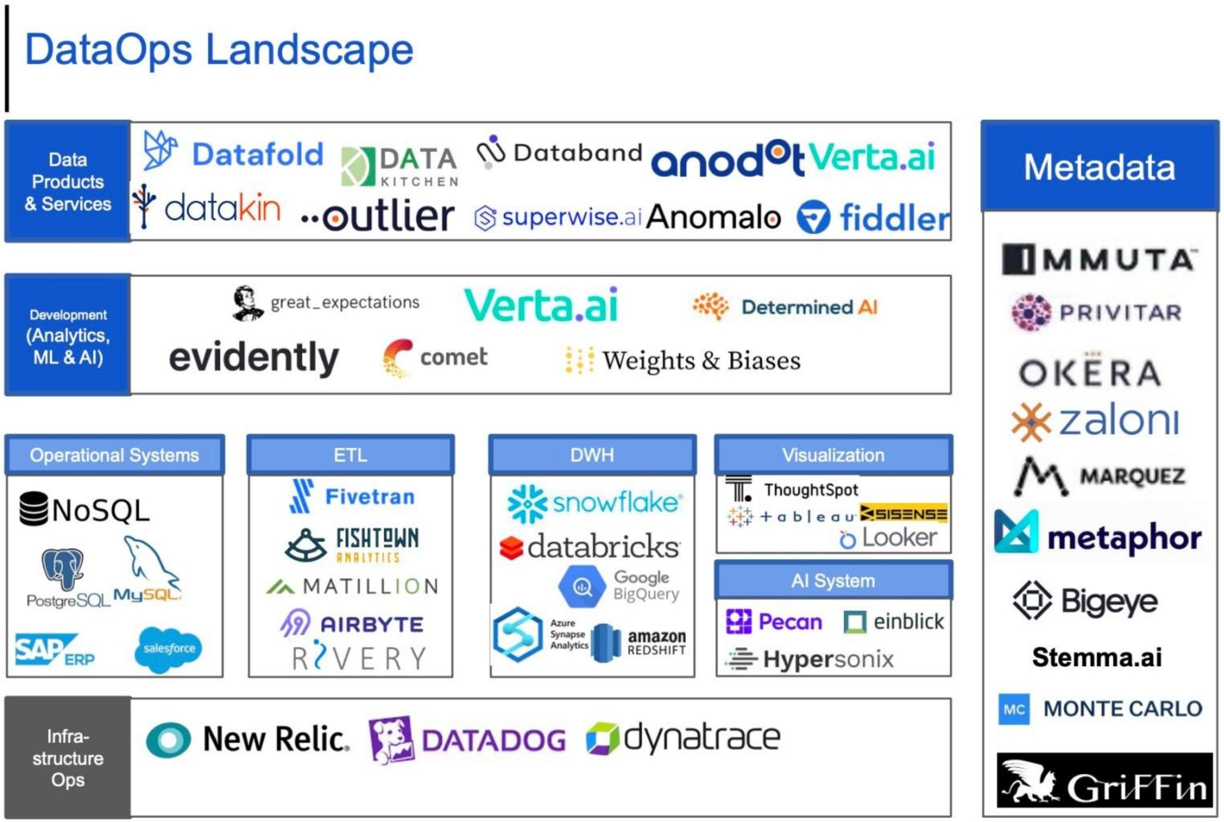
The DataOps Landscape



Source: What is DataOps? (by Gradient Flow)

Superb AI

The DataOps Landscape



*Other marks and brands are property of their respective owners. Image courtesy of Intel Capital and Gradient Flow.

Source: What is DataOps? (by Gradient Flow)

Why DataOps For

Computer Vision?

Superb AI

Why DataOps For Computer Vision?

Why DataOps For Computer Vision? (⅓)

Data Is More Important Than Models

This sentiment is conveyed by Francois Chollet - the creator of Keras (Source: Twitter)

Superb AI

Why DataOps For Computer Vision? (⅓)

Data Is More Important Than Models

Thread

François Chollet @fchollet · Jan 24

ML researchers work with fixed benchmark datasets, and spend all of their time searching over the knobs they do control: architecture & optimization. In applied ML, you're likely to spend most of your time on data collection and annotation -- where your investment will pay off.

23 430 2K

François Chollet
@fchollet

Replying to @fchollet

In general, there is very little research done on best practices for data curation / cleaning / annotation, even though these steps have more impact on applications than incremental architecture improvements. Preparing the data is an exercise left to the reader

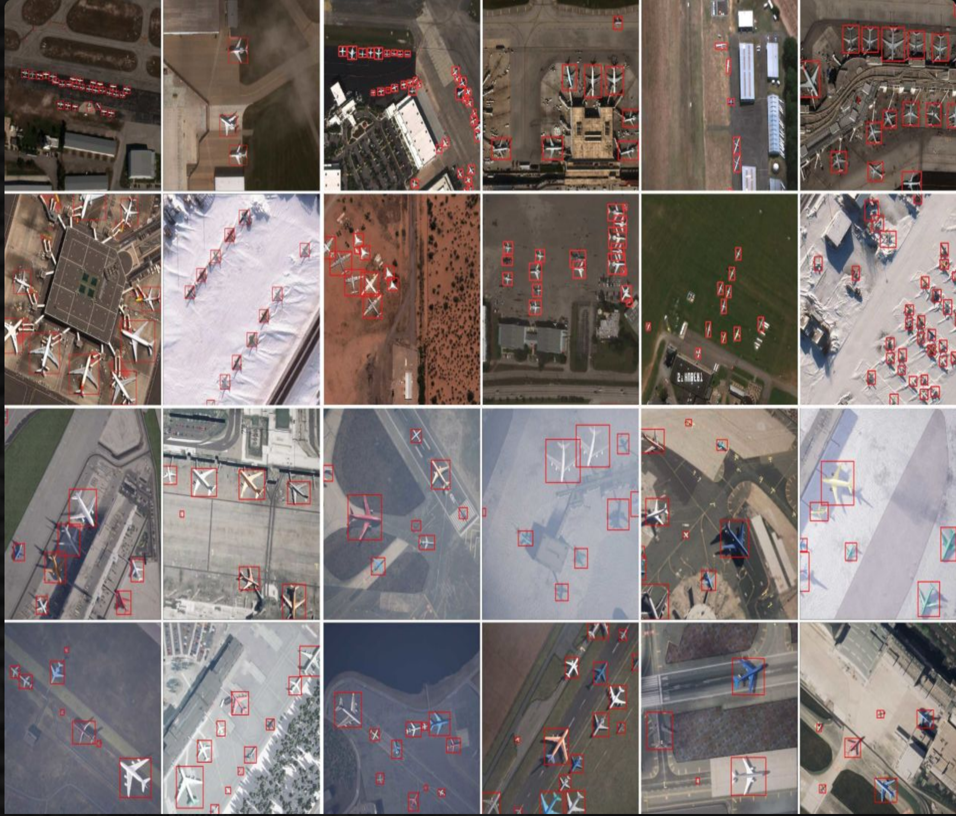
11:22 AM · Jan 24, 2021 · Twitter for Android

176 Retweets 35 Quote Tweets 1,382 Likes

This sentiment is conveyed by Francois Chollet - the creator of Keras (Source: Twitter)

Why DataOps For Computer Vision? (⅔)

Unstructured Data Preparation Is Challenging



Rareplane dataset that incorporates both real and synthetically generated satellite imagery (Source: Superb AI)

Superb AI

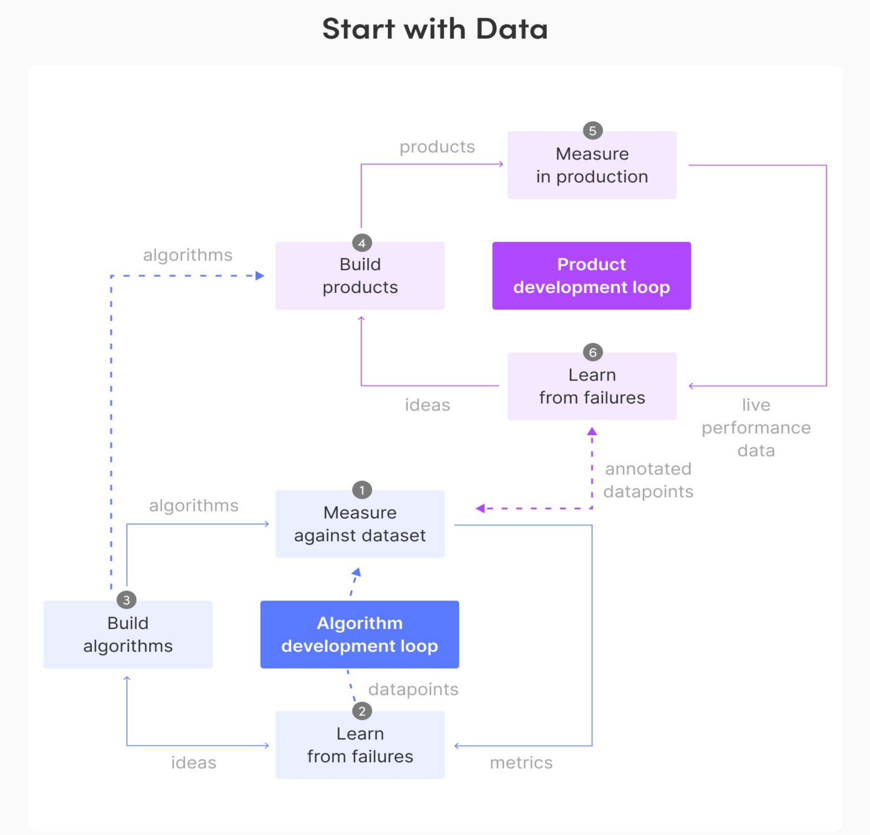
Why DataOps For Computer Vision? (⅔)

Unstructured Data Preparation Is Challenging

Rareplanes dataset that incorporates both real and synthetically generated satellite imagery (Source: Superb AI)

Why DataOps For Computer Vision? (3/3)

Building Computer Vision Applications Is Iterative

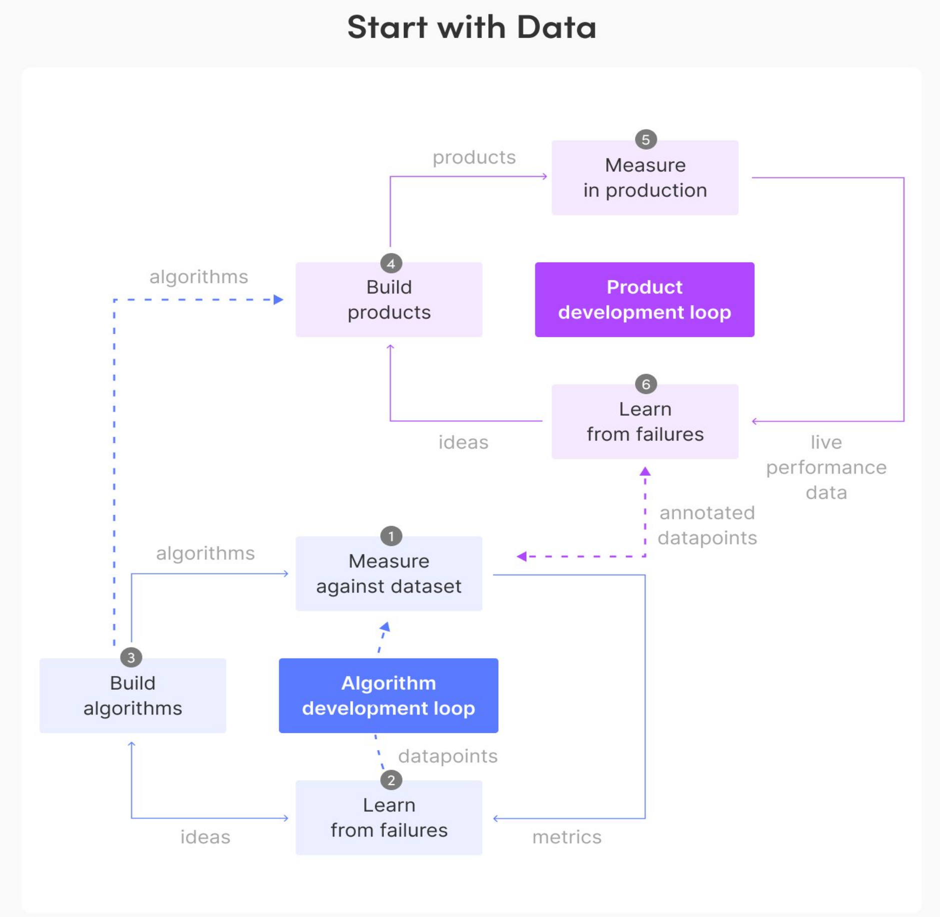


The Two Loops of Building Algorithmic Products (Source: Taivo Pungas)

Superb AI

Why DataOps For Computer Vision? (3/3)

Building Computer Vision Applications is Iterative



The Two Loops of Building Algorithmic Products (Source: [Taivo Pungas]())

DataOps

Key Principles

Superb AI

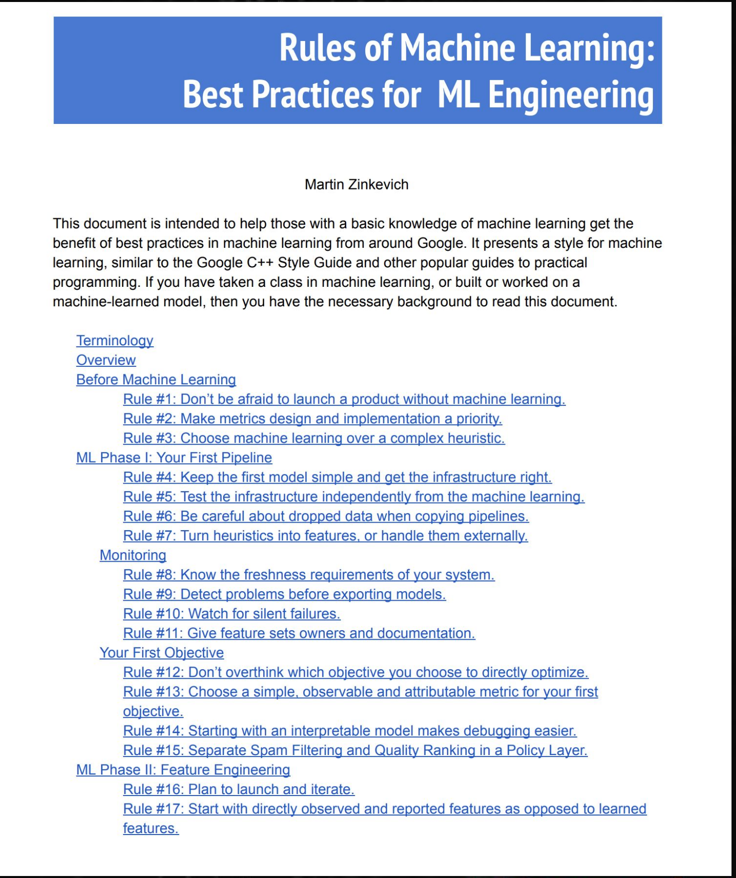
DataOps Key Principles

Principle 1 - Implement Best Practices for Development

Follow Software Engineering Cycle Guidelines

- Version control
- Code reviews
- Unit testing
- Artifacts management
- Release automation
- Infrastructure as code
- OSS Tools: Git, Docker, Terraform

Source: Engineering Best Practices for ML (by Alex Serban)



Rules of Machine Learning:
Best Practices for ML Engineering

Martin Zinkevich

This document is intended to help those with a basic knowledge of machine learning get the benefit of best practices in machine learning from around Google. It presents a style for machine learning, similar to the Google C++ Style Guide and other popular guides to practical programming. If you have taken a class in machine learning, or built or worked on a machine-learned model, then you have the necessary background to read this document.

Terminology
Overview
Before Machine Learning
Rule #1: Don't be afraid to launch a product without machine learning.
Rule #2: Make metrics design and implementation a priority.
Rule #3: Choose machine learning over a complex heuristic.
ML Phase I: Your First Pipeline
Rule #4: Keep the first model simple and get the infrastructure right.
Rule #5: Test the infrastructure independently from the machine learning.
Rule #6: Be careful about dropped data when copying pipelines.
Rule #7: Turn heuristics into features, or handle them externally.
Monitoring
Rule #8: Know the freshness requirements of your system.
Rule #9: Detect problems before exporting models.
Rule #10: Watch for silent failures.
Rule #11: Give feature sets owners and documentation.
Your First Objective
Rule #12: Don't overthink which objective you choose to directly optimize.
Rule #13: Choose a simple, observable and attributable metric for your first objective.
Rule #14: Starting with an interpretable model makes debugging easier.
Rule #15: Separate Spam Filtering and Quality Ranking in a Policy Layer.
ML Phase II: Feature Engineering
Rule #16: Plan to launch and iterate.
Rule #17: Start with directly observed and reported features as opposed to learned features.

Source: Rules of ML (by Google)

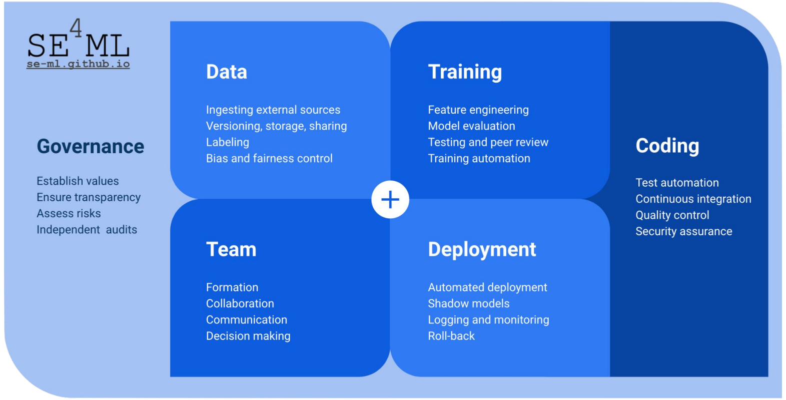
Superb AI

Principle 1 - Implement Best Practices for Development

Superb AI

Follow Software Engineering Cycle Guidelines

- Version control
- Code reviews
- Unit testing
- Artifacts management
- Release automation
- Infrastructure as code
- OSS Tools: Git, Docker, Terraform



Source: Engineering Best Practices for ML (by Alex Serban)

Rules of Machine Learning: Best Practices for ML Engineering

Martin Zinkevich

This document is intended to help those with a basic knowledge of machine learning get the benefit of best practices in machine learning from around Google. It presents a style for machine learning, similar to the Google C++ Style Guide and other popular guides to practical programming. If you have taken a class in machine learning, or built or worked on a machine-learned model, then you have the necessary background to read this document.

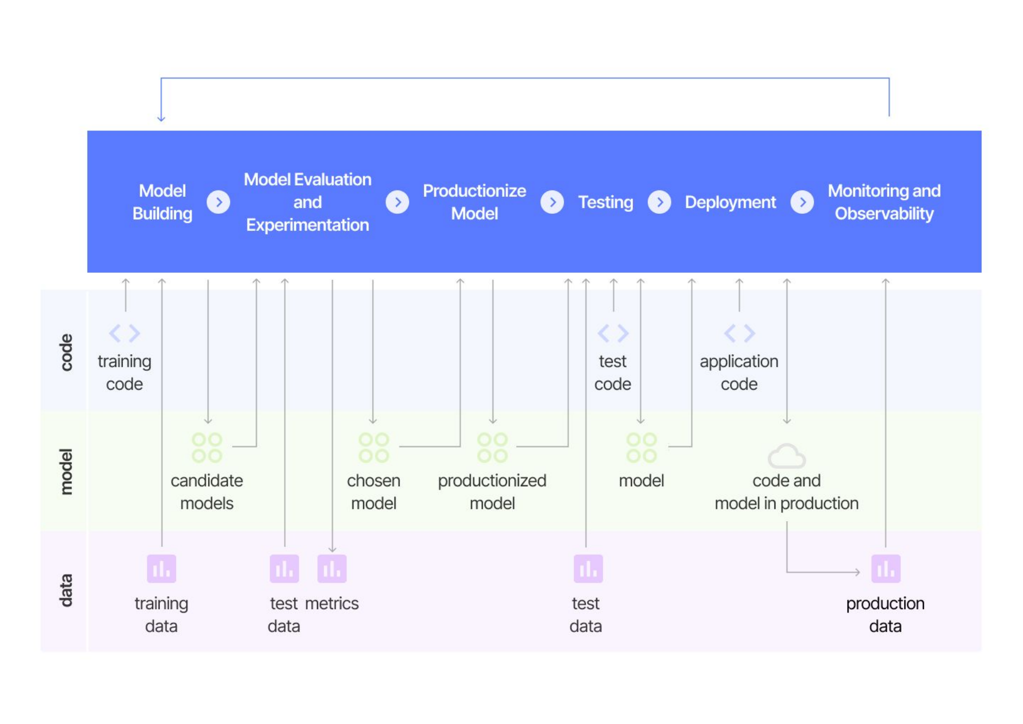
Terminology
Overview
Before Machine Learning
Rule #1: Don't be afraid to launch a product without machine learning.
Rule #2: Make metrics design and implementation a priority.
Rule #3: Choose machine learning over a complex heuristic.
ML Phase I: Your First Pipeline
Rule #4: Keep the first model simple and get the infrastructure right.
Rule #5: Test the infrastructure independently from the machine learning.
Rule #6: Be careful about dropped data when copying pipelines.
Rule #7: Turn heuristics into features, or handle them externally.
Monitoring
Rule #8: Know the freshness requirements of your system.
Rule #9: Detect problems before exporting models.
Rule #10: Watch for silent failures.
Rule #11: Give feature sets owners and documentation.
Your First Objective
Rule #12: Don't overthink which objective you choose to directly optimize.
Rule #13: Choose a simple, observable and attributable metric for your first objective.
Rule #14: Starting with an interpretable model makes debugging easier.
Rule #15: Separate Spam Filtering and Quality Ranking in a Policy Layer.
ML Phase II: Feature Engineering
Rule #16: Plan to launch and iterate.
Rule #17: Start with directly observed and reported features as opposed to learned features.

Source: Rules of ML (by Google)

Principle 2 - Automate and Orchestrate All Data Flows

Continuous Integration and Continuous Delivery

- Automate deployment with CI/CD pipelines
- Discourage manual data wrangling
- Run the data flows using an orchestrator
    - Backfilling
    - Scheduling
    - Pipeline metrics
- OSS Tools: Airflow, Dagster, Prefect



Source: *Continuous Delivery for Machine Learning* *(by ThoughtWorks)*

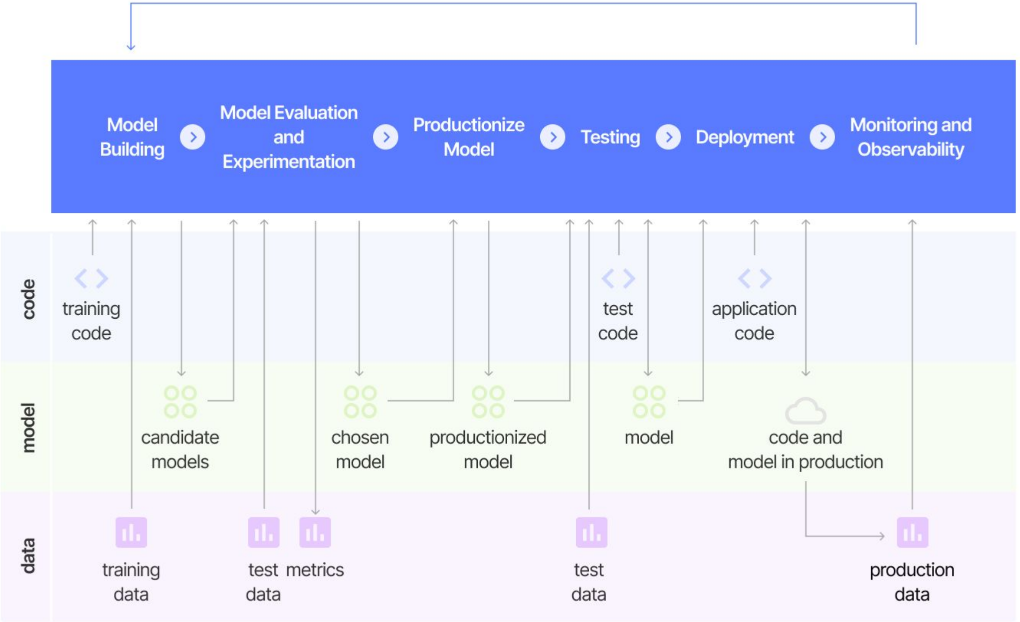
Superb AI

Principle 2 - Automate and Orchestrate All Data Flows

Superb AI

Continuous Integration and Continuous Delivery

- Automate deployment with CI/CD pipelines
- Discourage manual data wrangling
- Run the data flows using an orchestrator
    - Backfilling
    - Scheduling
    - Pipeline metrics
- OSS Tools: Airflow, Dagster, Prefect

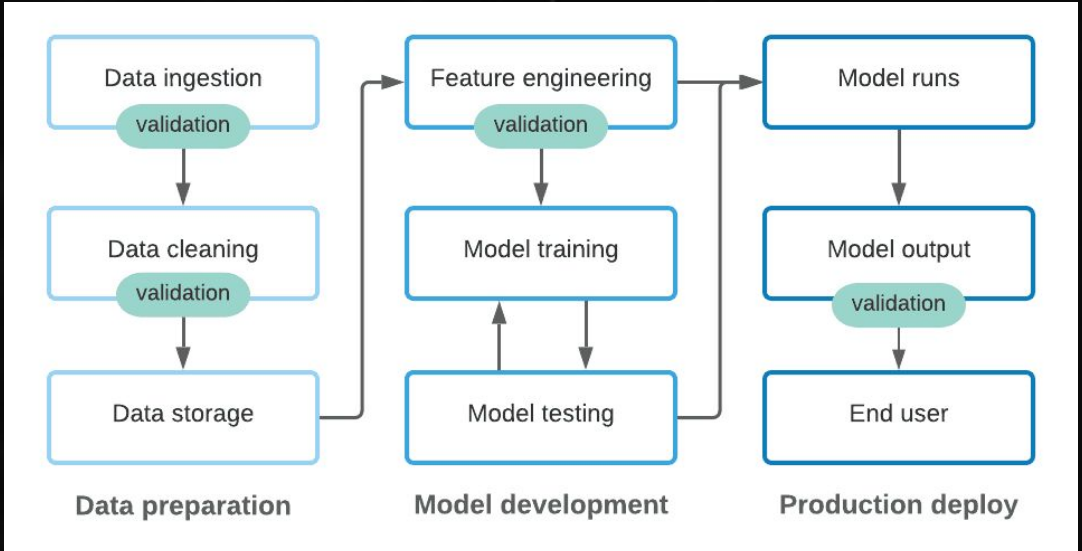


Source: Continuous Delivery for Machine Learning (by ThoughtWorks)

Principle 3 - Test Data Quality In All Stages of Data Lifecycle

Continuous Testing

- Test the data arriving from sources
  - Data unit tests
  - Schema/SQL/Streaming tests
- Validate data at different stages in the data flow
- Capture and publish metrics
- Reuse test tools across projects
- OSS Tool: great_expectations



Source: Why Data Quality Is Key to Successful MLOps (by Superconductive)

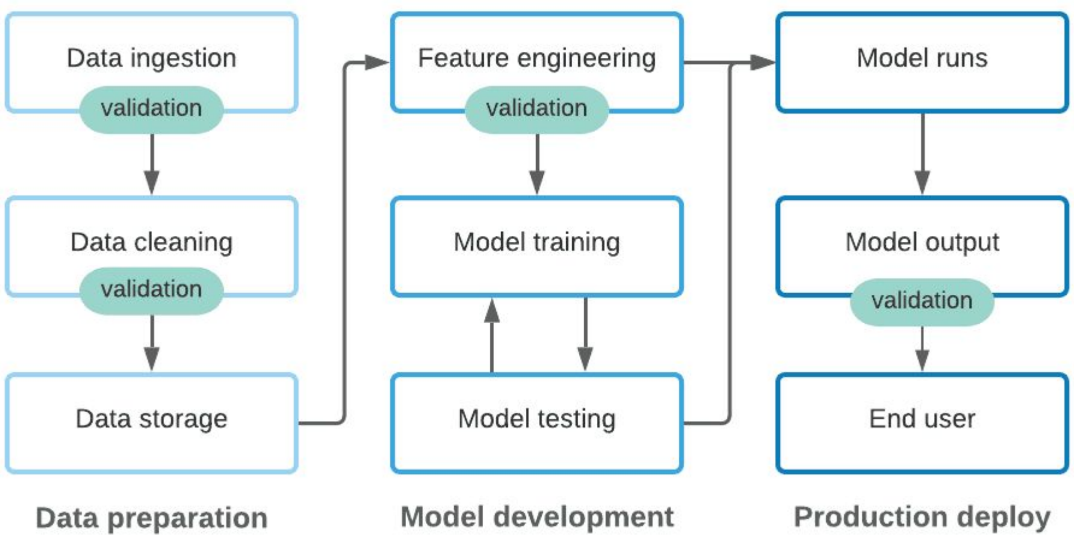
Superb AI

Principle 3 - Test Data Quality In All Stages of Data Lifecycle

Superb AI

Continuous Testing

- Test the data arriving from sources
  - Data unit tests
  - Schema/SQL/Streaming tests
- Validate data at different stages in the data flow
- Capture and publish metrics
- Reuse test tools across projects
- OSS Tool: great_expectations

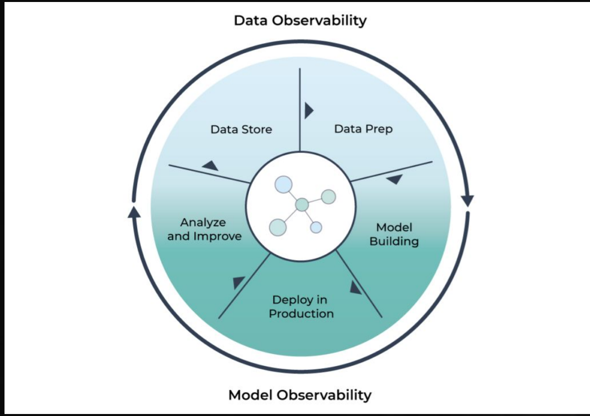


Source: [Why Data Quality Is Key to Successful MLOps]() (by Superconductive)

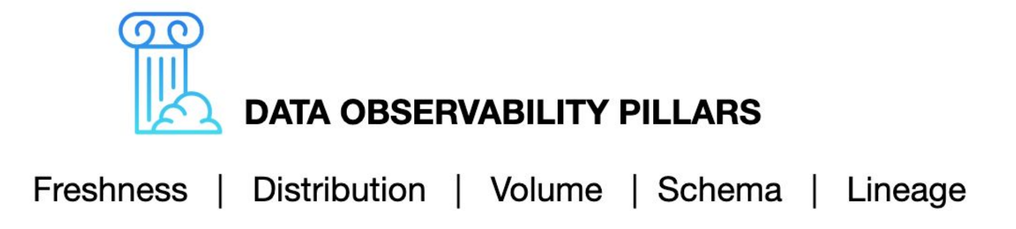
Principle 4 - Monitor Quality and Performance Metrics Across Data Flows

Improve Observability

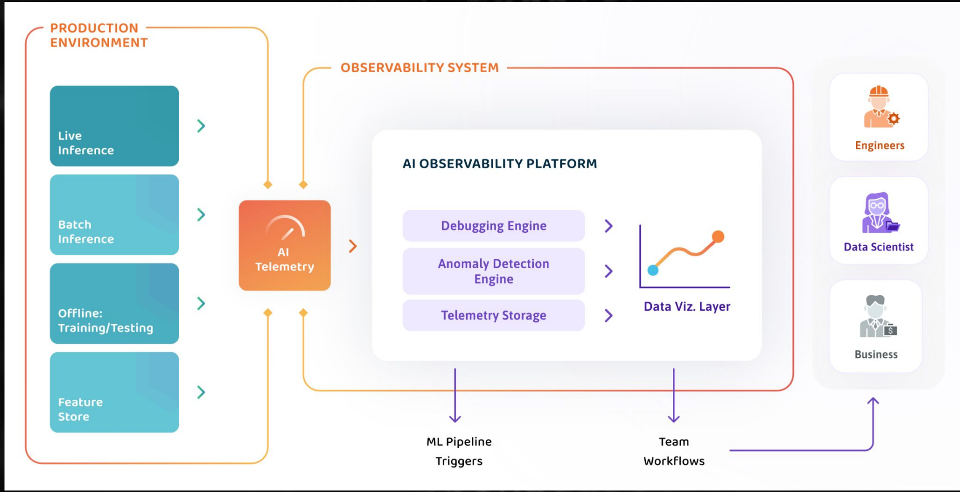
- Define data quality metrics
  - Technical metrics
  - Functional metrics
  - Performance metrics
- Visualize metrics
- Configure meaningful alerts



Source: *Beyond Monitoring: The Rise of Observability* (by Arize AI)



Source: *What is Data Observability?* (by Monte Carlo)



Source: *Anatomy of an Enterprise AI Observability Platform* (by WhyLabs)

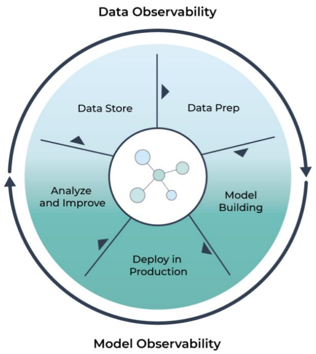
Superb AI

Principle 4 - Monitor Quality and Performance Metrics Across Data Flows

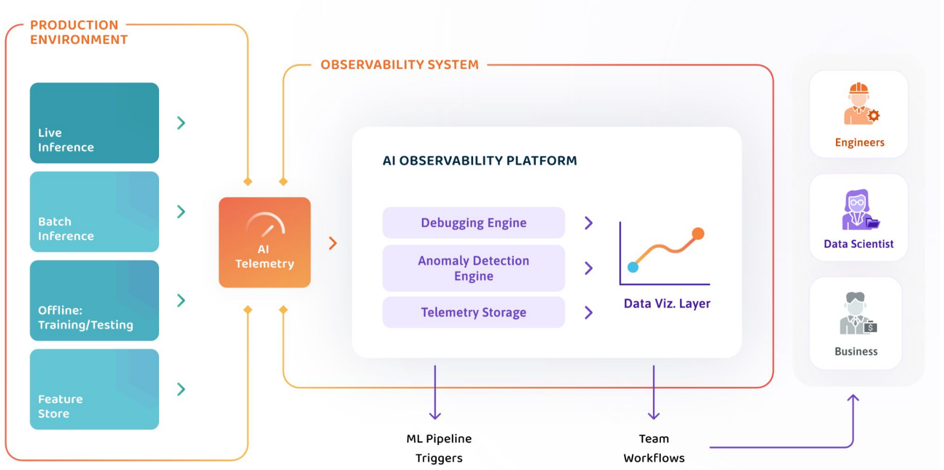
Improve Observability

- Define data quality metrics
  - Technical metrics
  - Functional metrics
  - Performance metrics
- Visualize metrics
- Configure meaningful alerts

Source: *[What is Data Observability?](#) (by Monte Carlo)*



Source: *[Beyond Monitoring: The Rise of Observability](#) (by Arize AI)*

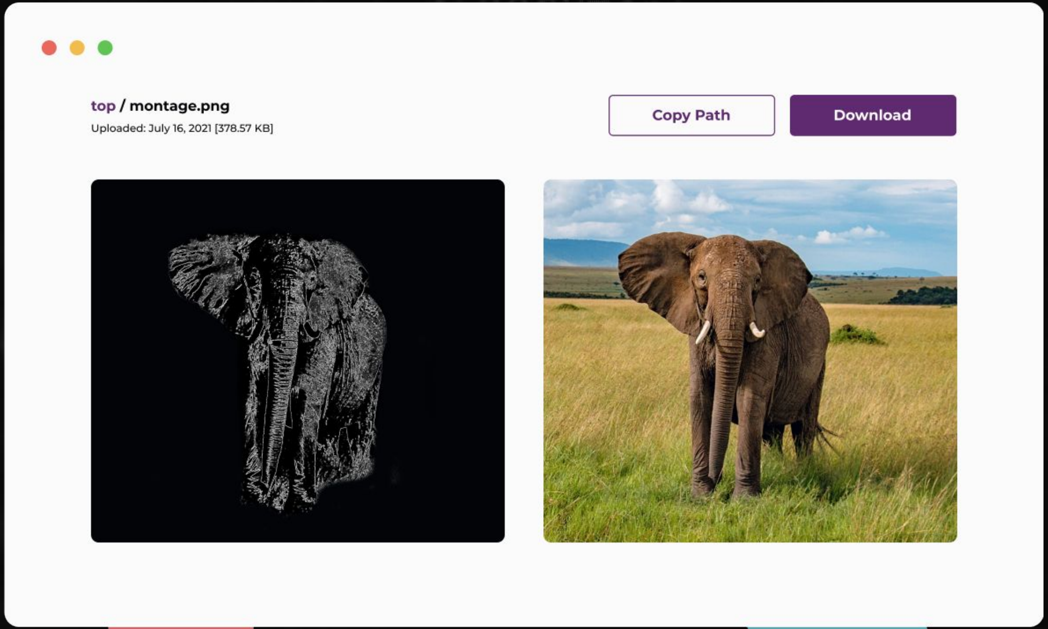


Source: *[Anatomy of an Enterprise AI Observability Platform](#) (by WhyLabs)*

Principle 5 - Build a Common Data and Metadata Model

Focus on Data Semantics

- Create a common data model
- Share the same terminology and schemas
  - Development teams
  - Data teams
  - Business teams
- Use a data catalog to share knowledge
- OSS Tools: dbt, Amundsen, DataHub, Marquez



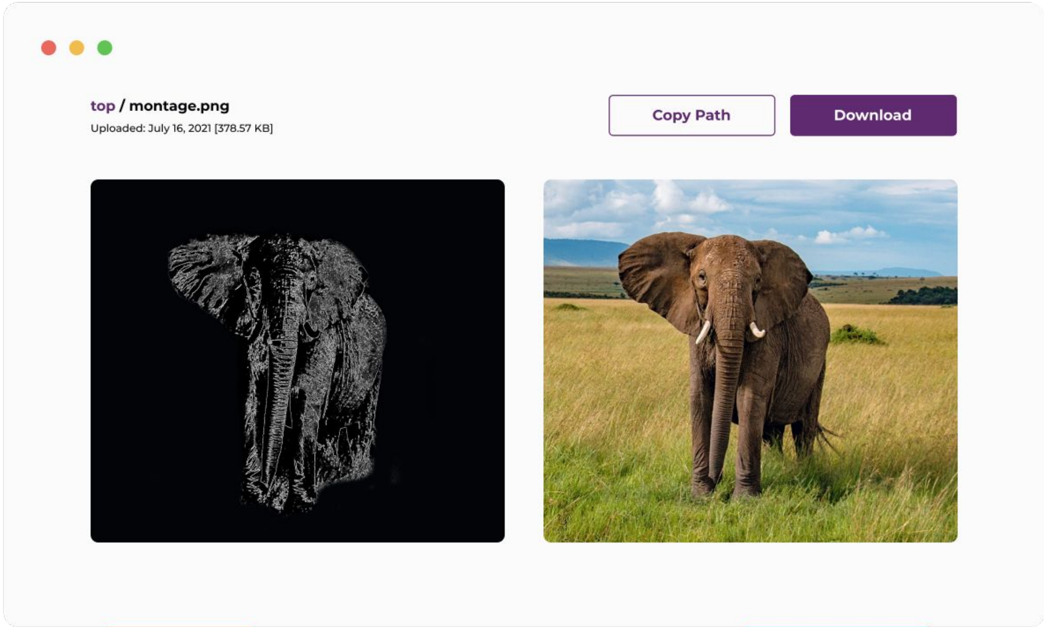
Source: Automated Data Versioning (by Pachyderm)

Principle 5 - Build a Common Data and Metadata Model

Superb AI

Focus on Data Semantics

- Create a common data model
- Share the same terminology and schemas
  - Development teams
  - Data teams
  - Business teams
- Use a data catalog to share knowledge
- OSS Tools: dbt, Amundsen, DataHub, Marquez

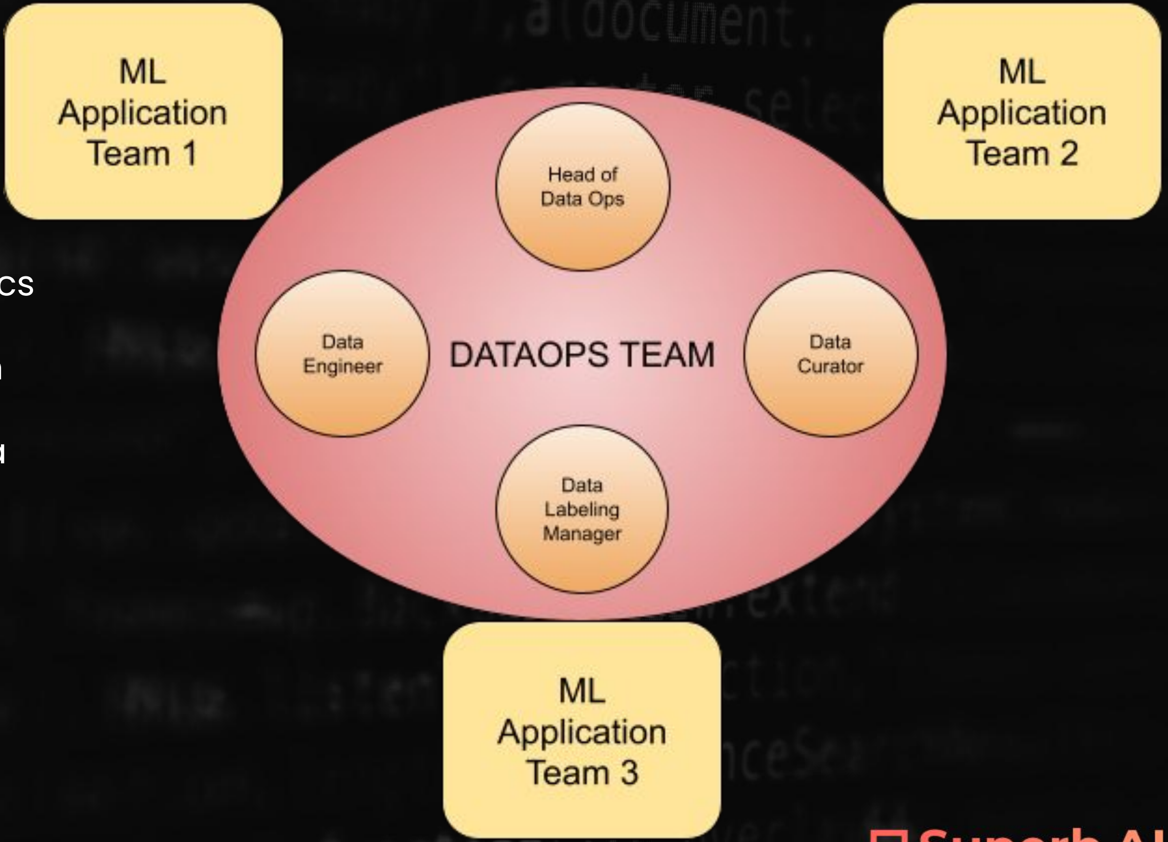


Source: Automated Data Versioning (by Pachyderm)

Principle 6 - Empower Collaboration Among Data Stakeholders

Cross-Functional Teams

- Use knowledge in cross-functional teams
  - Define important metrics and KPIs
  - Shared-objectives with business goals
- Remove bottlenecks for data usage
  - Self-service data monitoring
  - Democratize access to the data

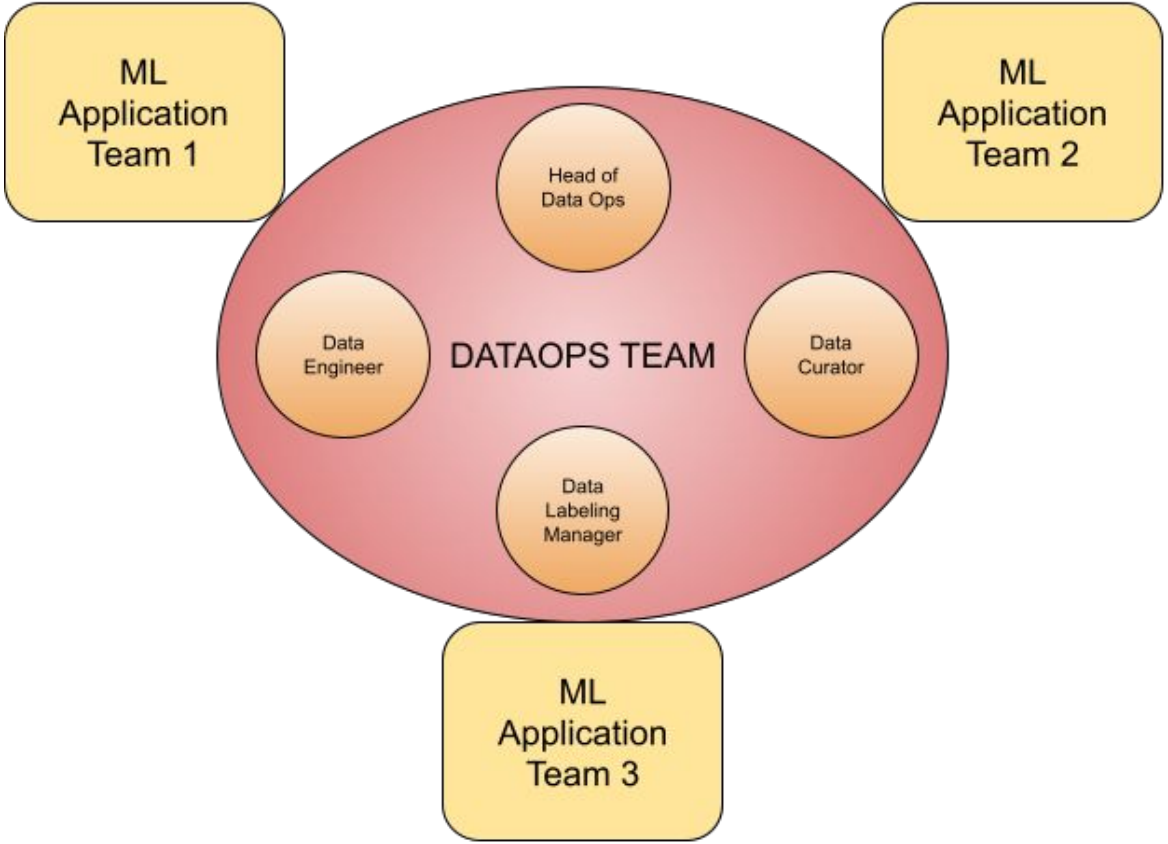


Principle 6 - Empower Collaboration Among Data Stakeholders

Superb AI

Cross-Functional Teams

- Use knowledge in cross-functional teams
  - Define important metrics and KPIs
  - Shared-objectives with business goals
- Remove bottlenecks for data usage
  - Self-service data monitoring
  - Democratize access to the data

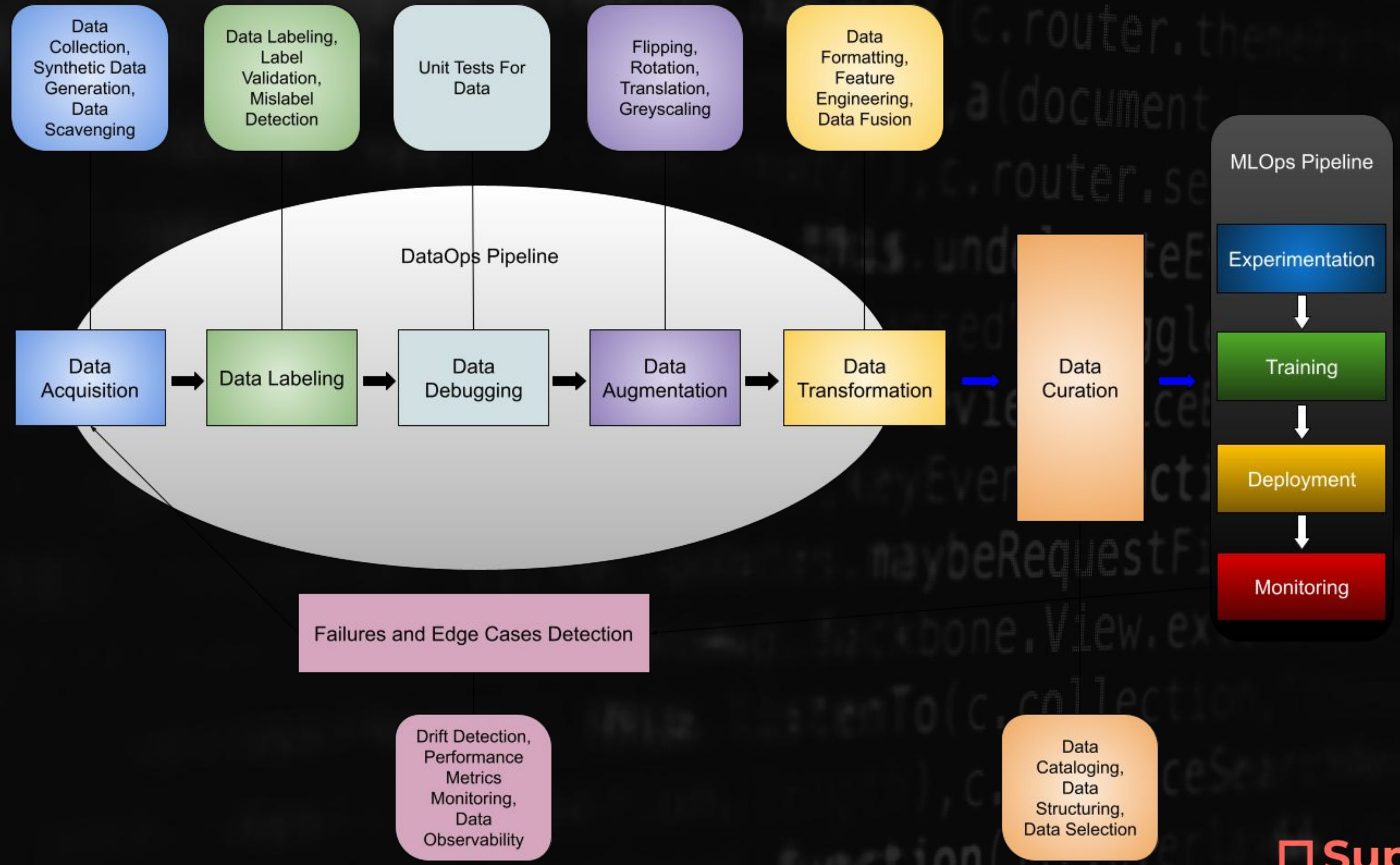


DataOps For

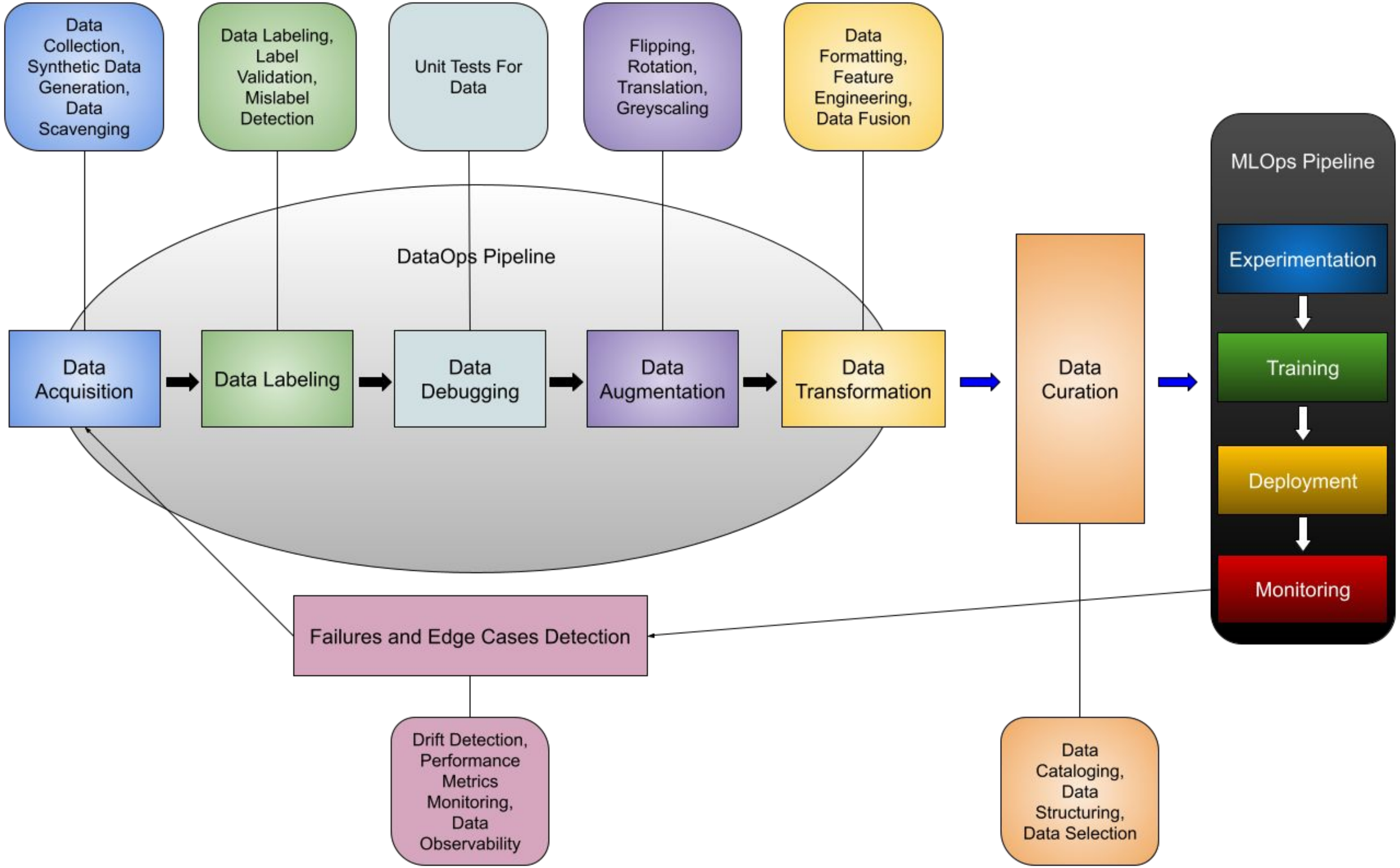
Computer Vision Stack?

Superb AI

DataOps For Computer Vision Stack?


Data Collection, Synthetic Data Generation, Data Scavenging
Data Labeling, Label Validation, Mislabel Detection
Unit Tests For Data
Flipping, Rotation, Translation, Greyscaling
Data Formatting, Feature Engineering, Data Fusion
DataOps Pipeline
Data Acquisition
Data Labeling
Data Debugging
Data Augmentation
Data Transformation
Data Curation
MLOps Pipeline
Experimentation
Training
Deployment
Monitoring
Failures and Edge Cases Detection
Drift Detection, Performance Metrics Monitoring, Data Observability
Data Cataloging, Data Structuring, Data Selection
Superb AI

Proposed DataOps for the Modern Computer Vision Stack



Key Data Challenges For

Computer Vision Teams

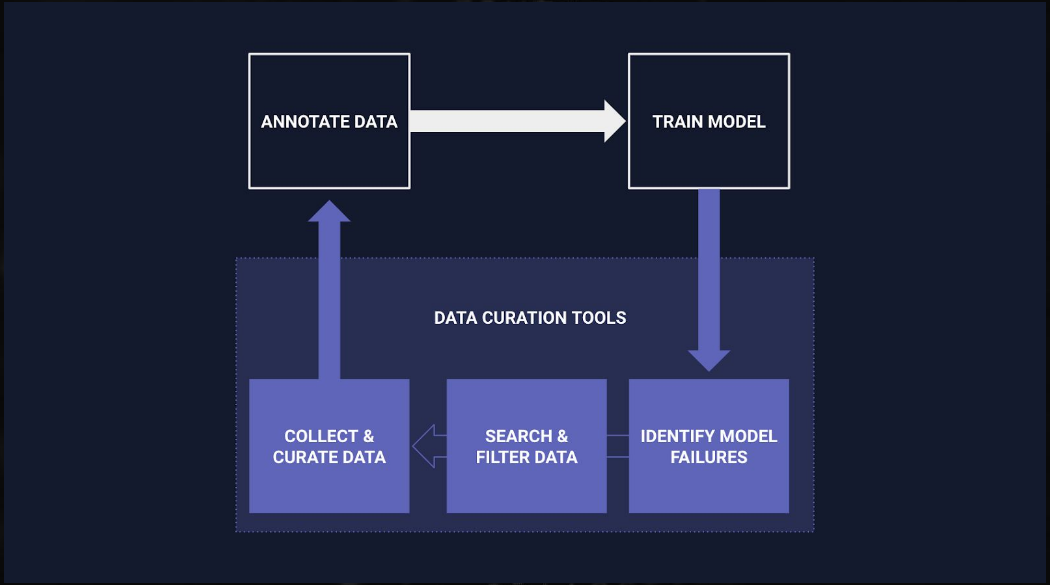
Superb AI

Key Data Challenges For Computer Vision Teams

Challenge 1: Curate High-Quality Data Points

Pain Points
1. Require domain knowledge
2. Can't deal with the 4 Vs of big data (Volume, Velocity, Variety, Veracity)
3. Narrow solutions

Solutions
1. Visualize massive datasets
2. Discover and retrieve data with ease
3. Curate diverse scenarios
4. Integrate seamlessly with existing workflows and tools

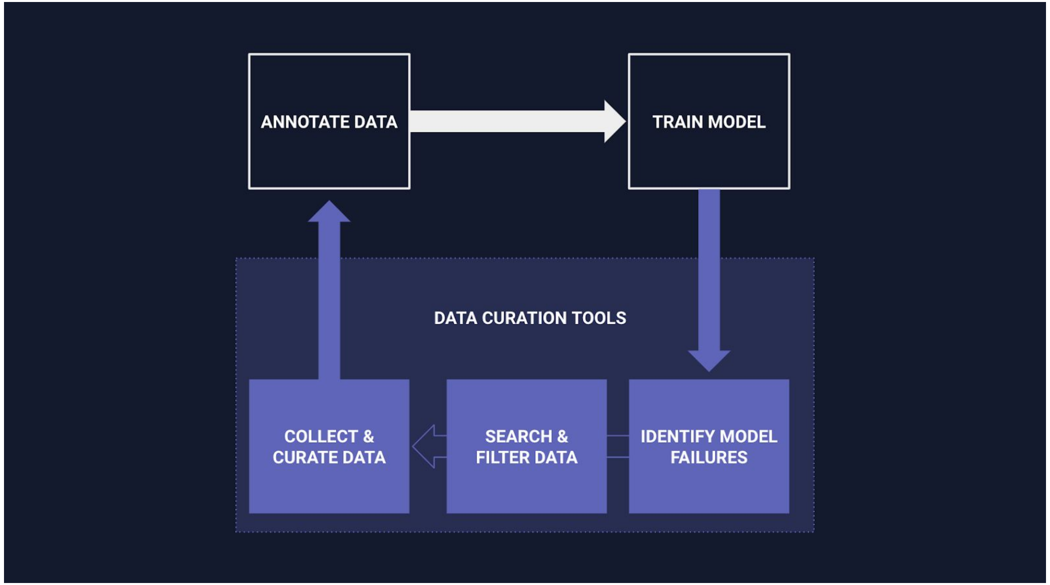


Source: The Best Data Curation Tools for Computer Vision (by Siasearch)

Challenge 1: Curate High-Quality Data Points

Superb AI

- Pain Points
  - Require domain knowledge
  - Can't deal with the 4 Vs of big data
  - Narrow solutions
- Solutions
  - Visualize massive datasets
  - Discover and retrieve data with ease
  - Curate diverse scenarios
  - Integrate seamlessly with existing workflows and tools



Source: The Best Data Curation Tools for Computer Vision (by Siasearch)

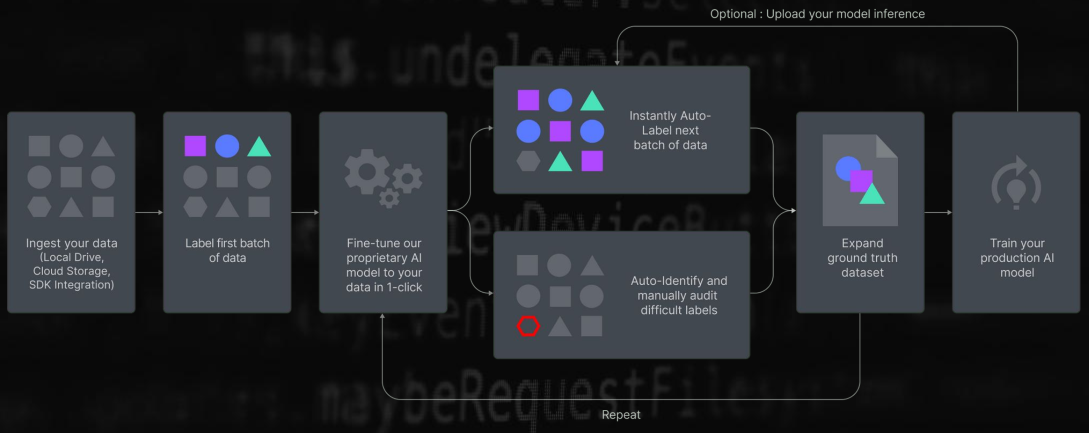
Challenge 2: Label and Audit Data at Massive Scale

Pain Points

1. Manual labeling and quality assurance is painfully slow
2. Label quality is bad when dealing with domain-specific datasets and hard edge cases

Solutions

1. Automatically label data
2. Identify and audit hard labels
3. Use active learning for human verification of labels



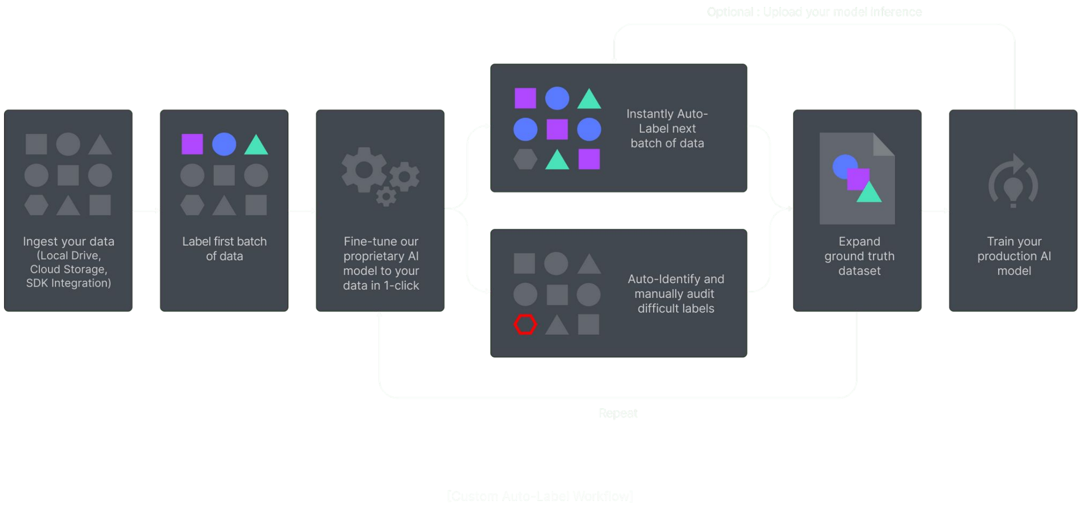
[Custom Auto-Label Workflow]

Source: Automate Data Preparation for Computer Vision (by Superb AI)

Superb AI

Challenge 2: Label and Audit Data at Massive Scale

- Pain Points
  - Manual labeling and quality assurance is painfully slow
  - Label quality is bad when dealing with (1) domain-specific datasets and (2) hard edge cases
- Solutions
  - Automatically label data
  - Identify and audit hard labels
  - Use active learning for human verification of labels



Source: *[Automate Data Preparation for Computer Vision](#) (by Superb AI)*

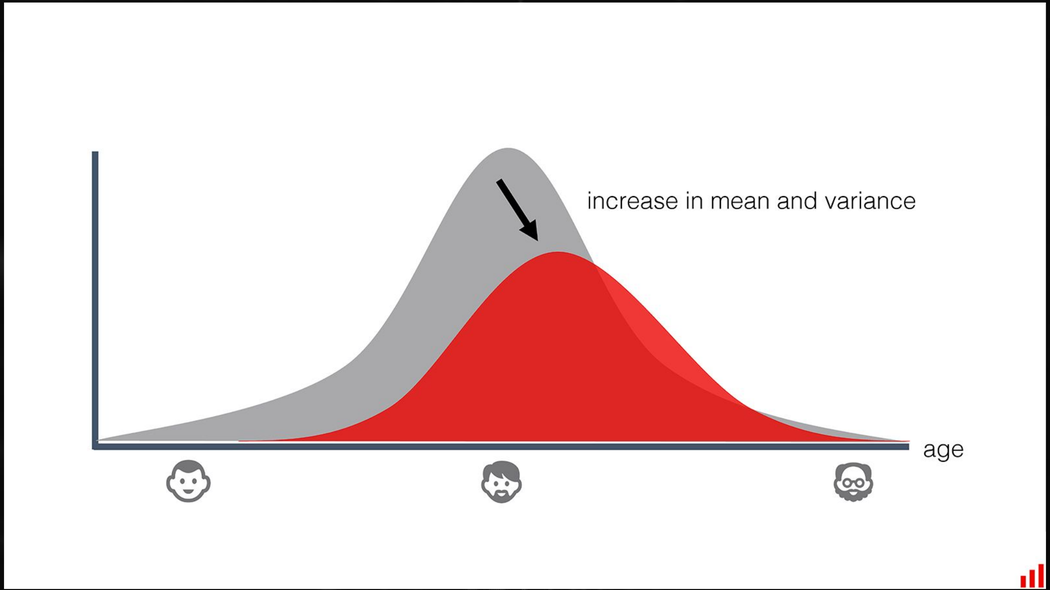
Challenge 3: Account For Data Drift

Pain Points

1. Upstream process changes
2. Data quality issues
3. Natural drift in the data
4. Covariate shift

Solutions

1. Detect data drifts and raise alerts
2. Analyze where and why drift happens
3. Adapt to drift and improve model performance

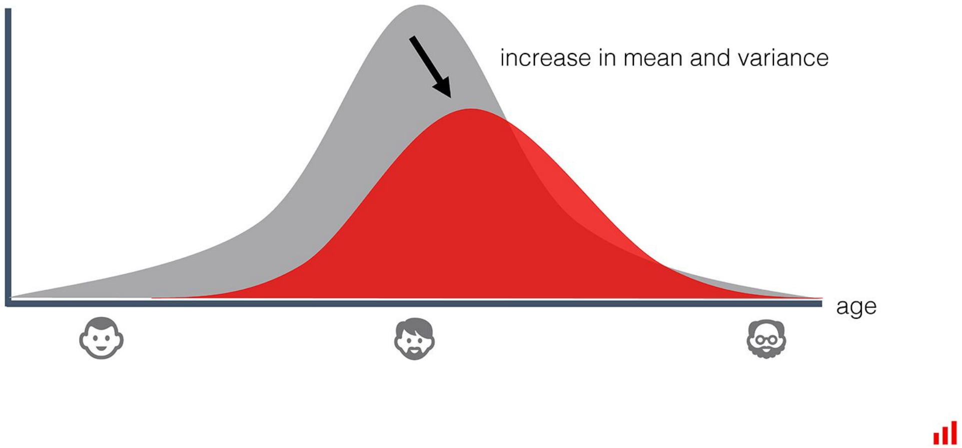


Source: Why Should You Care About Data and Concept Drift (by Evidently AI)

Challenge 3: Account for Data Drift

Superb AI

- Pain Points
  - Upstream process changes
  - Data quality issues
  - Natural drift in the data
  - Covariate shift
- Solutions
  - Detect data drifts and raise alerts
  - Analyze where and why drift happens
  - Adapt to drift and improve model performance



Source: Why Should You Care About Data and Concept Drift (by Evidently AI)

The Future Of The

Modern Computer Vision Stack

Superb AI

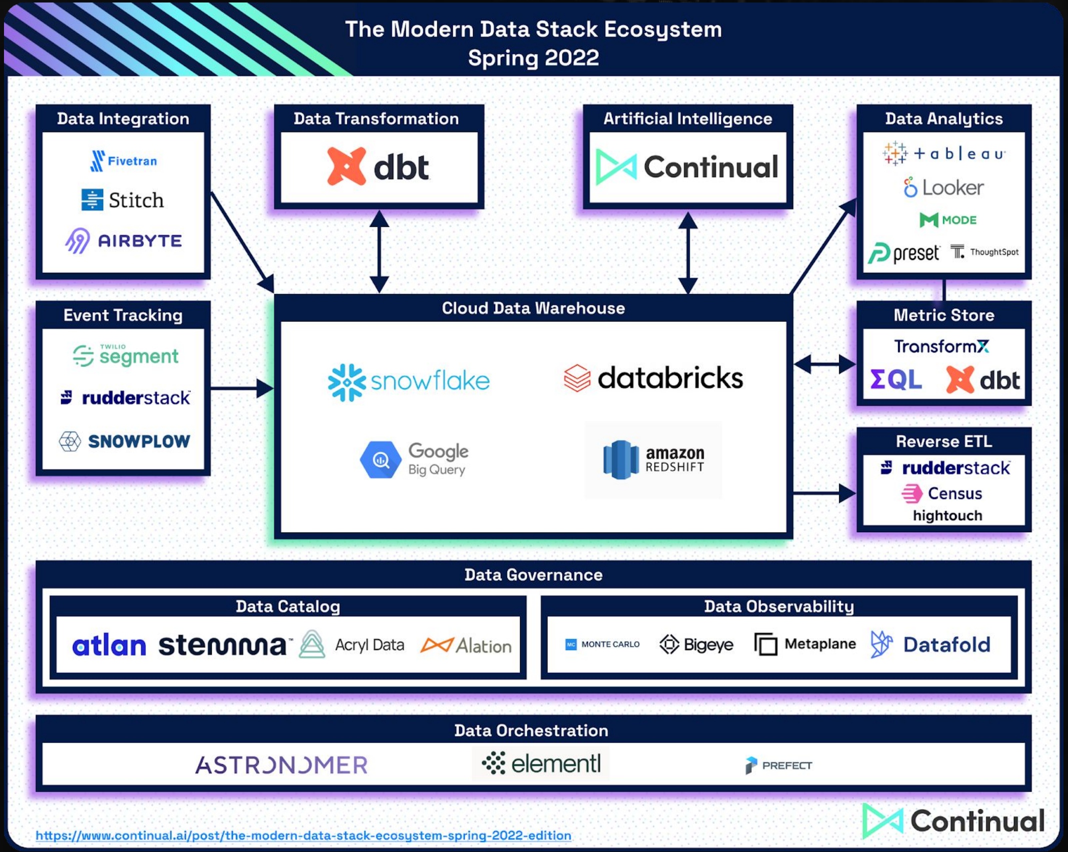
The Future of The Modern Computer Vision Stack

Following The Footsteps of The Modern Data Stack

The Modern Data Stack is a collection of cloud-native tools centered around a cloud data warehouse.

Benefits:

1. Ease of Use
2. Wide Adoption
3. Automation
4. Cost



Source: The Modern Data Stack Ecosystem - Spring 2022 Edition (by Continual)

Superb AI

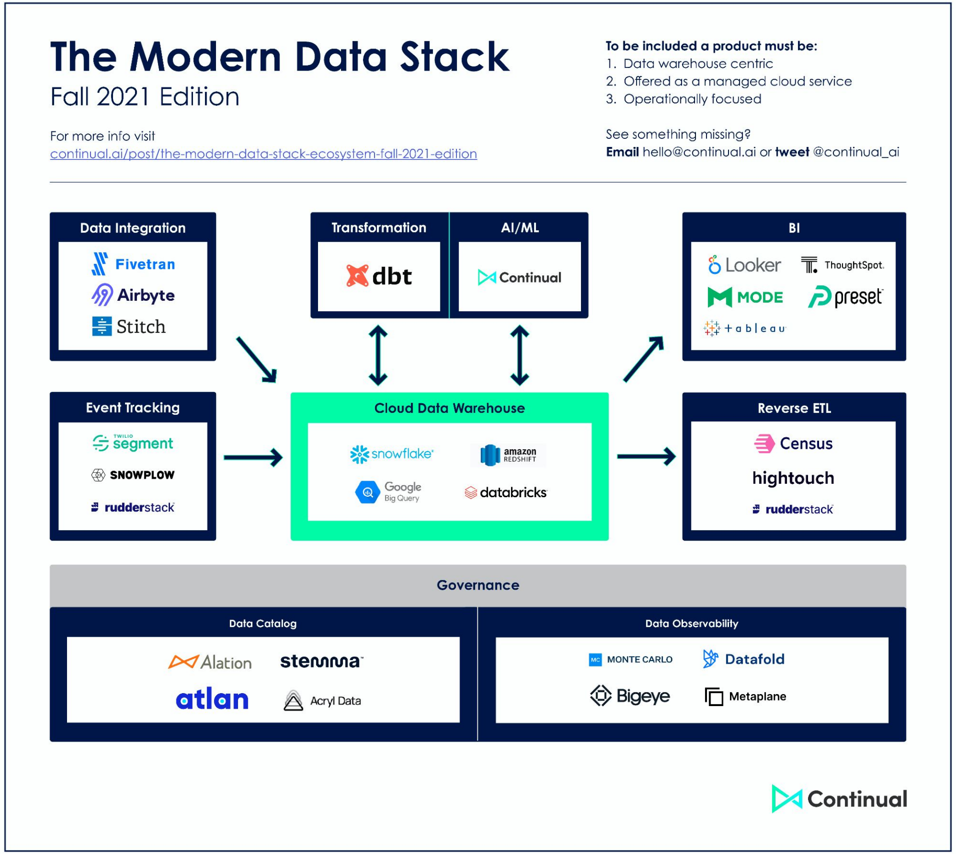
Following The Footstep of "The Modern Data Stack"

Superb AI

The Modern Data Stack is a collection of cloud-native tools centered around a cloud data warehouse.

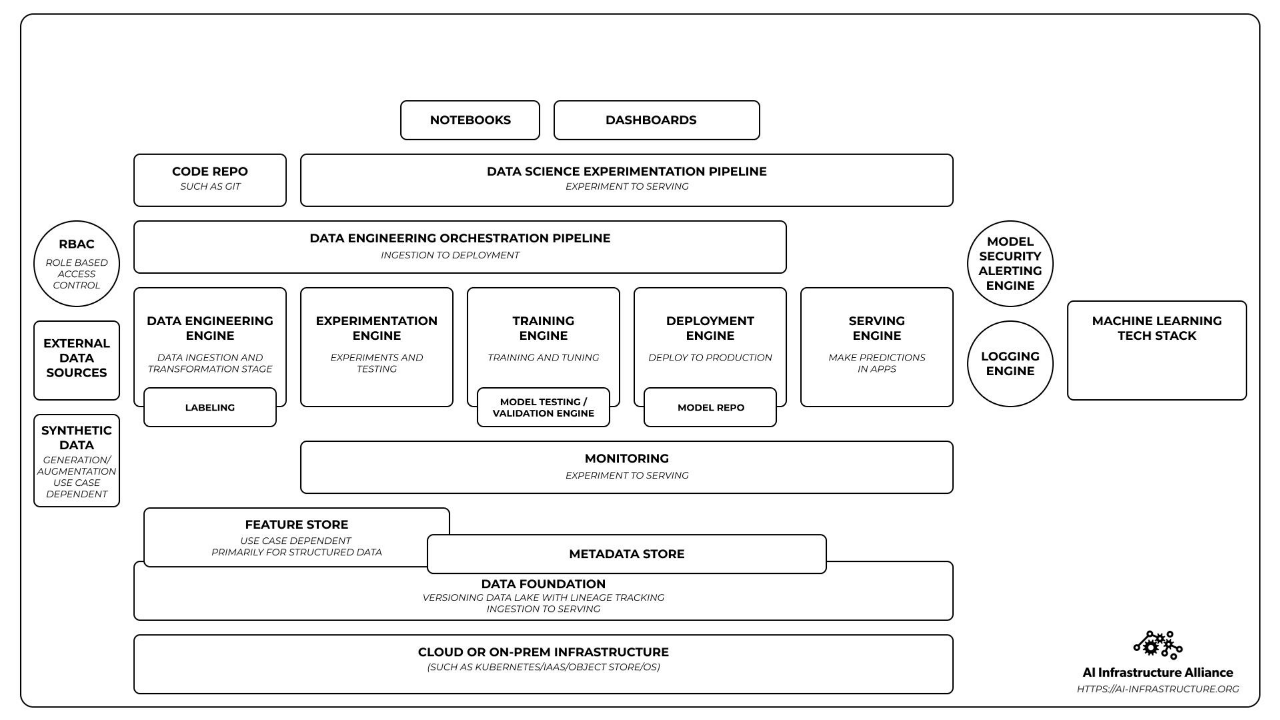
Benefits:

1. Ease of Use
2. Wide Adoption
3. Automation
4. Cost



Source: *The Modern Data Stack Ecosystem - Fall 2021 Edition (by Continual)*

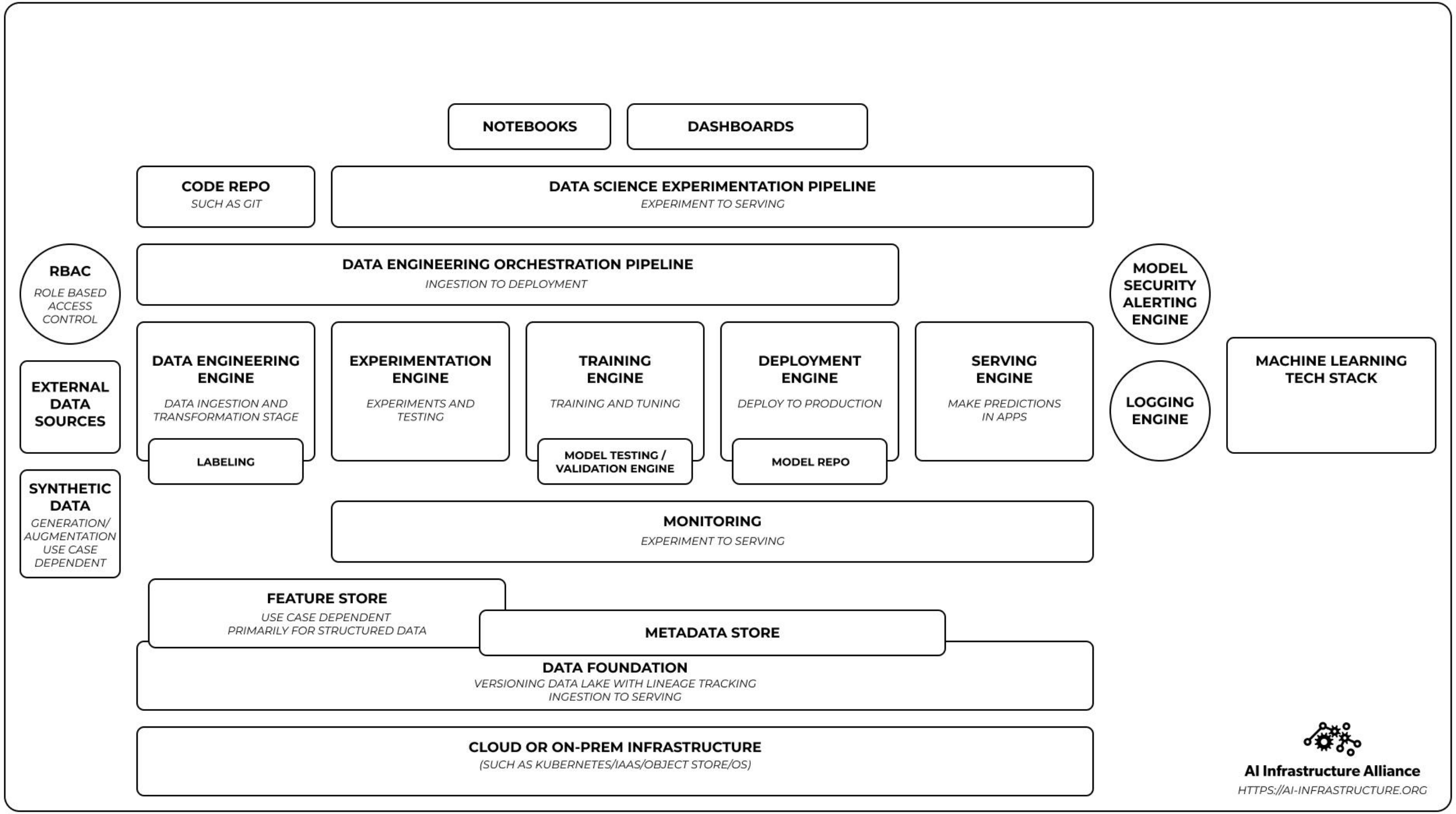
The Canonical Stack for ML



Source: The Rapid Evolution of the Canonical Stack for Machine Learning (by Daniel Jeffries)

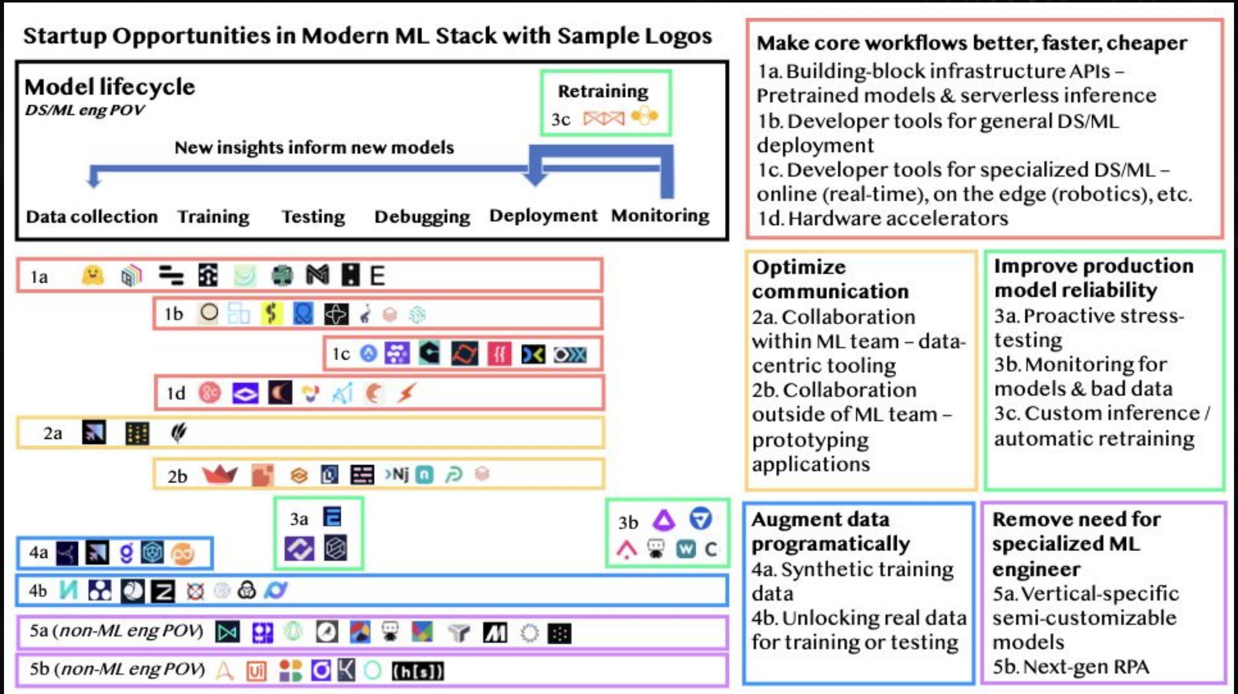
Superb AI

The Canonical Stack for Machine Learning



Source: The Rapid Evolution of the Canonical Stack for Machine Learning
(by Daniel Jeffries)

Startup Opportunities in ML Infrastructure

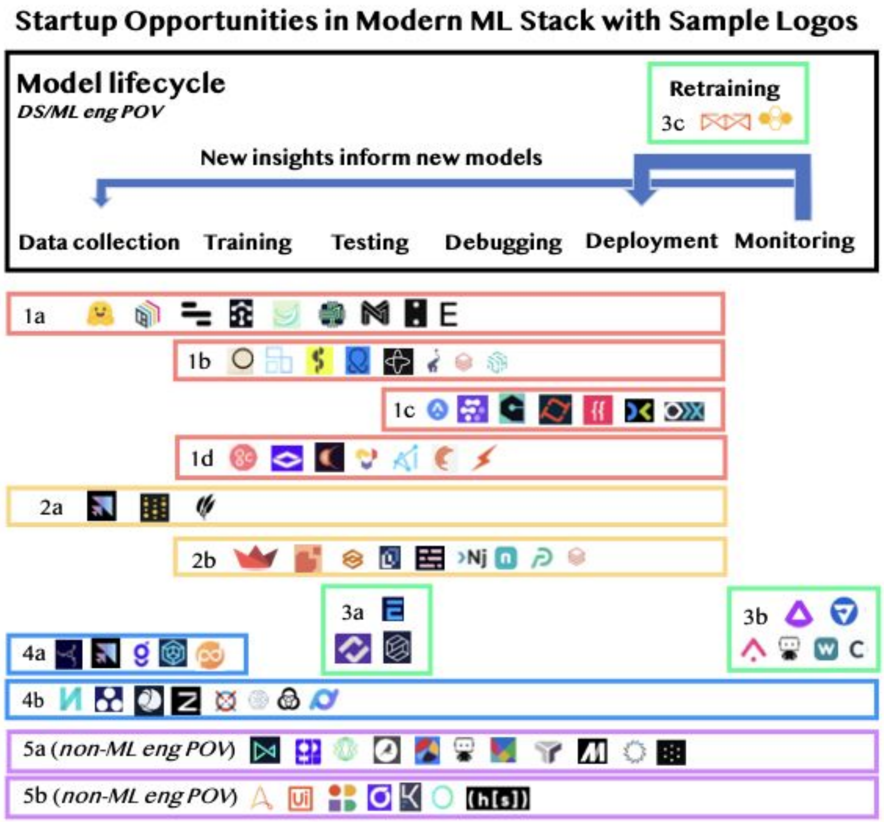


Source: *Startup Opportunities in ML Infrastructure* (by Leigh-Marie Braswell)

Superb AI

Startup Opportunities in Machine Learning Infrastructure

Superb AI



Make core workflows better, faster, cheaper
1a. Building-block infrastructure APIs – Pretrained models & serverless inference
1b. Developer tools for general DS/ML deployment
1c. Developer tools for specialized DS/ML – online (real-time), on the edge (robotics), etc.
1d. Hardware accelerators

Optimize communication
2a. Collaboration within ML team – data-centric tooling
2b. Collaboration outside of ML team – prototyping applications

Improve production model reliability
3a. Proactive stress-testing
3b. Monitoring for models & bad data
3c. Custom inference / automatic retraining

Augment data programmatically
4a. Synthetic training data
4b. Unlocking real data for training or testing

Remove need for specialized ML engineer
5a. Vertical-specific semi-customizable models
5b. Next-gen RPA

Source: [Startup Opportunities in ML Infrastructure](#) (by Leigh-Marie Braswell)

Thank you!

James Le
Website: jameskle.com
Twitter: @le_james94
Email: james.le@superb-ai.com

Superb AI

Thank you!

James Le
Website: jameskle.com
Twitter: @le_james94
Email: james.le@superb-ai.com