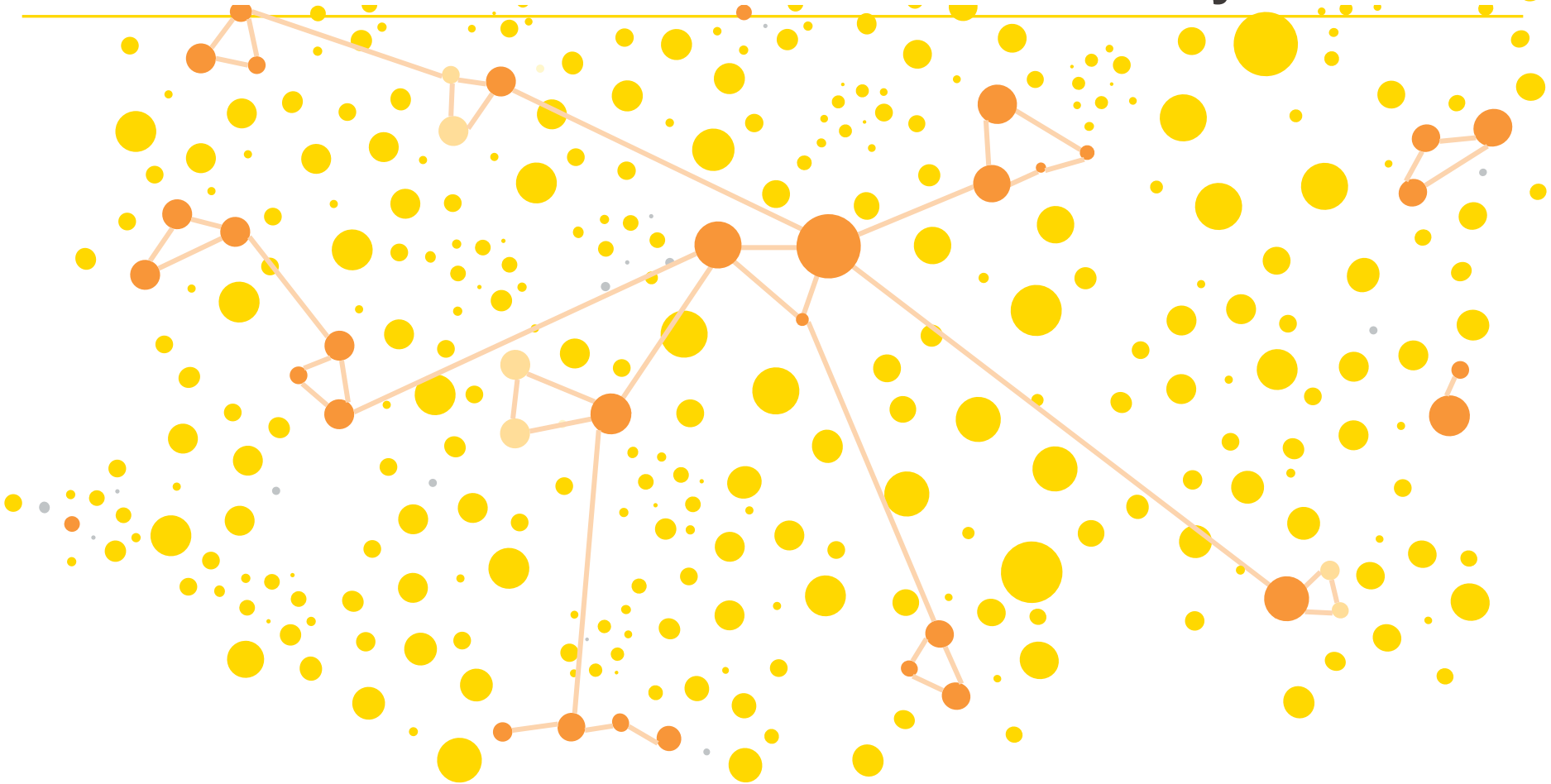
From Creating to Disseminating Data Science

Michael Berthold

KNIME Spring Summit 2023
April, 18, 2023

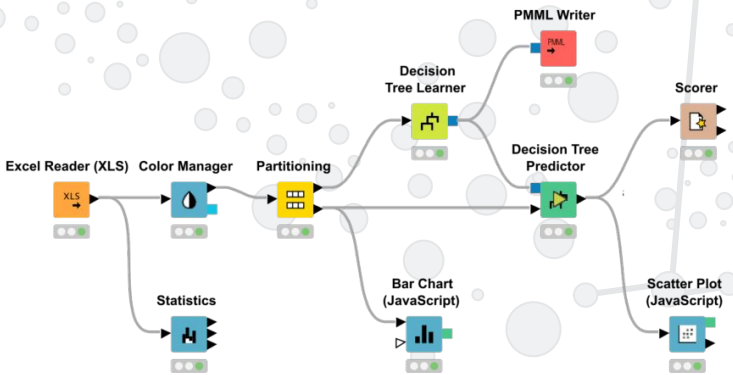
The Ultimate Goal: Data Driven Decisions Everywhere



Open for Innovation KNIME

How Do We Get There?

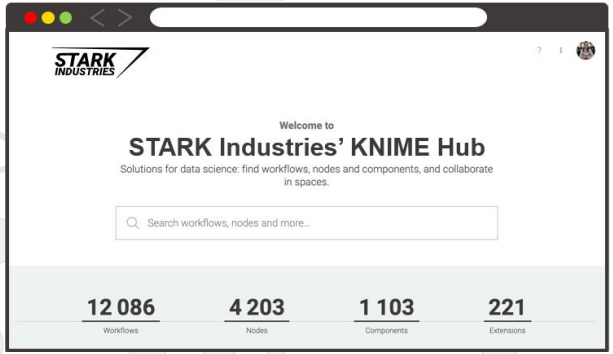
Enabling Experts to Collaborate

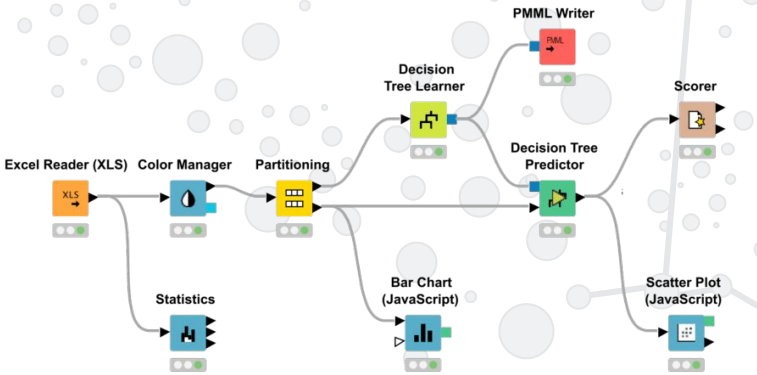


Low Code / No Code
Visual Workflows for:

- Data Engineers
- Machine Learning / AI Experts
- Visualization Gurus
- Python, R, and other Coders

Enabling Others in the Organization

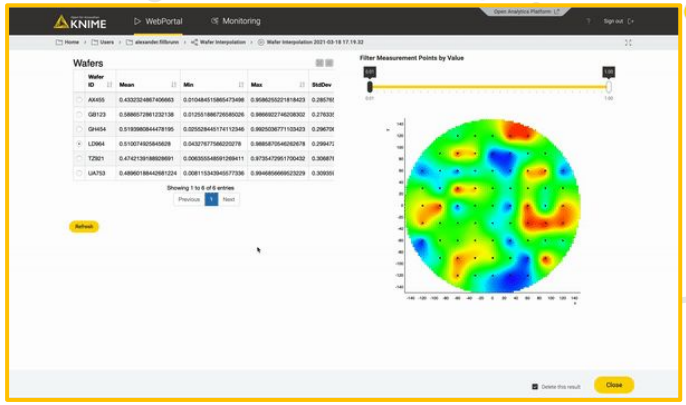




No Code Visual Workflows for:

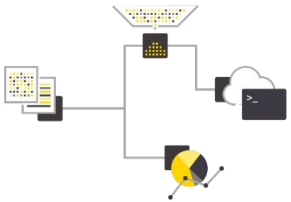
- Expertise Sharing
- Spreading Data Literacy
- …getting started!

Delivering Insights to Colleagues



Deploy Visual Workflows as
- Interactive DataApps
- Application Services

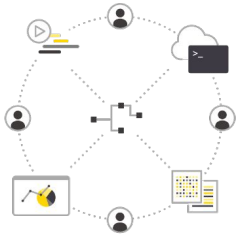
Intermezzo: KNIME Software



KNIME Analytics Platform

FREE, OPEN SOURCE

Open-source software for creating data science. Intuitive, open, and continuously integrating new developments, KNIME makes understanding data and designing data science workflows and reusable components accessible to everyone.



KNIME Business Hub

COMMERCIAL

Enterprise software for team-based collaboration, automation, management, and deployment of data science workflows as analytical applications and services. Non experts are given access to data science via KNIME WebPortal or can use REST APIs.

Create

Blend & Transform

Access, merge, and transform all of your data

Model & Visualize

Make sense of your data with the tools you choose

Productionize

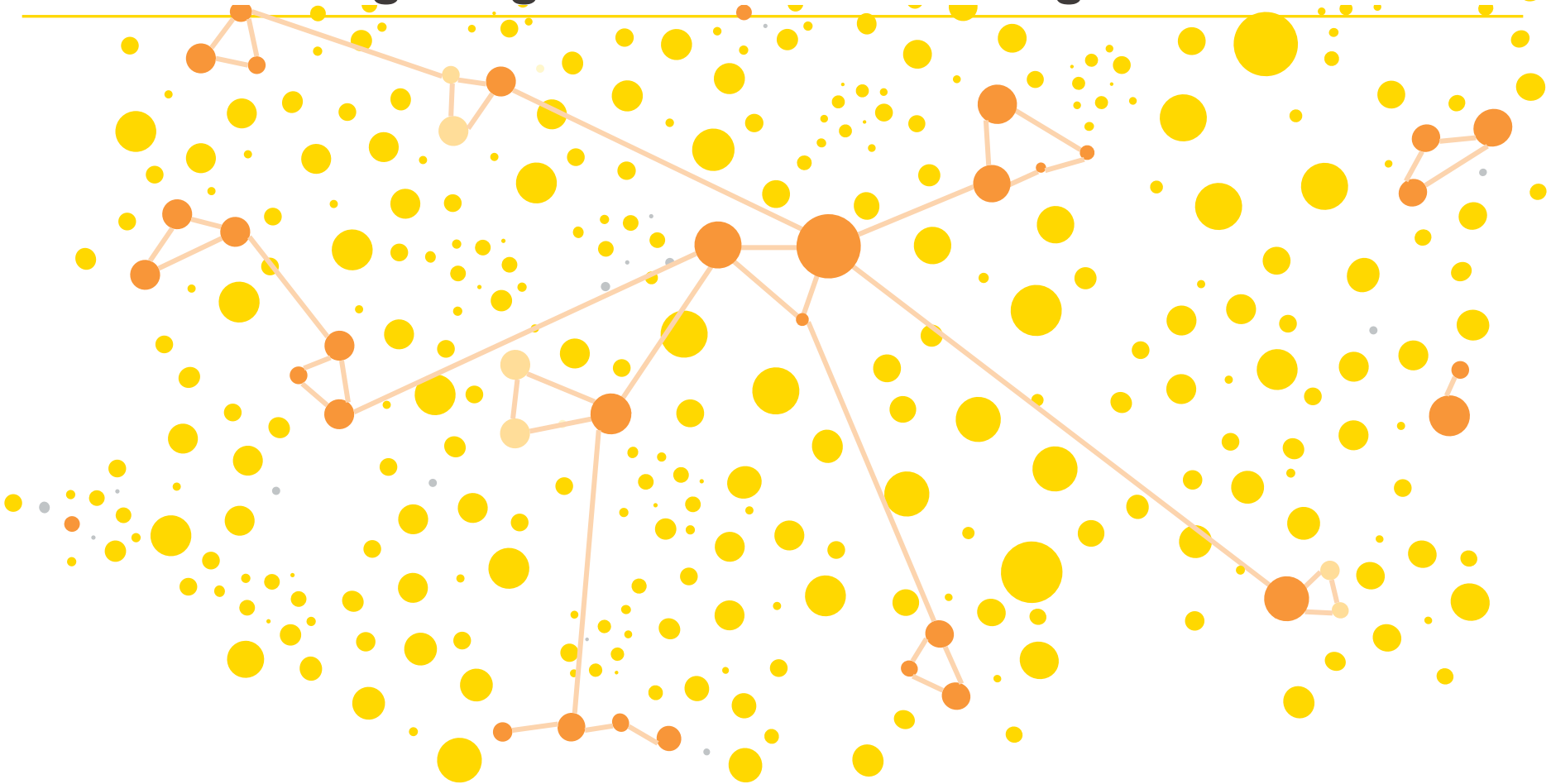
Deploy & Manage

Support enterprise-wide data science practices

Consume & Interact

Leverage insights gained from your data

Disseminating Insights Across Entire Organizations?



Open for Innovation KNIME

Deploying Data Science to Production: Status Quo

Scenario Wild West:

- Create Wrapper - Share it somewhere/somehow - Use it (if you can find it)

Scenario IT:

- Send Model to IT - IT recodes - IT deploys - IT forgets

Scenario Traditional Coding Practice:

- Make Data Science Production part of Software Engineering Department

- (Some) Issues:
  - Translation Errors & lack of functionality (model types & data preprocessing)
  - Long Deployment Cycle
  - Governance often incomplete/missing

Data Science at Scale: Requirements

- Automated **Packaging of all Dependencies** (“Integration”)
- Automated and/or Expert **Validation** of DS going into Production
- Automated move into Production (“**Deployment**”)
- **Archive** of history enables **Audit**, **Rollback**, and **Explainability**
- **Monitoring** of Data Science in Production
- Automatic **Retraining** when Reality shifts

Sounds Like Software Development & Deployment?

Fully Automated CI/CD Pipelines:

- Continuous Integration:
  Putting all the Pieces together
- Continuous Delivery/Deployment:
  Reliably (re)placing the new Version in Production

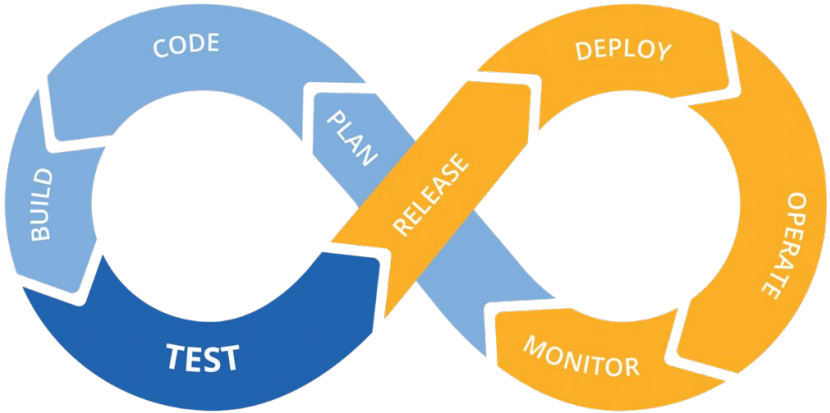


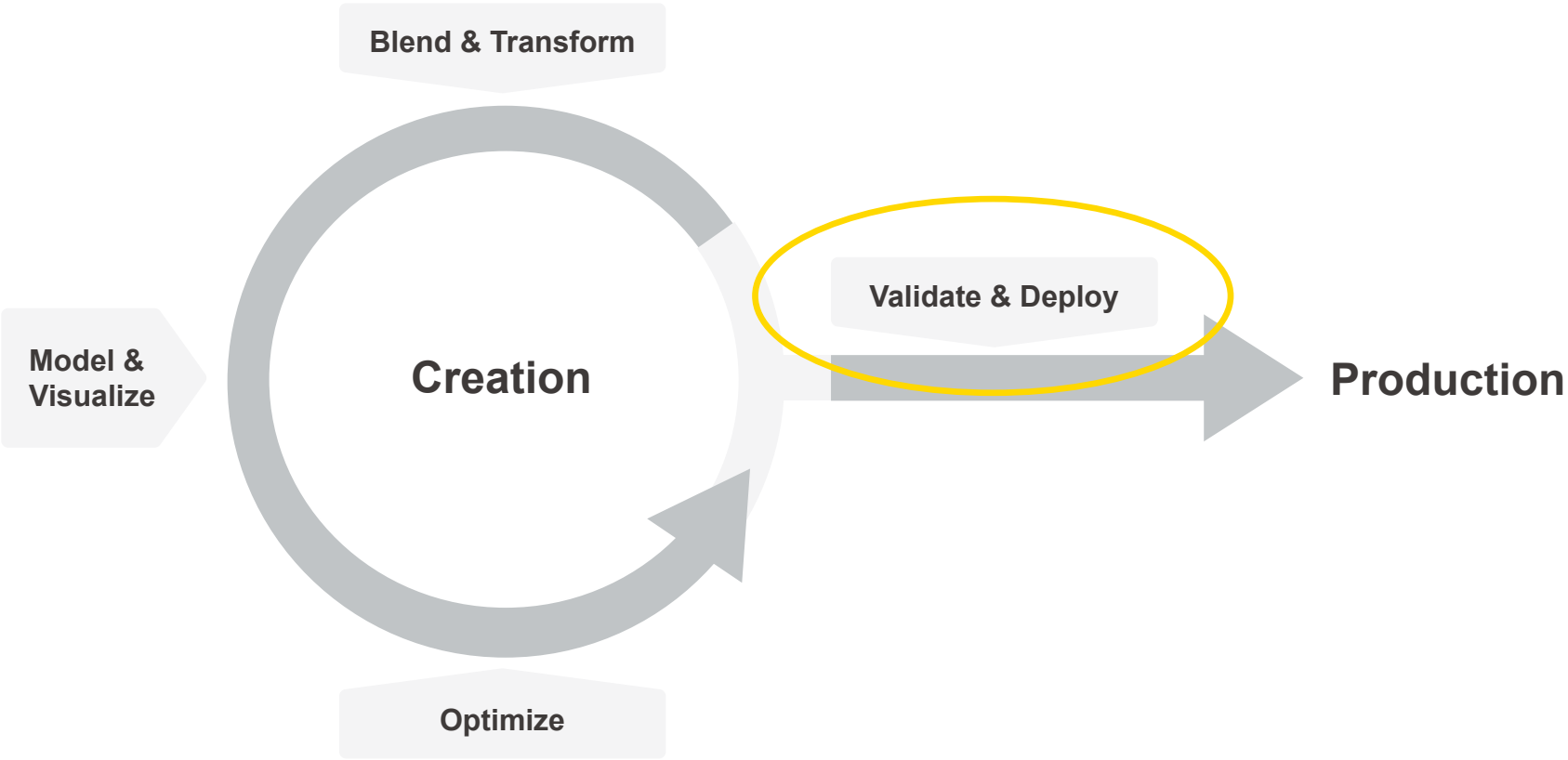
Image: Github

Data Science at Scale: Requirements

- Automated **Packaging of all Dependencies** (“Integration”)
- Automated and/or Expert **Validation** of DS going into Production
- Automated move into Production (“**Deployment**”)
- **Archive** of history enables **Audit**, **Rollback**, and **Explainability**
- **Monitoring** of Data Science in Production
- Automatic **Retraining** when Reality shifts

Step 1:
Validation and Deployment

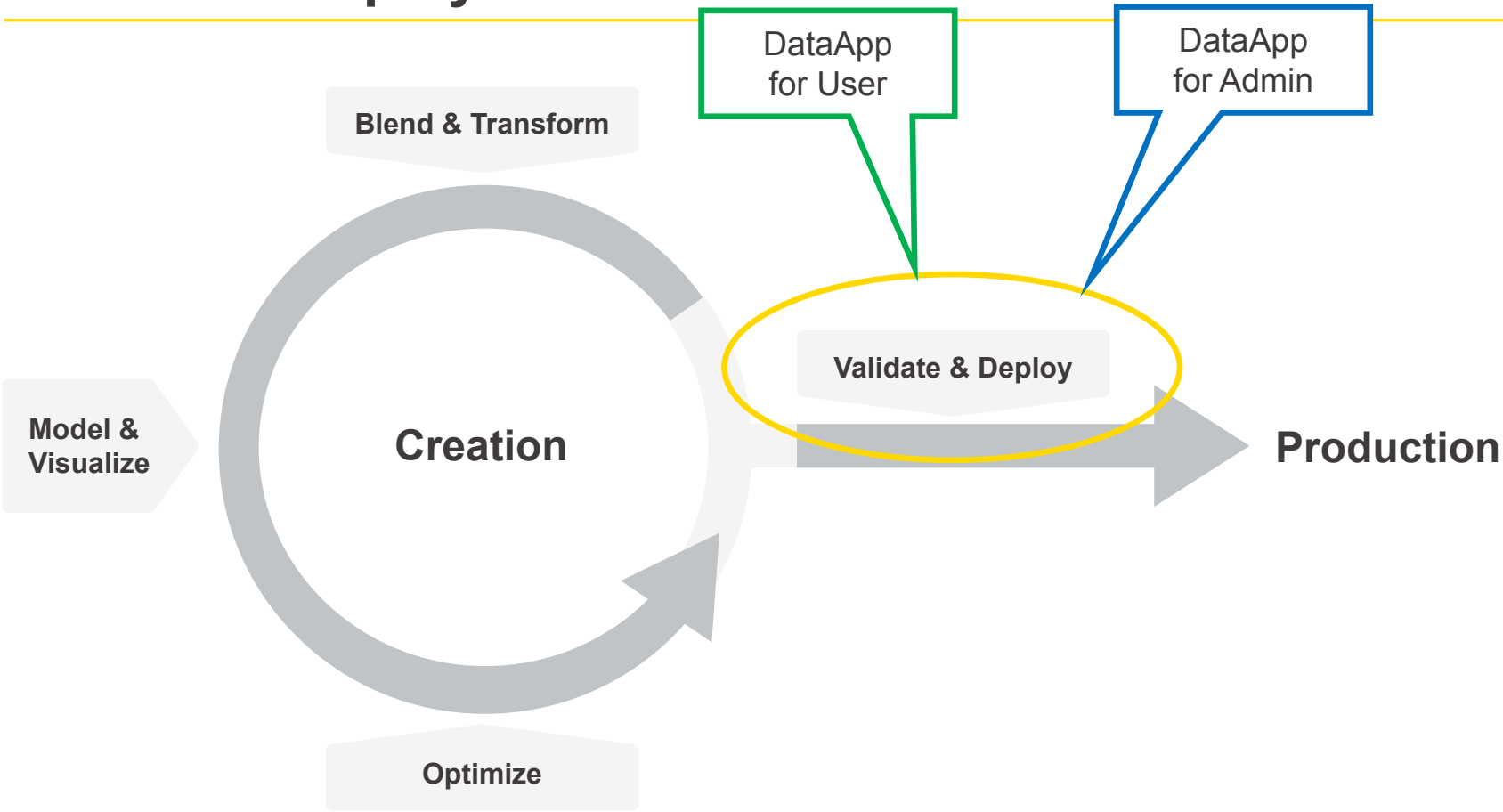
Validated Deployment



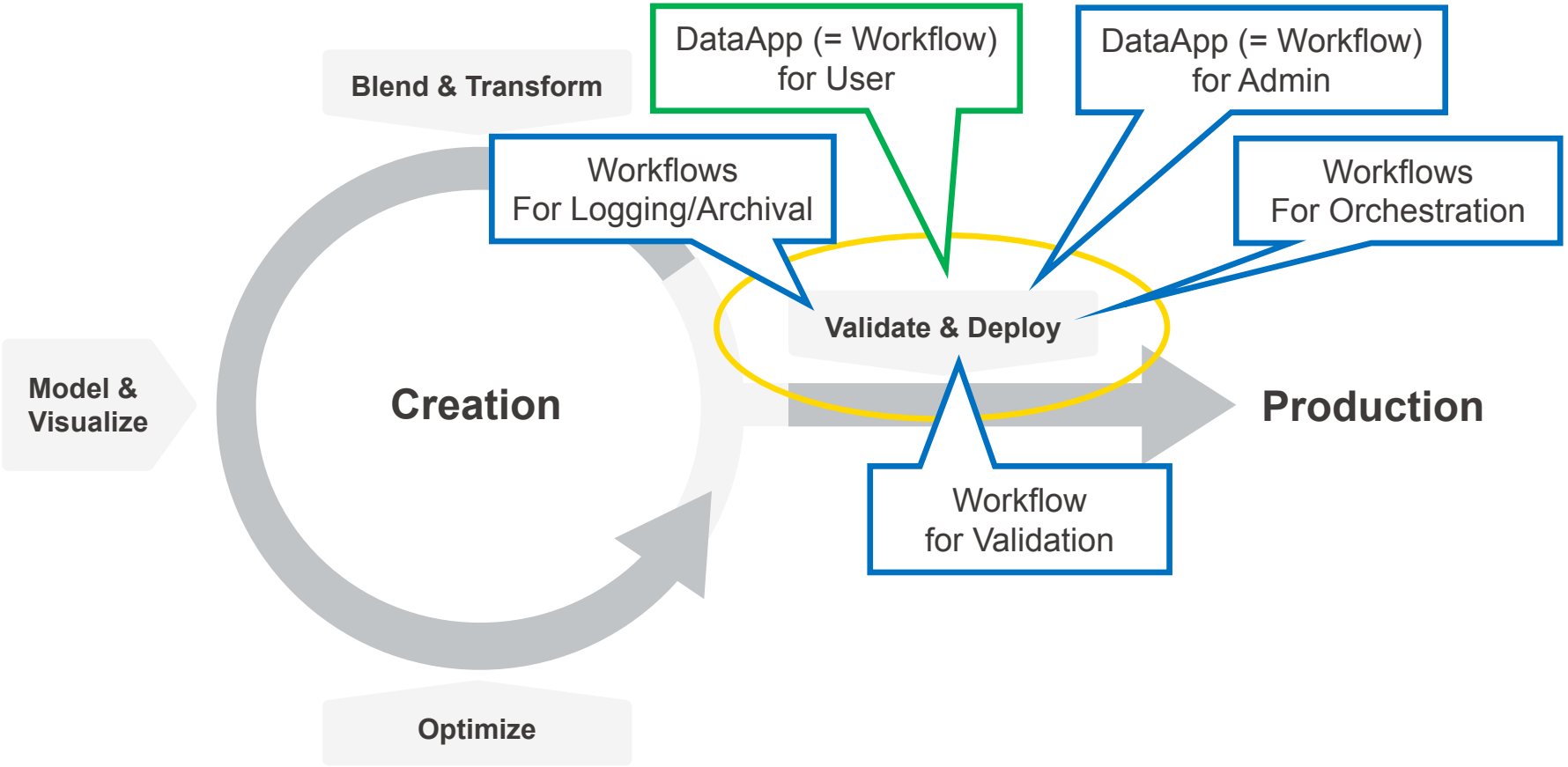
The CDDS Extension for KNIME Business Hub

Validated Deployment of KNIME Workflows

Validated Deployment



Validated Deployment: Fully Customizable



Validated Deployment

Usable out of the box (customizable setup via installation wizard)

…but also: everything else is also completely customizable

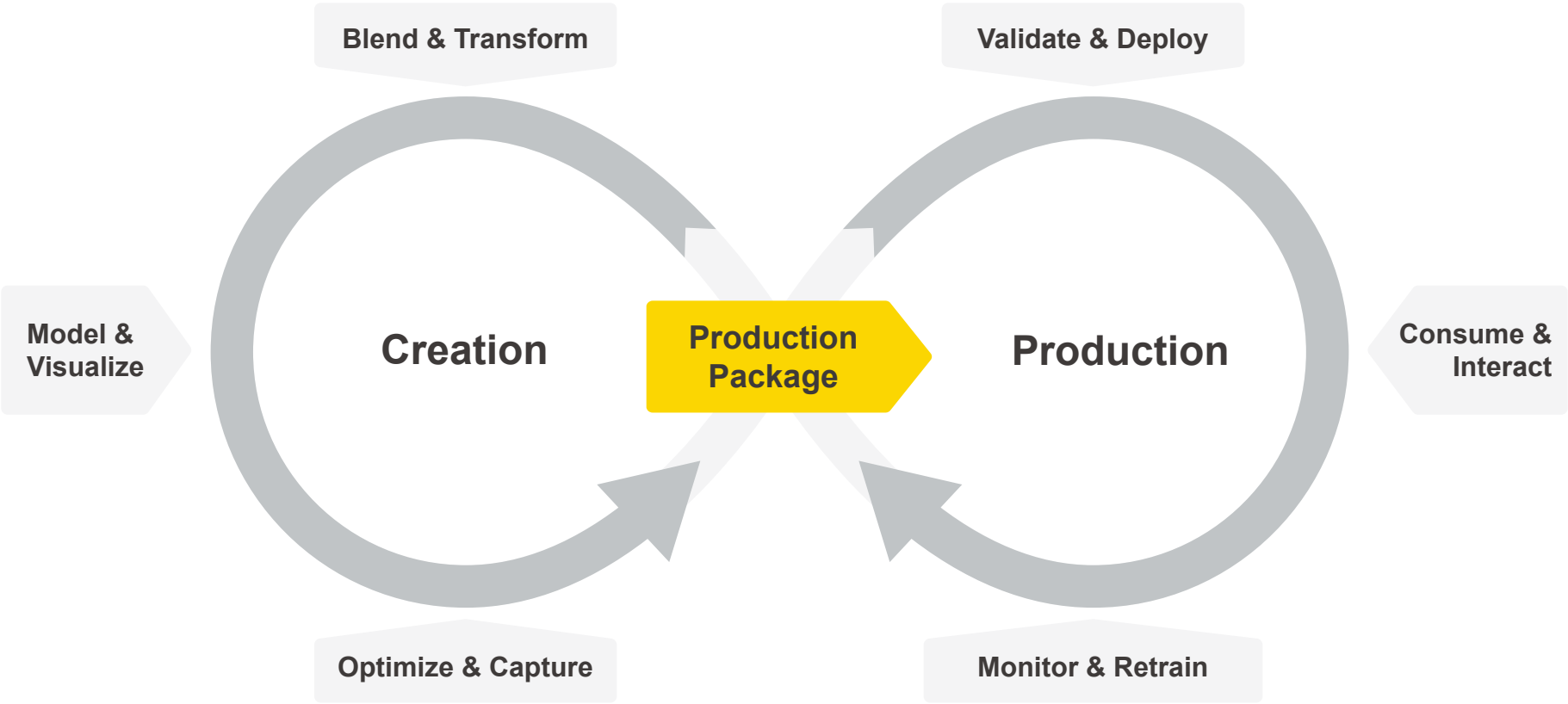
- Logging and Archiving Infrastructure
- Validation Workflow(s)
- Development, Validation, Production Envs (location, execution infrastructure…)

Example Validation Failures:

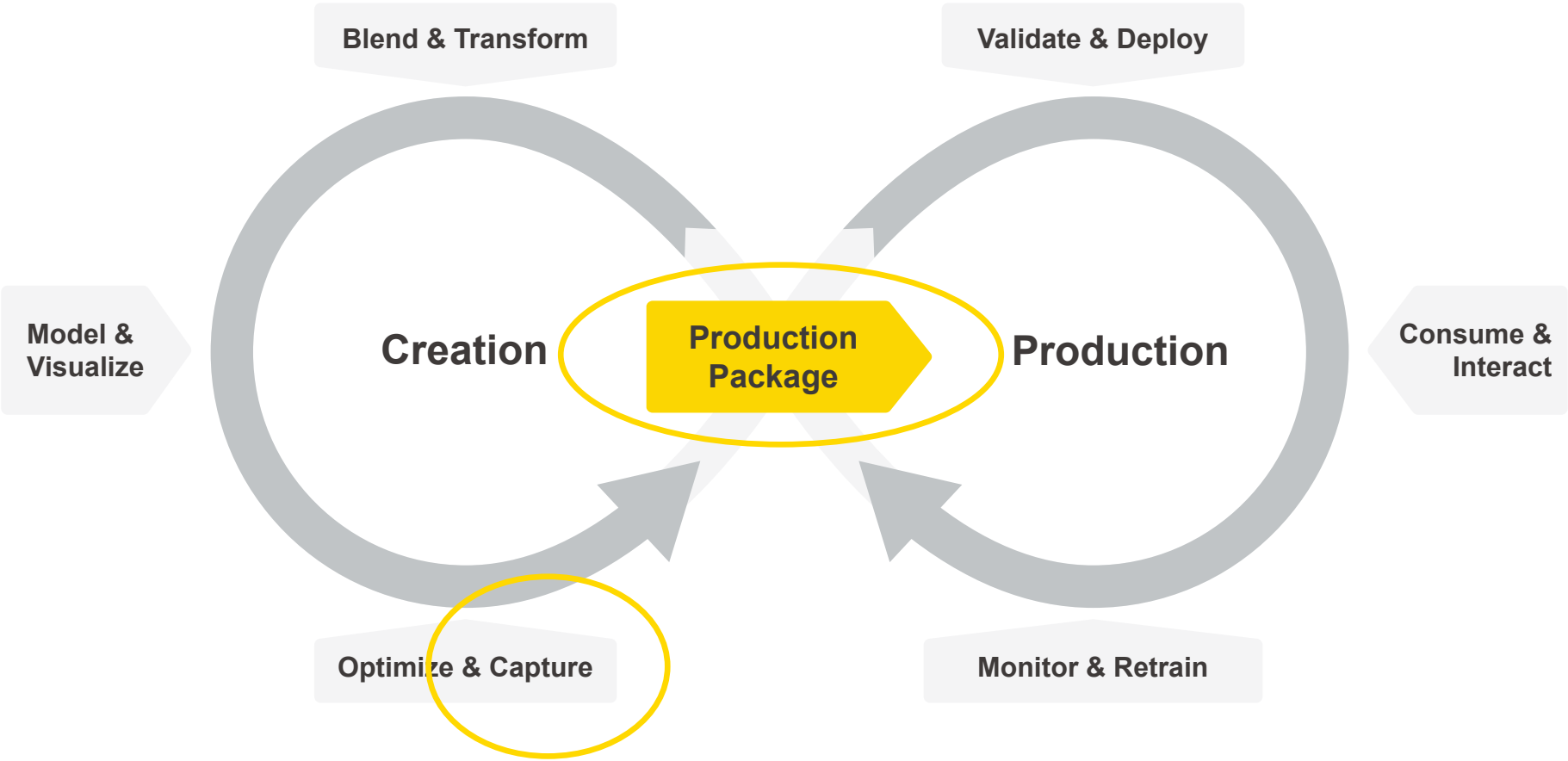
- Components out of Date
- Deprecated Nodes
- Unvalidated Python Code
- Not Authorized use of DB Nodes
- Workflow contains passwords
- GDPR standard components not used
- Non-verified Components
- Unknown data sources
- Call External Program Node used
- Illegal Write to External File System

Step 2:
Monitoring and Retraining

Data Science Deployment at Scale

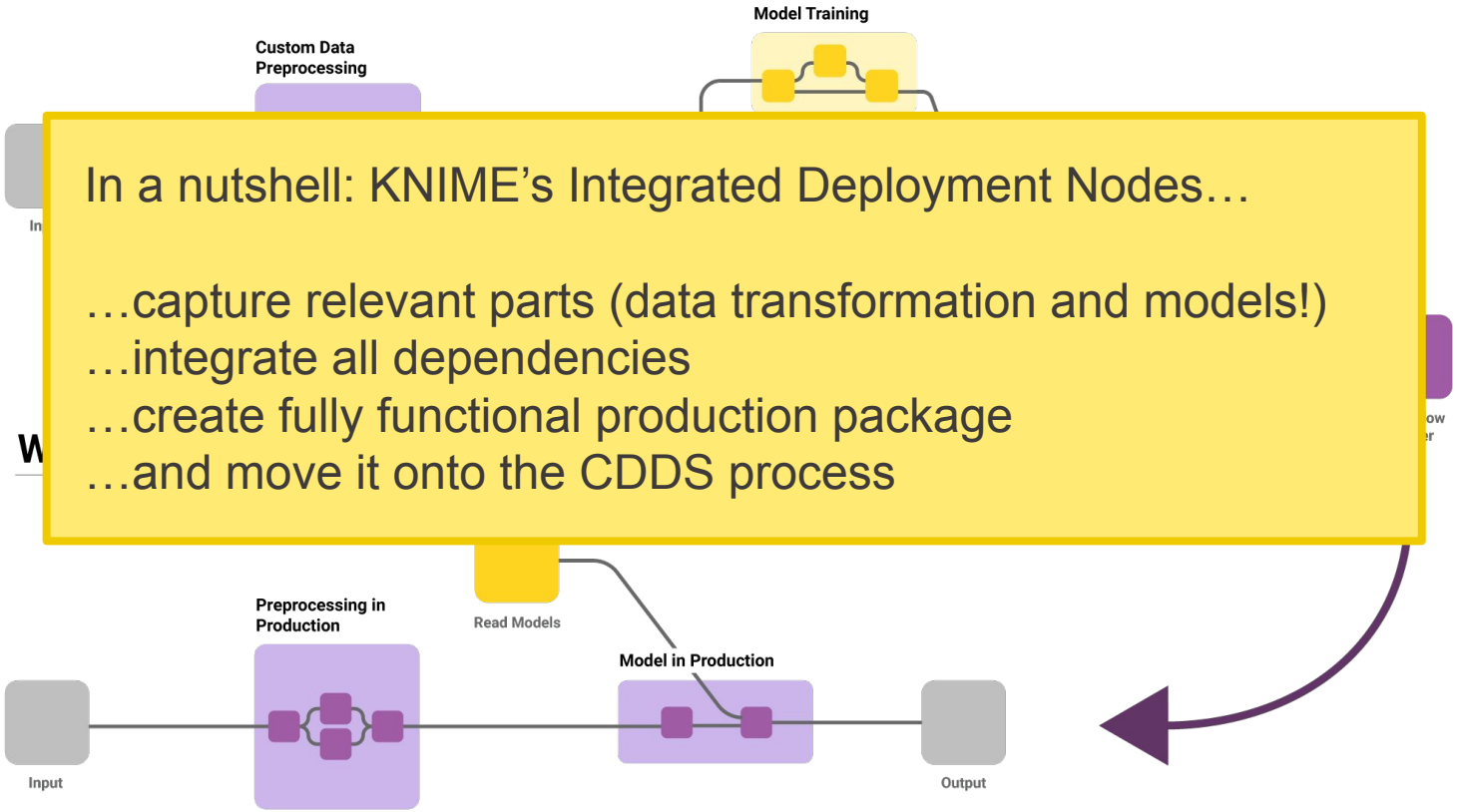


Data Science Deployment at Scale



Integrated Deployment (Continuous Integration “for free”)

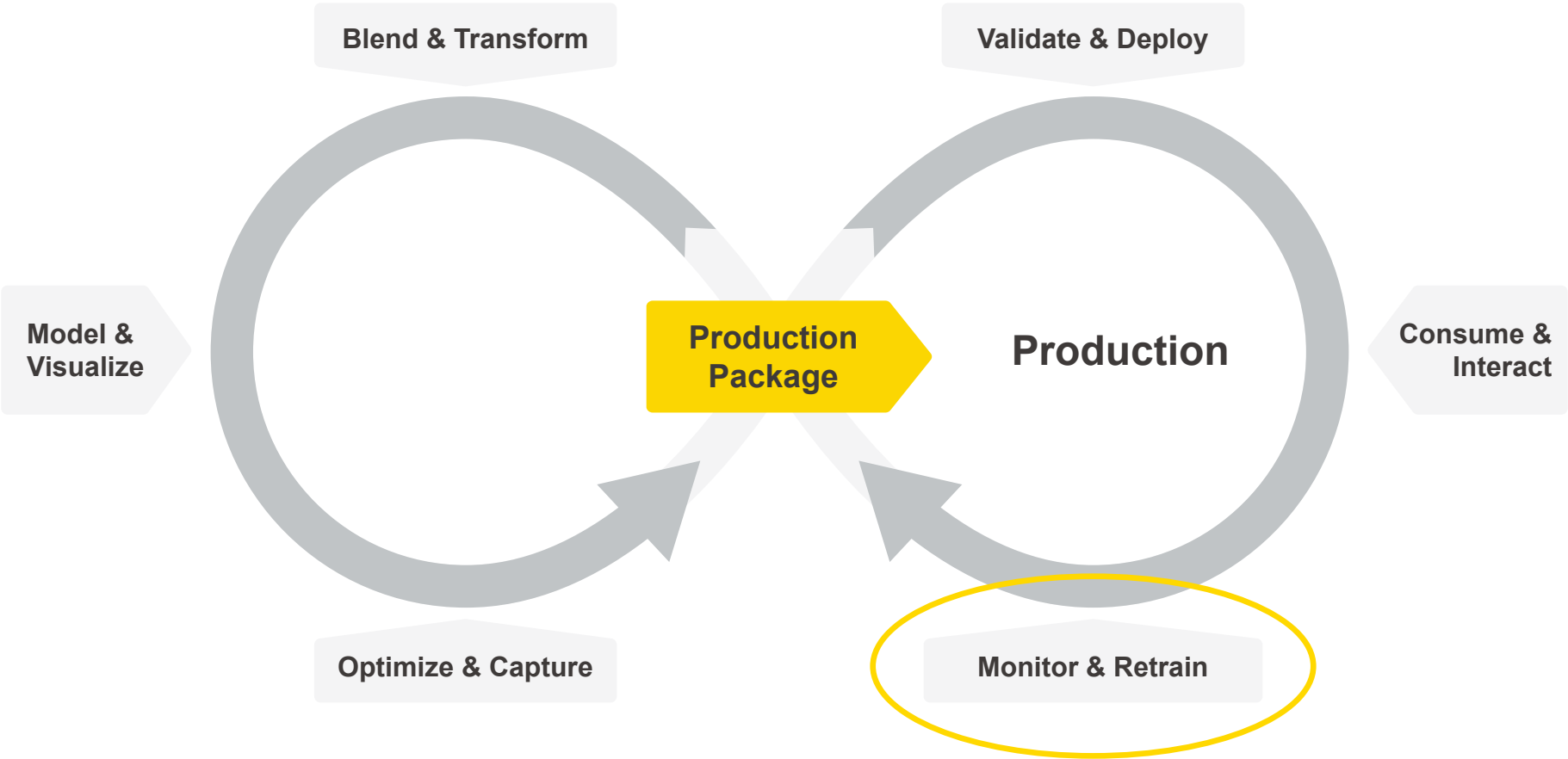
Creating Prediction Model



In a nutshell: KNIME’s Integrated Deployment Nodes…

…capture relevant parts (data transformation and models!)
…integrate all dependencies
…create fully functional production package
…and move it onto the CDDS process

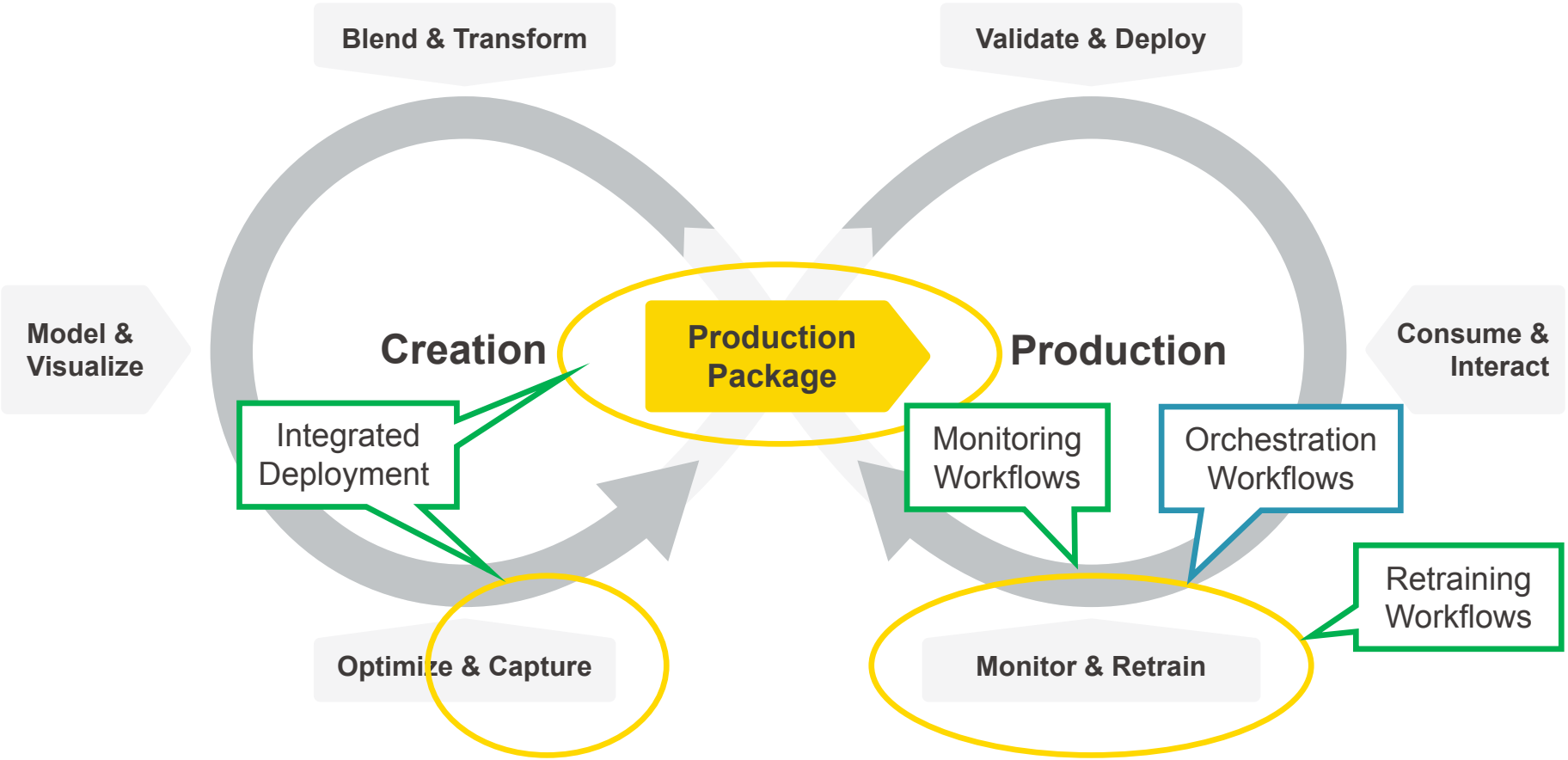
Data Science Deployment at Scale



The CDDS Extension for KNIME Business Hub

Monitoring and Retraining of Data Science in Production

Data Science Deployment at Scale



Continuous Deployment for Data Science

Still: Absolutely usable out of the box!

…but again: Completely Customizable

- Logging and Archiving Infrastructure
- Validation Workflow(s)
- Development, Validation, Production Environments (location, execution infrastructure…)
- Continuous Integration
- Monitoring Workflow(s)
- Retraining Workflow(s)
- Orchestration Workflow(s)

typically provided by the Data Science Team!

Summary

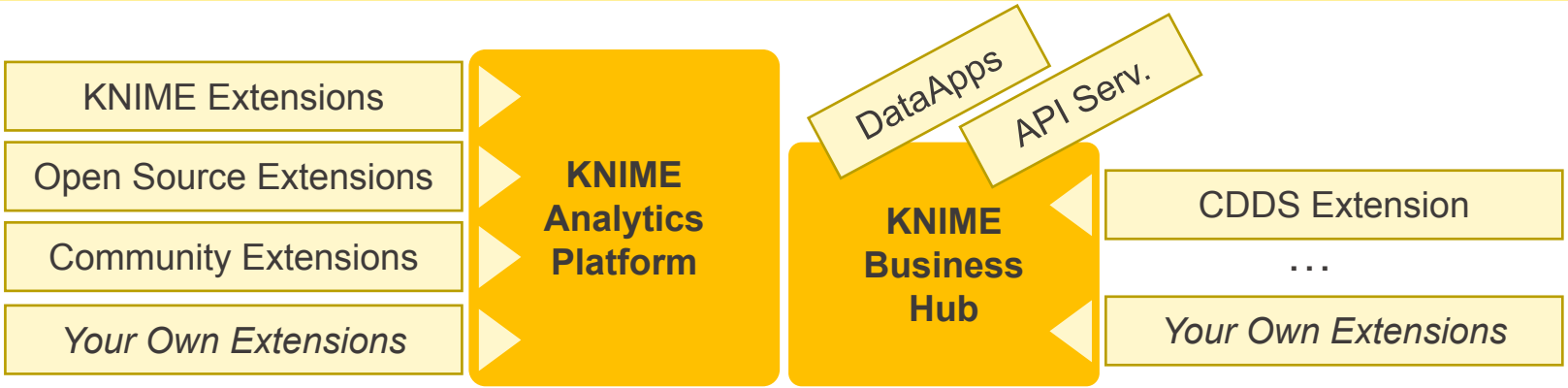
CDDS: Continuous Deployment for Data Science

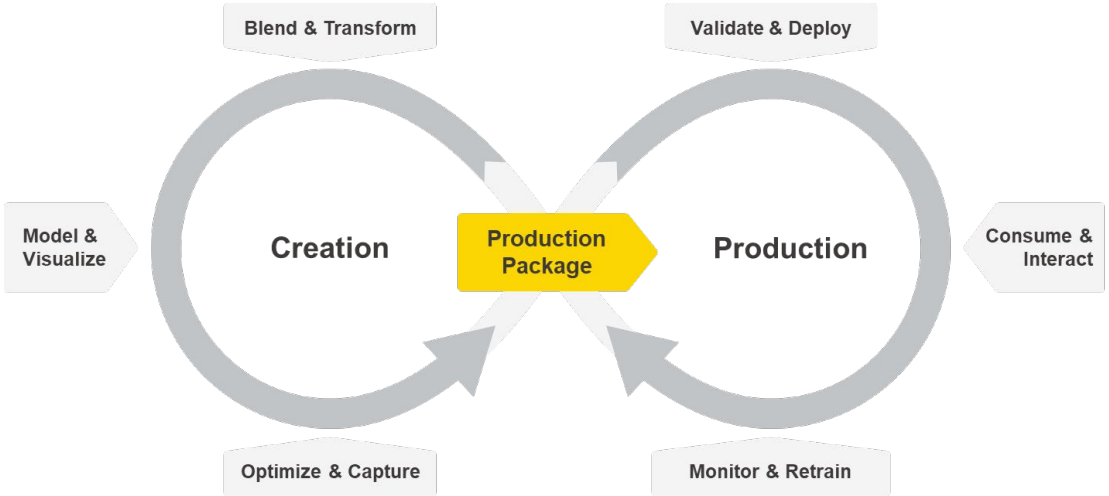
Complete & Customizable Process as an Extension for KNIME Business Hub

With complete choice of implementation:

- Environments (dev, test, validation, prod, …)
- Execution (local, cloud, hybrid, …)
- Storage / Archive (DB, Git, Cloud Storage, …)

Flexible Data Science





Thank You!