

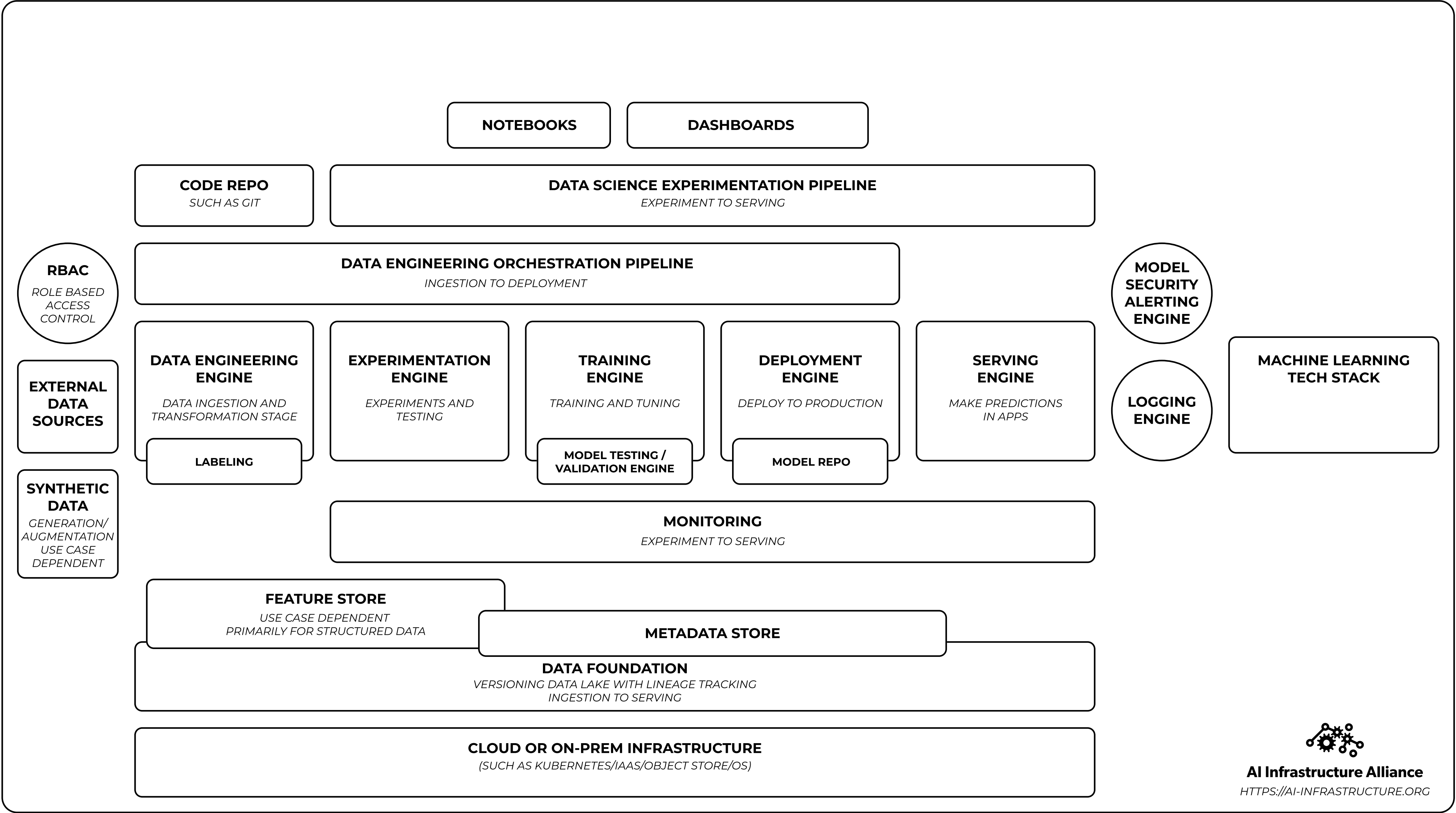


Enterprise MLOps 101

Michael Balint and William Benton



“MLOps is confusing”



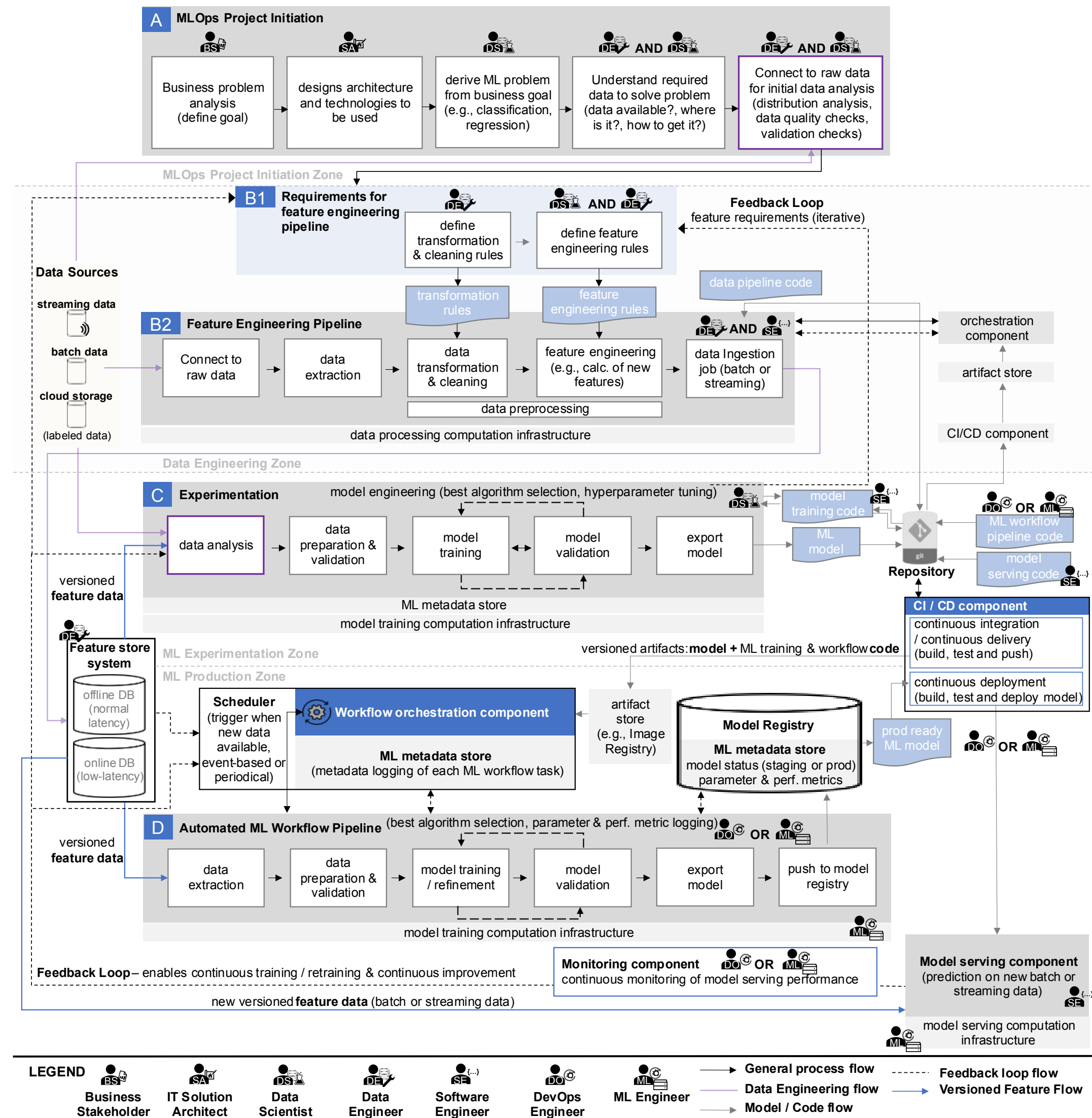
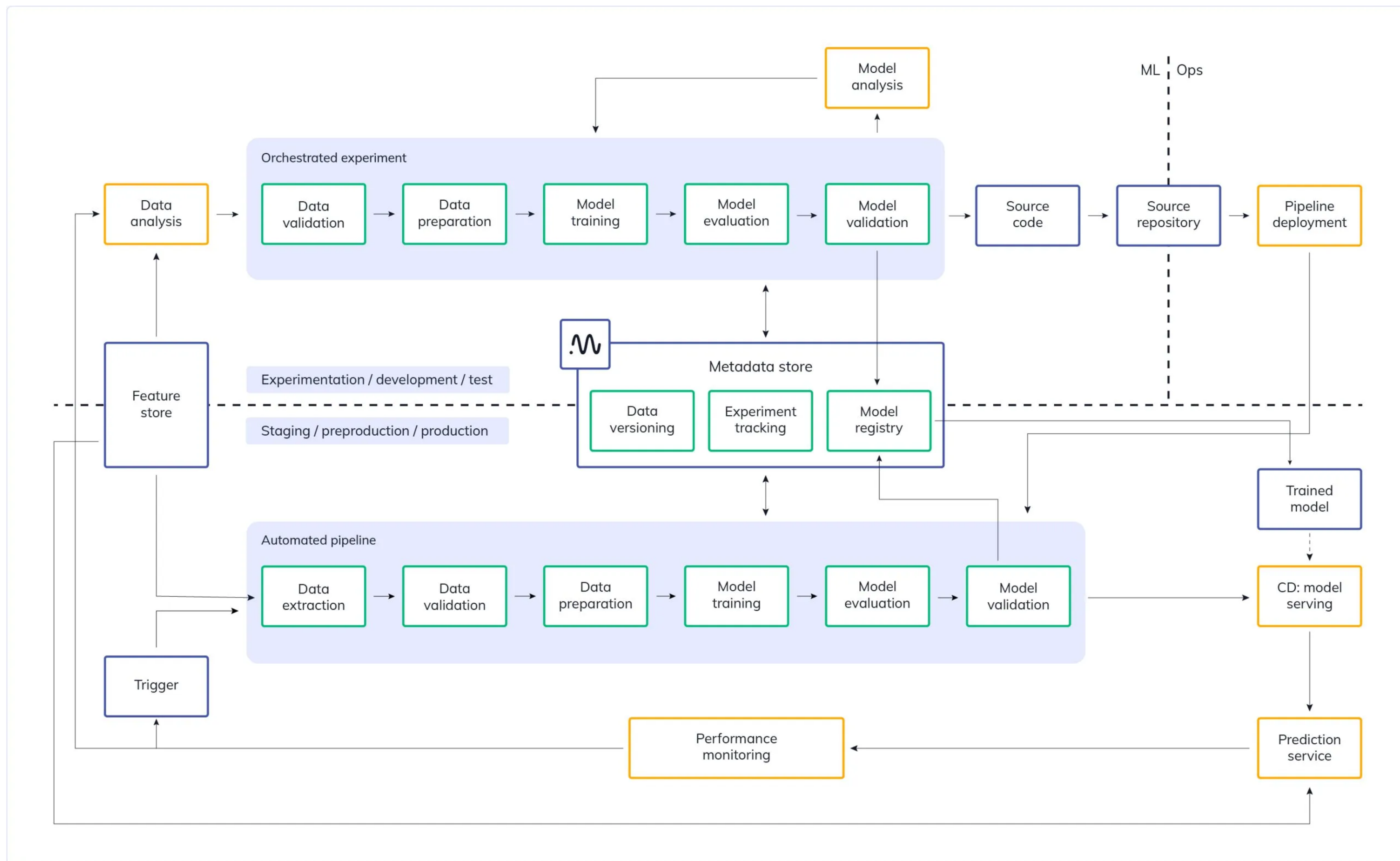



Figure 4. End-to-end MLOps architecture and workflow with functional components and roles

(source: <https://arxiv.org/abs/2205.02302>)

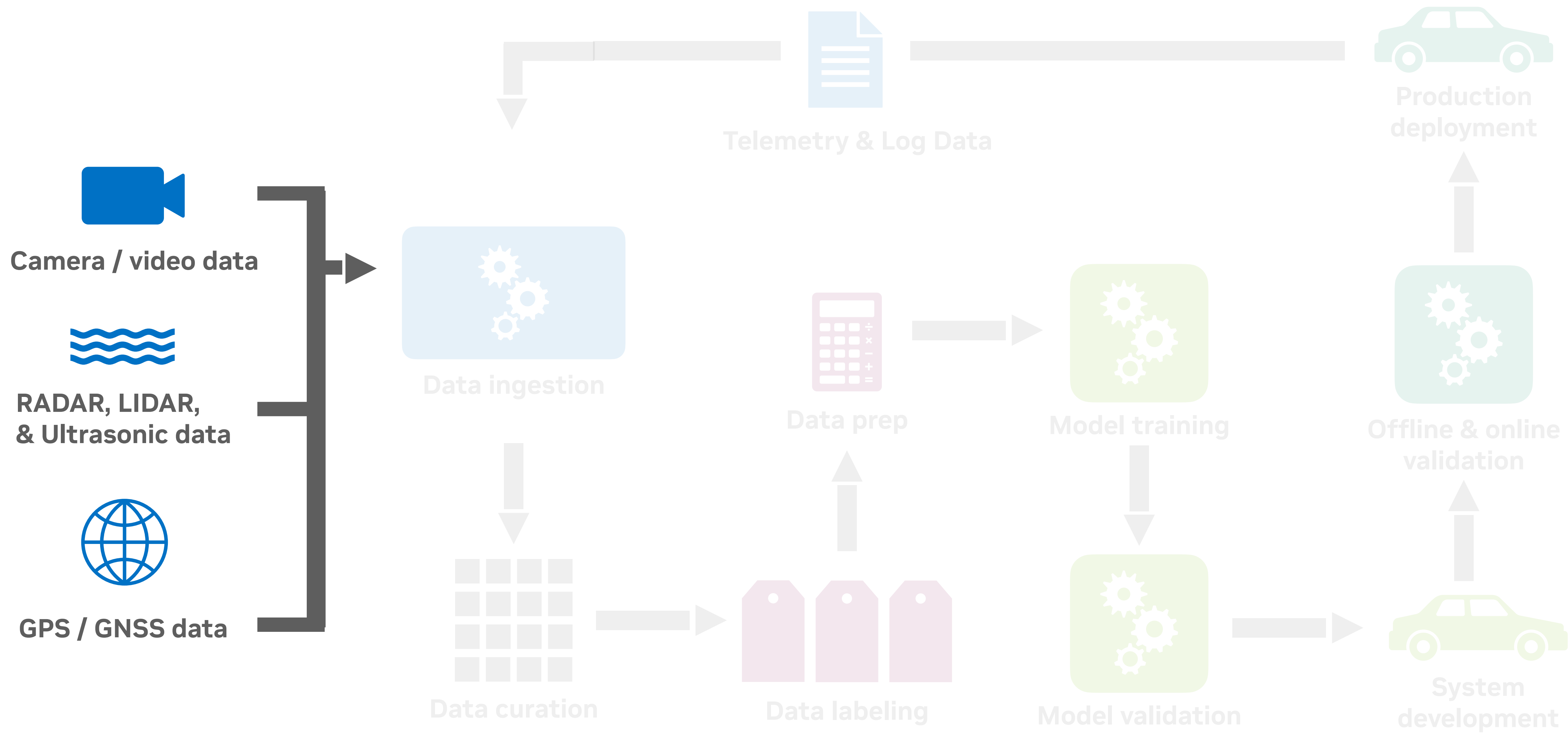


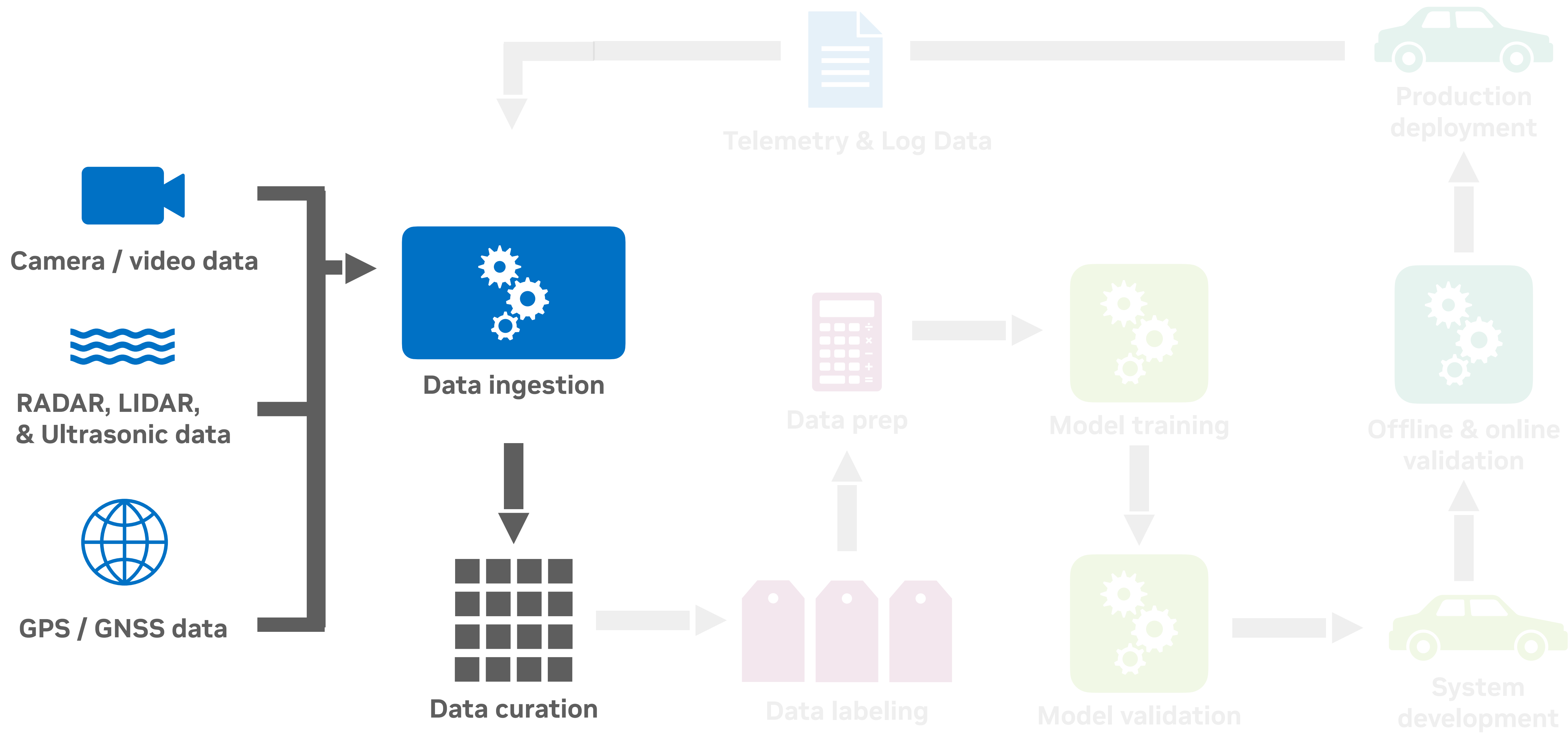


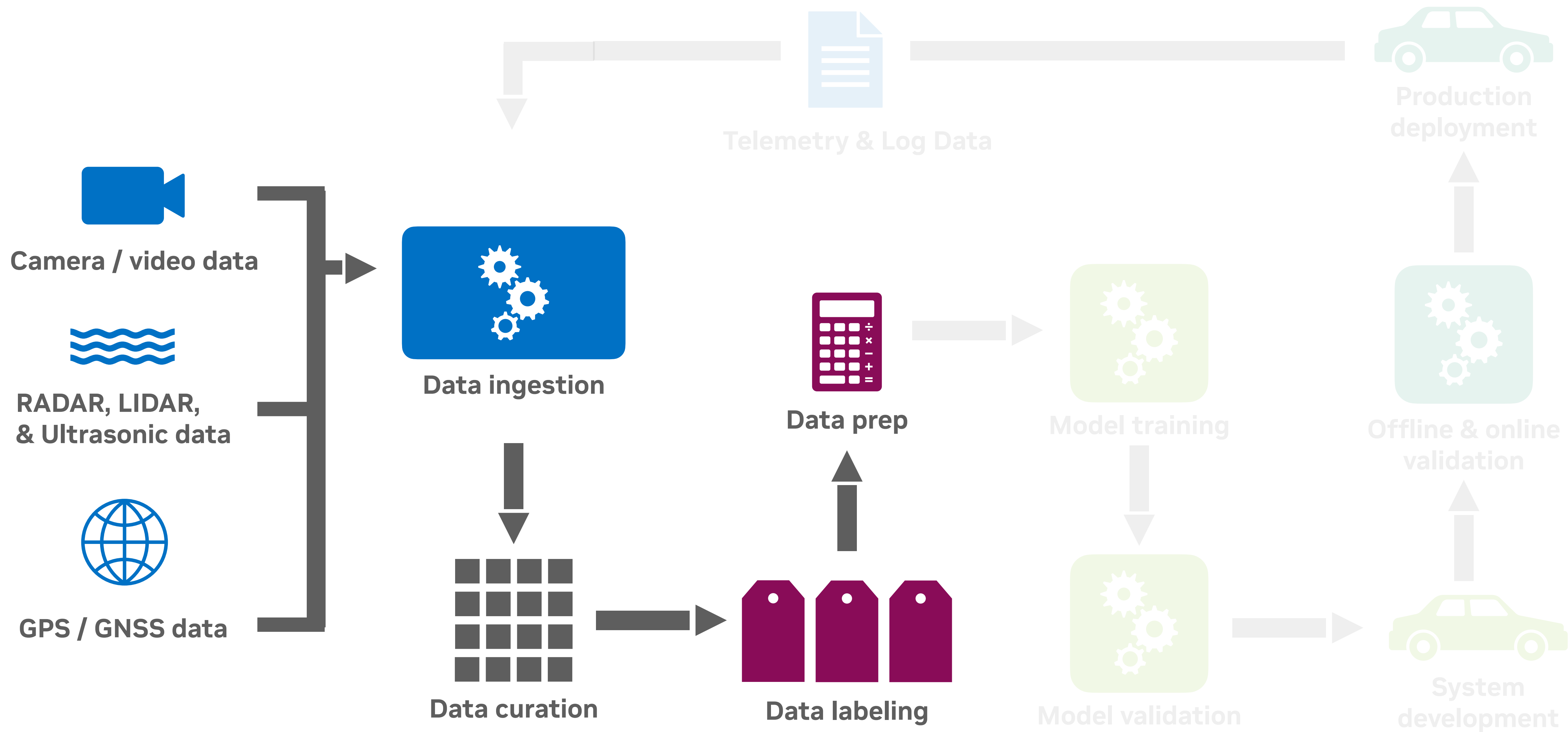
Example ML Systems and Lifecycles

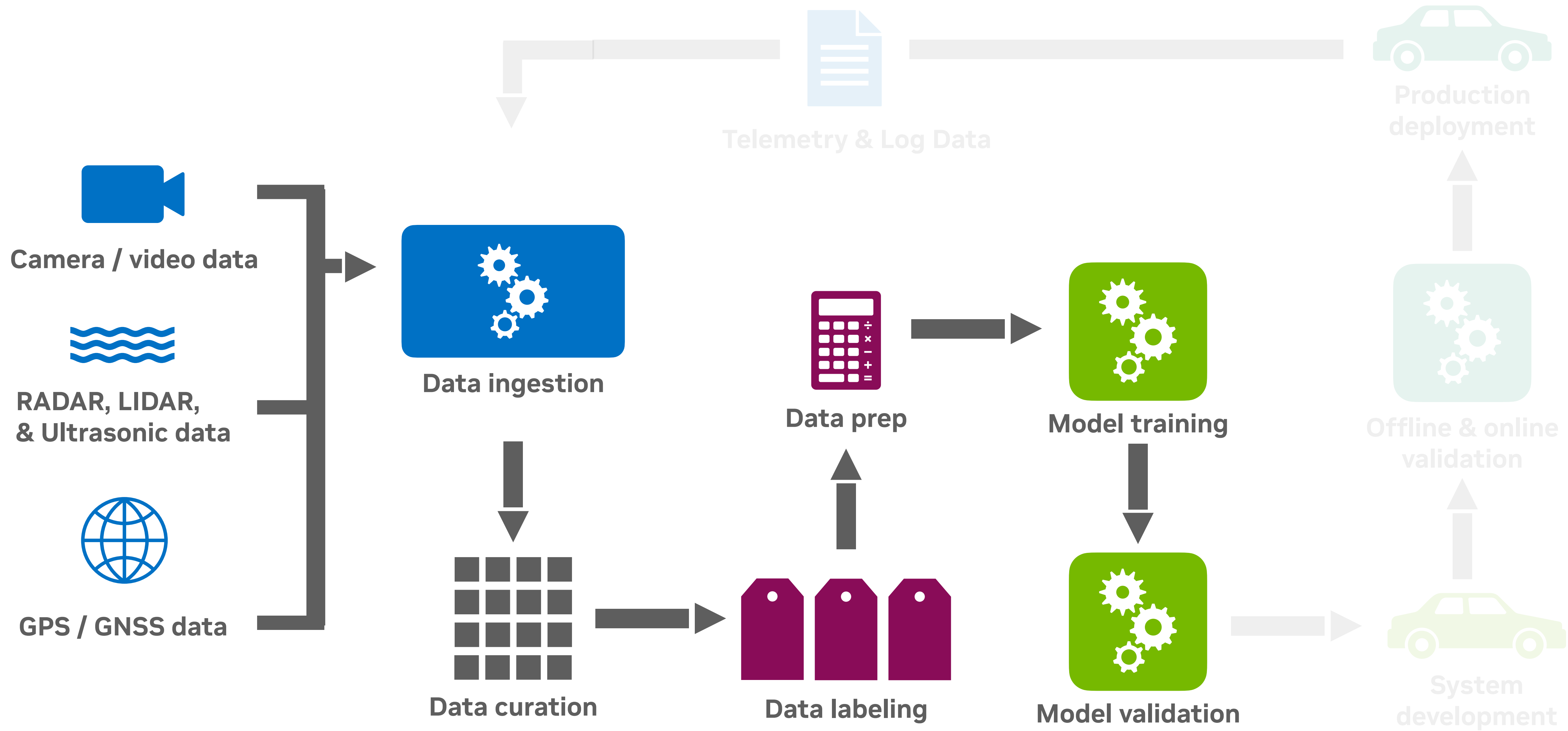


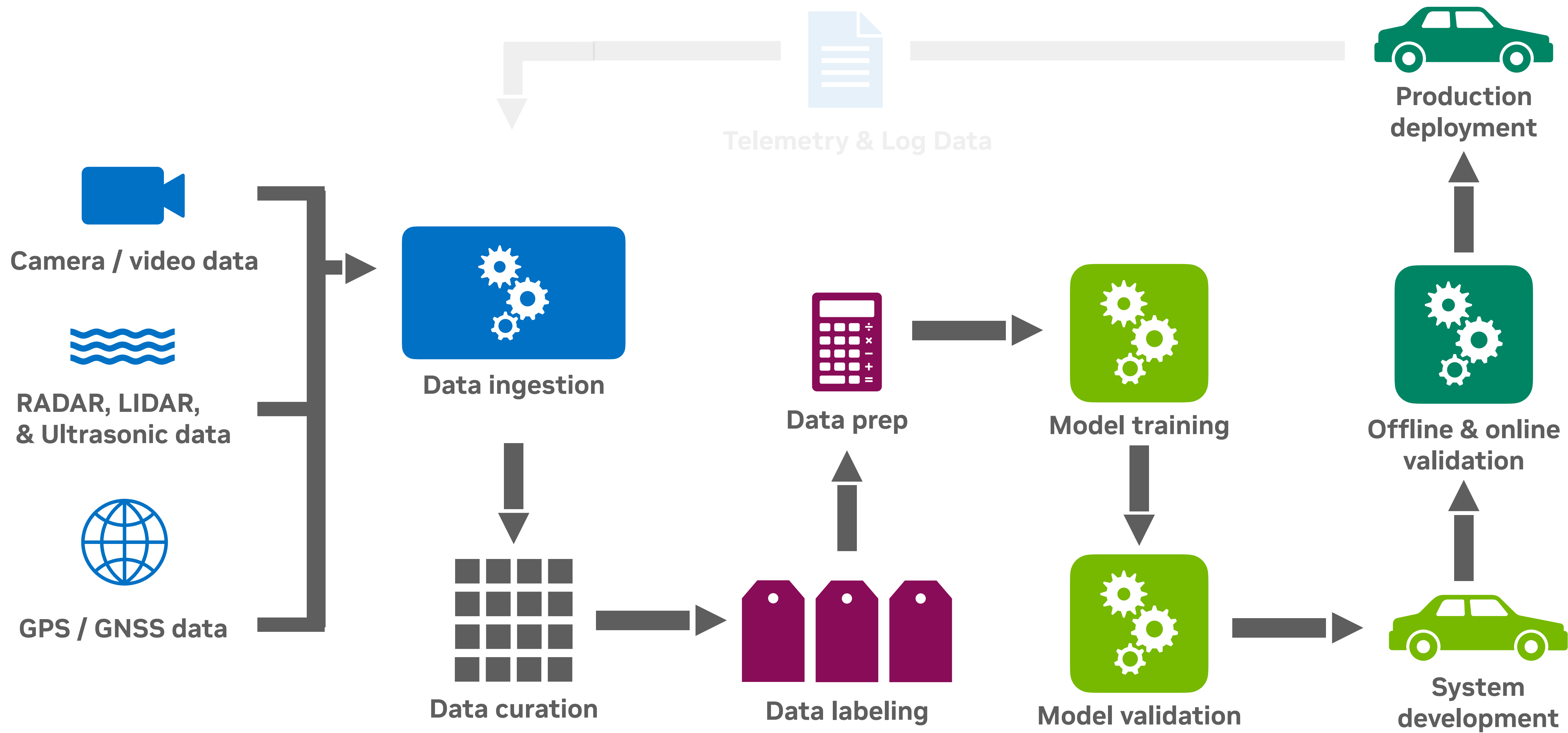
Case Study: Autonomous Vehicles

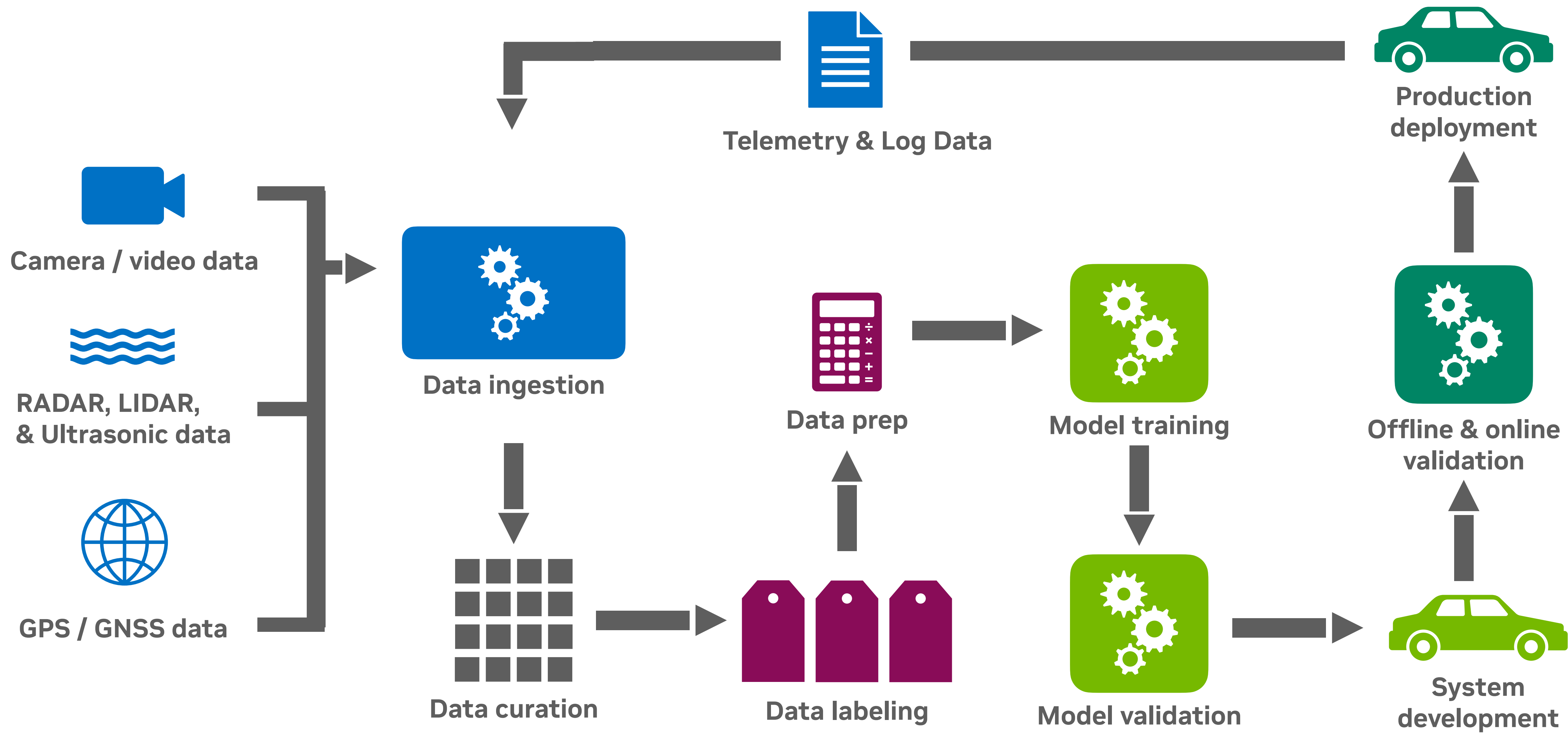






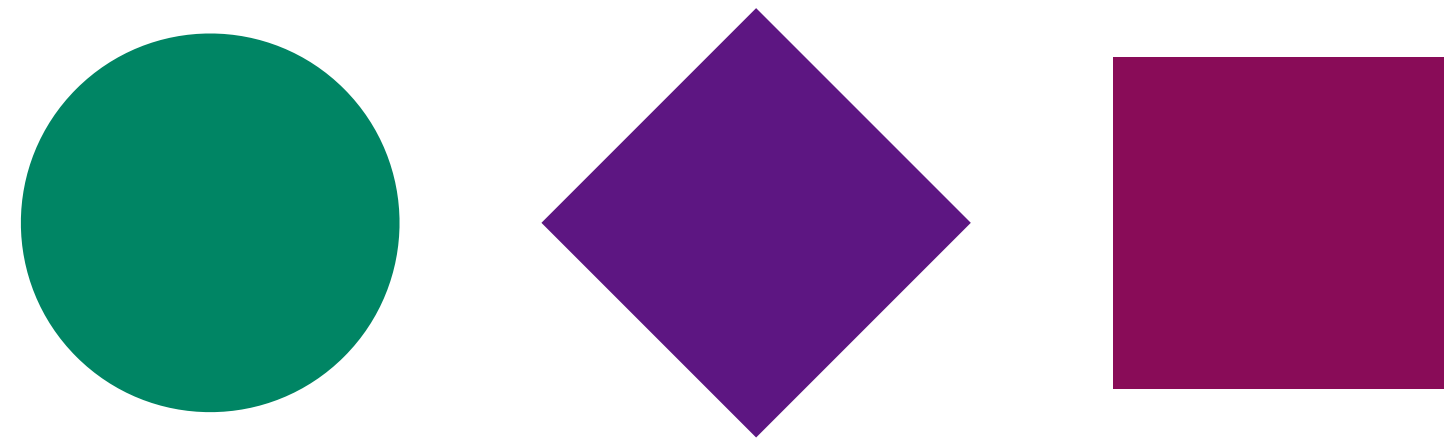


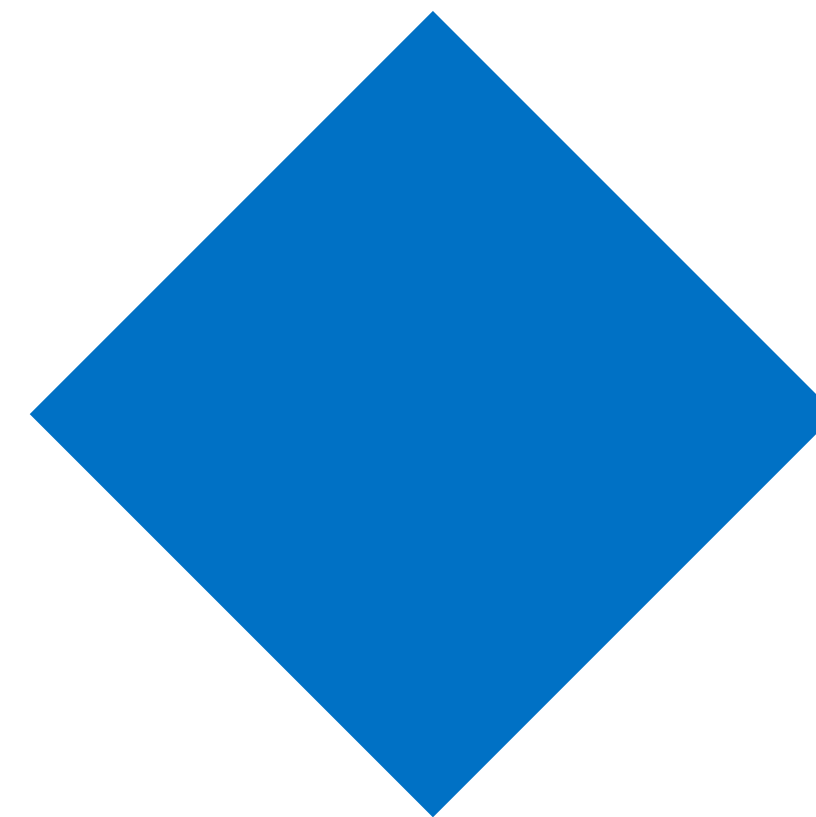
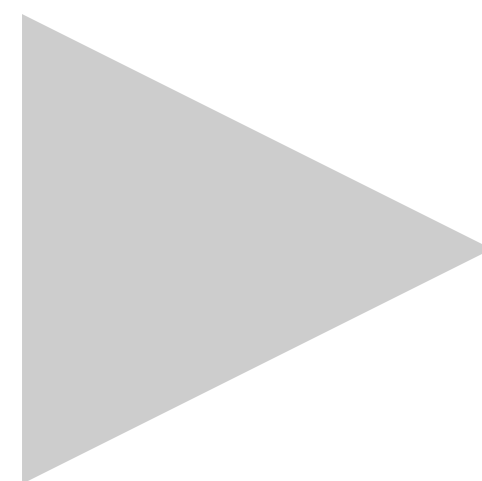
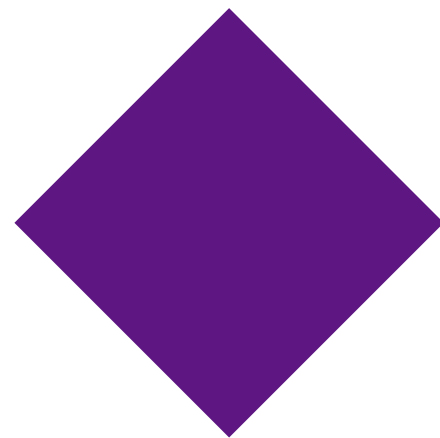
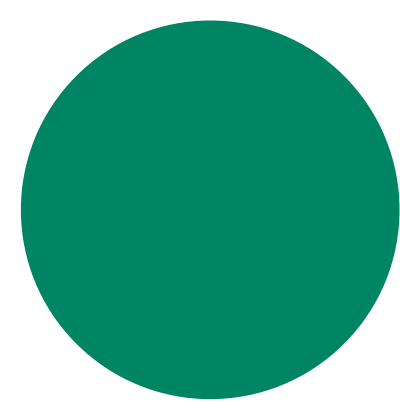






Case Study: Retail Recommendations

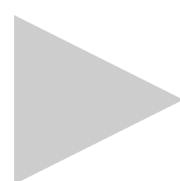




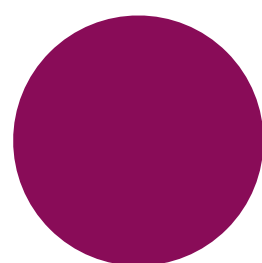
1



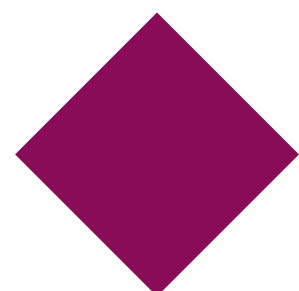
1



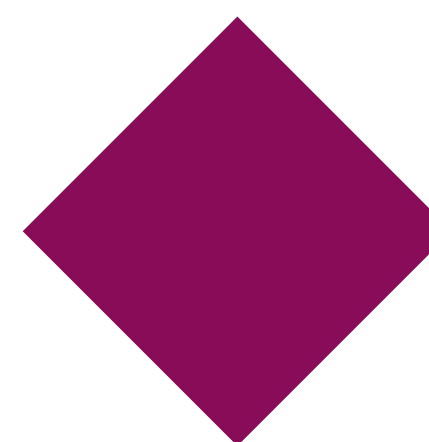
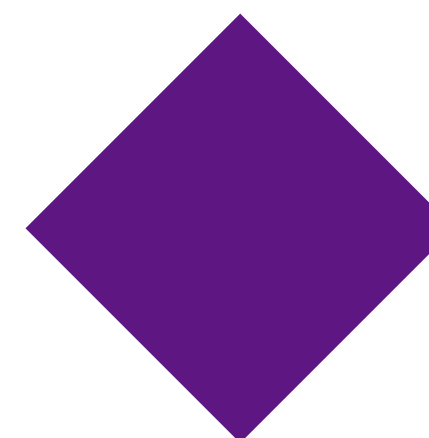
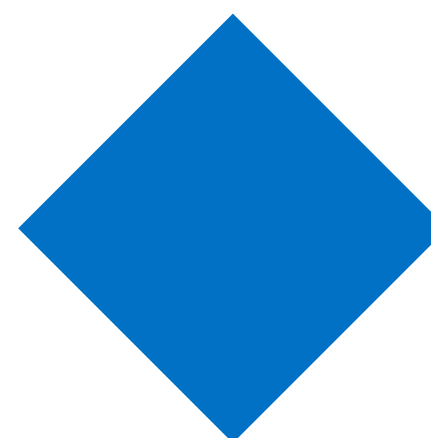
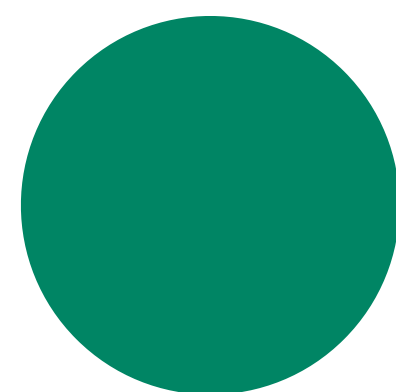
2



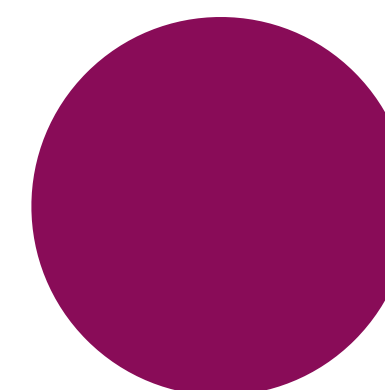
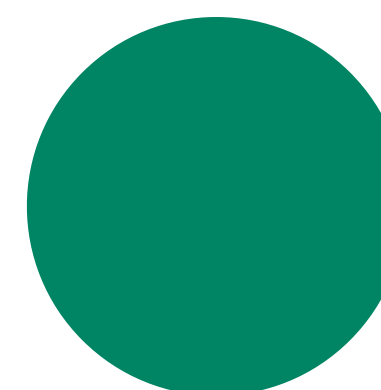
1



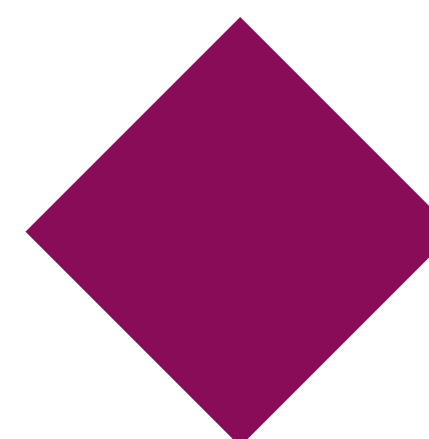
1

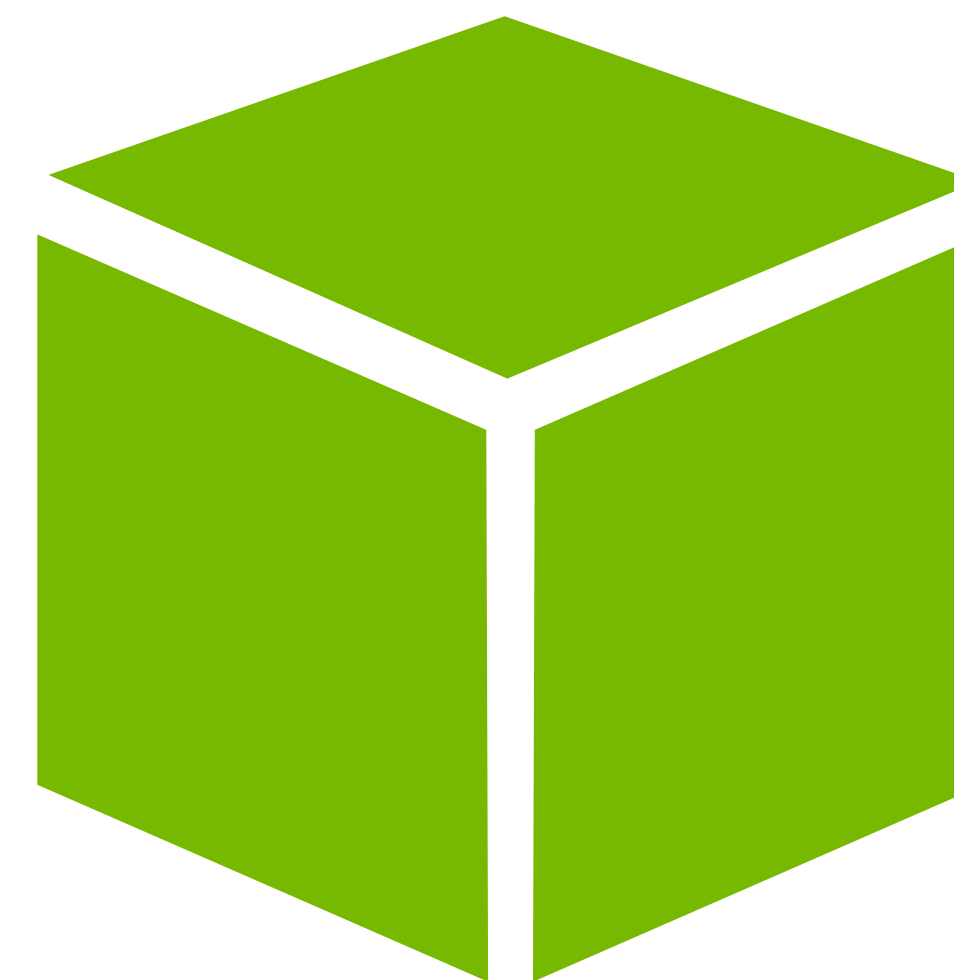
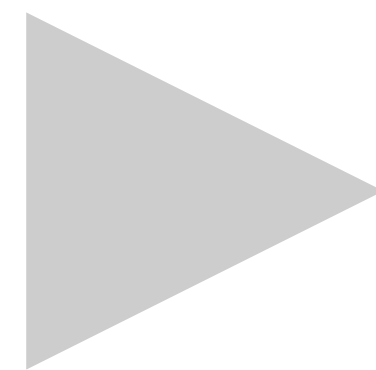
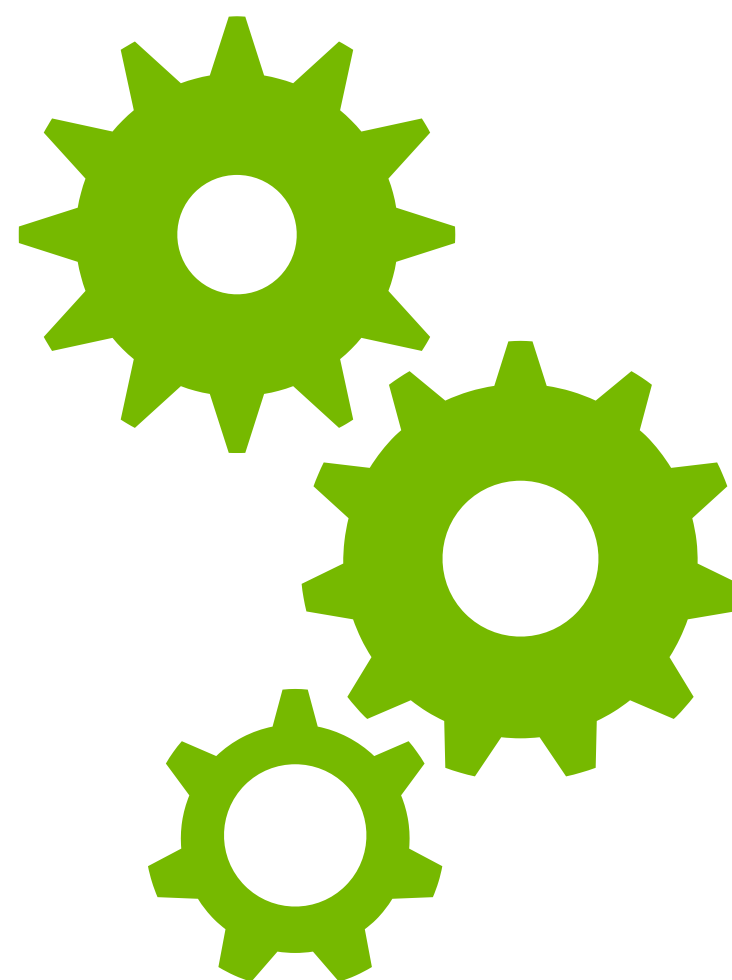
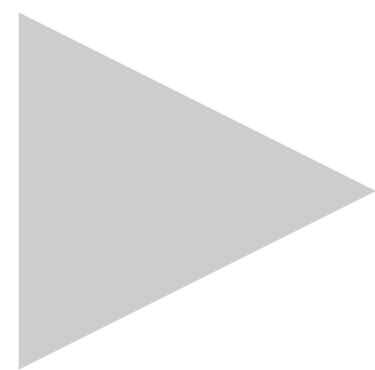
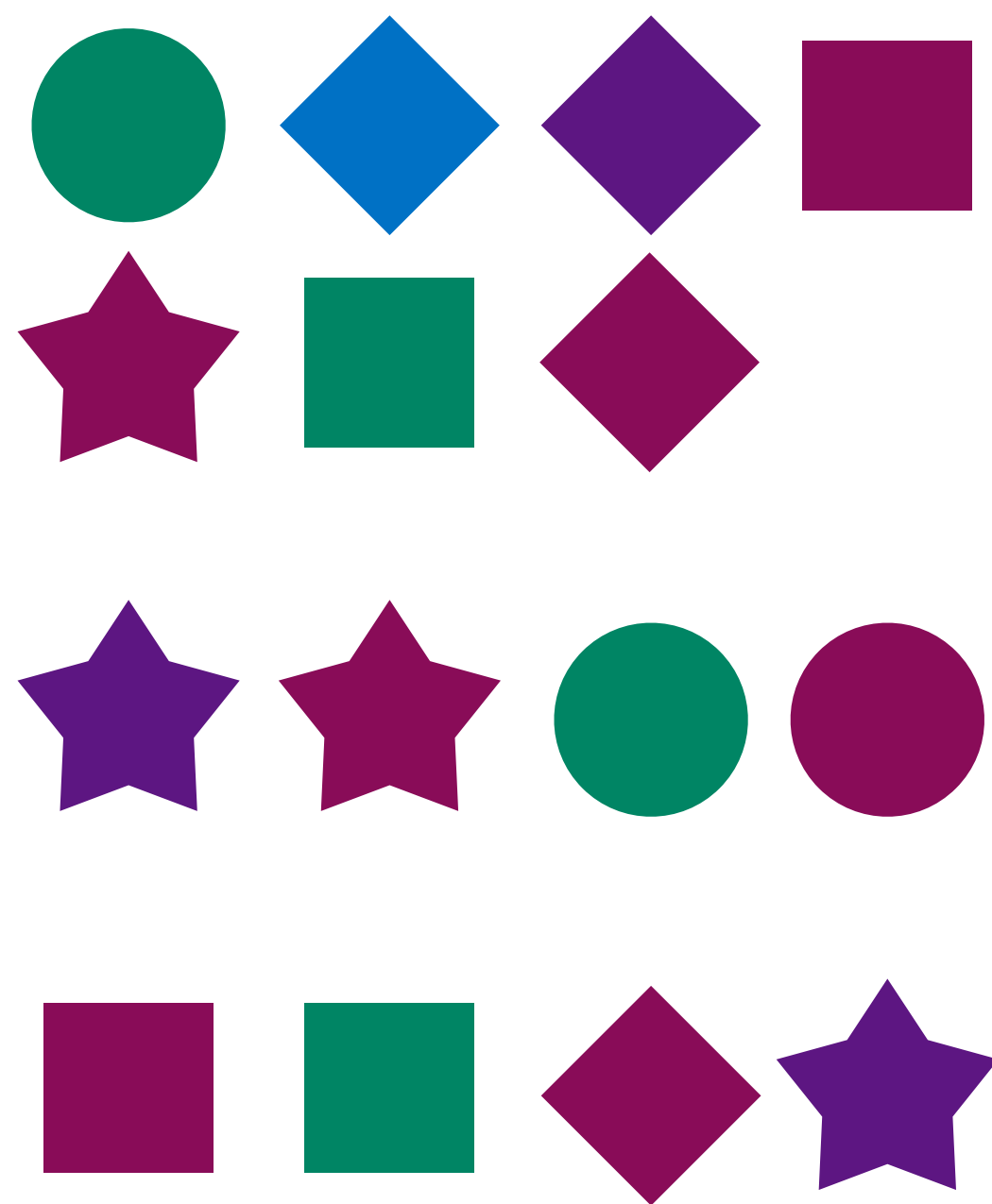


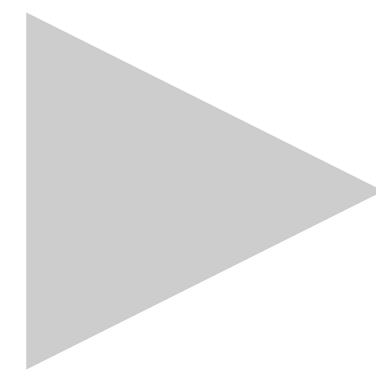
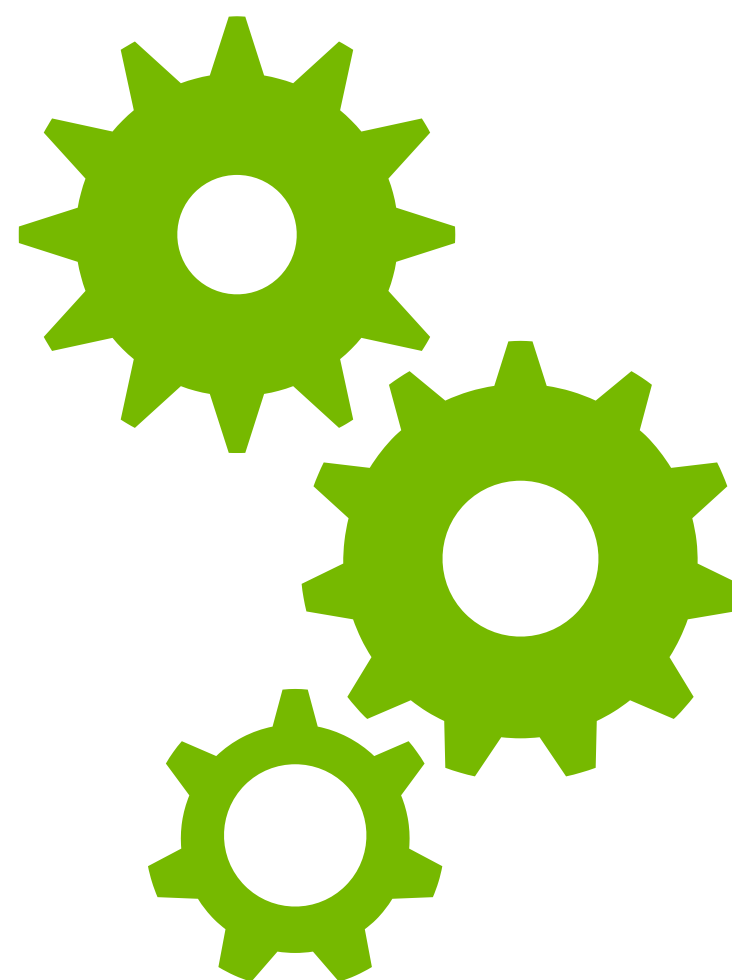
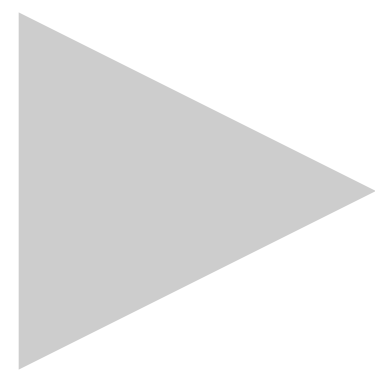
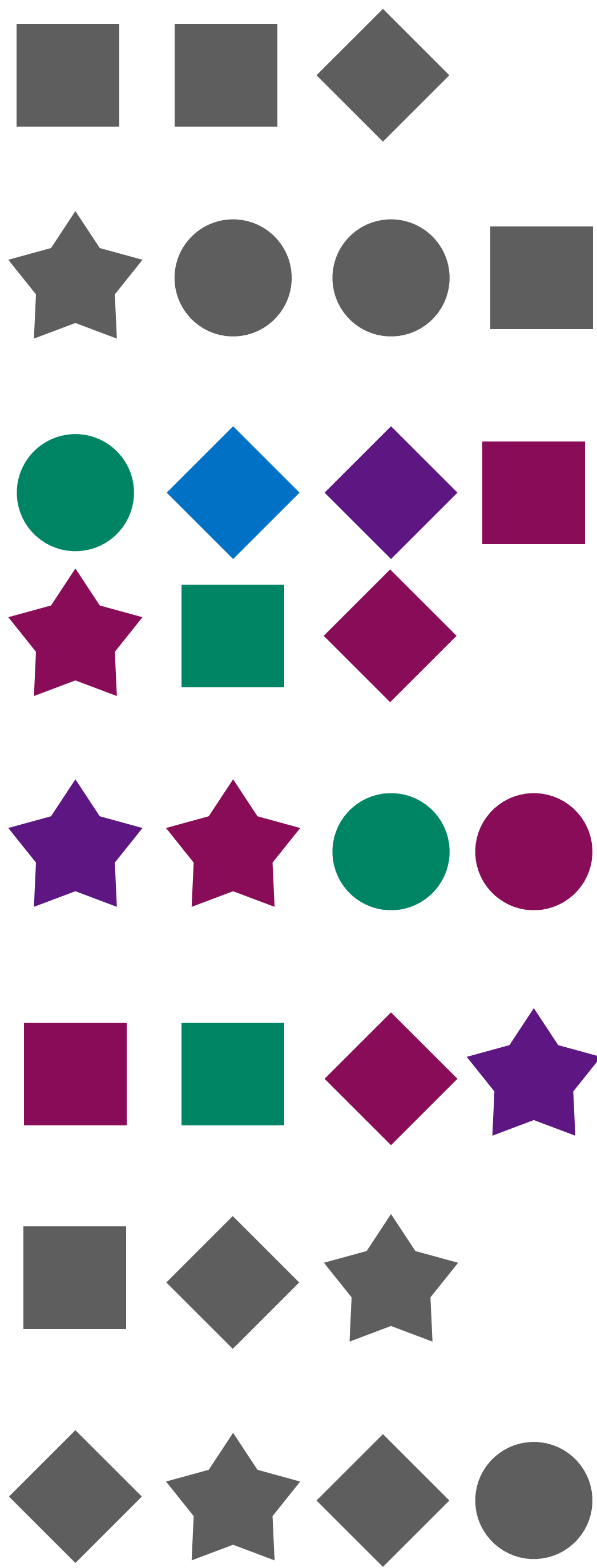
2

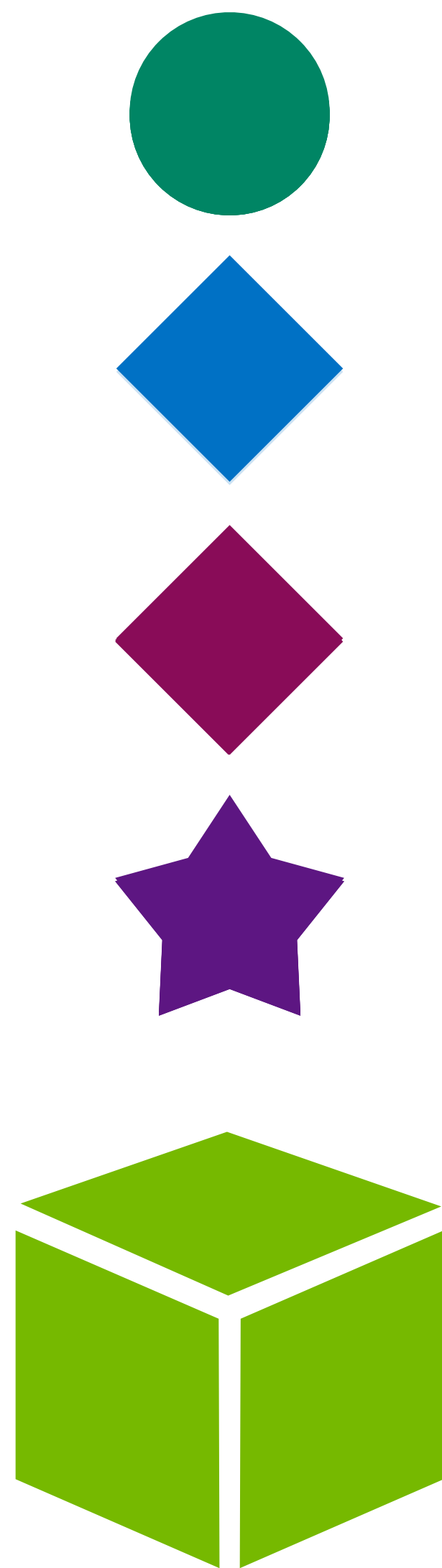


3

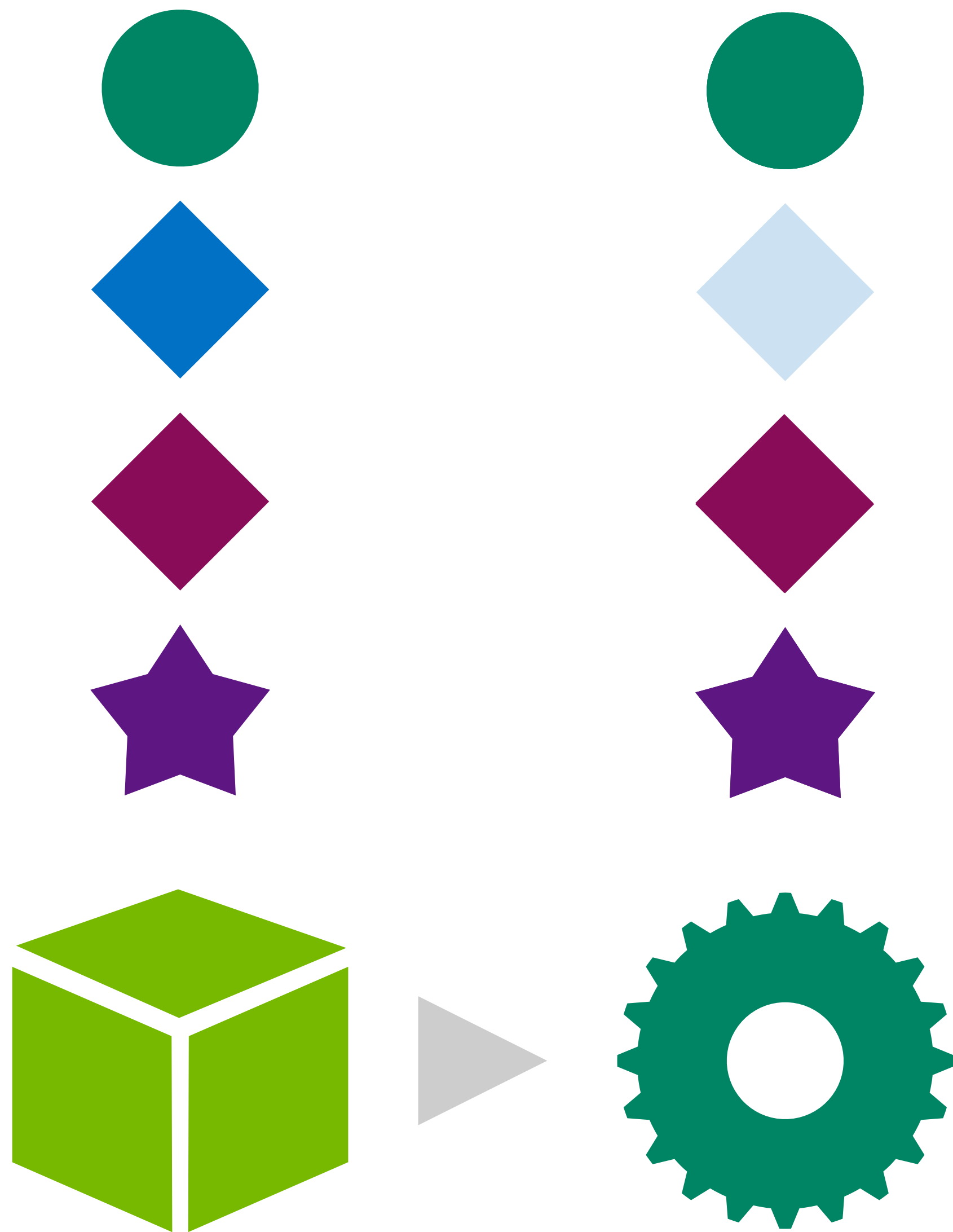






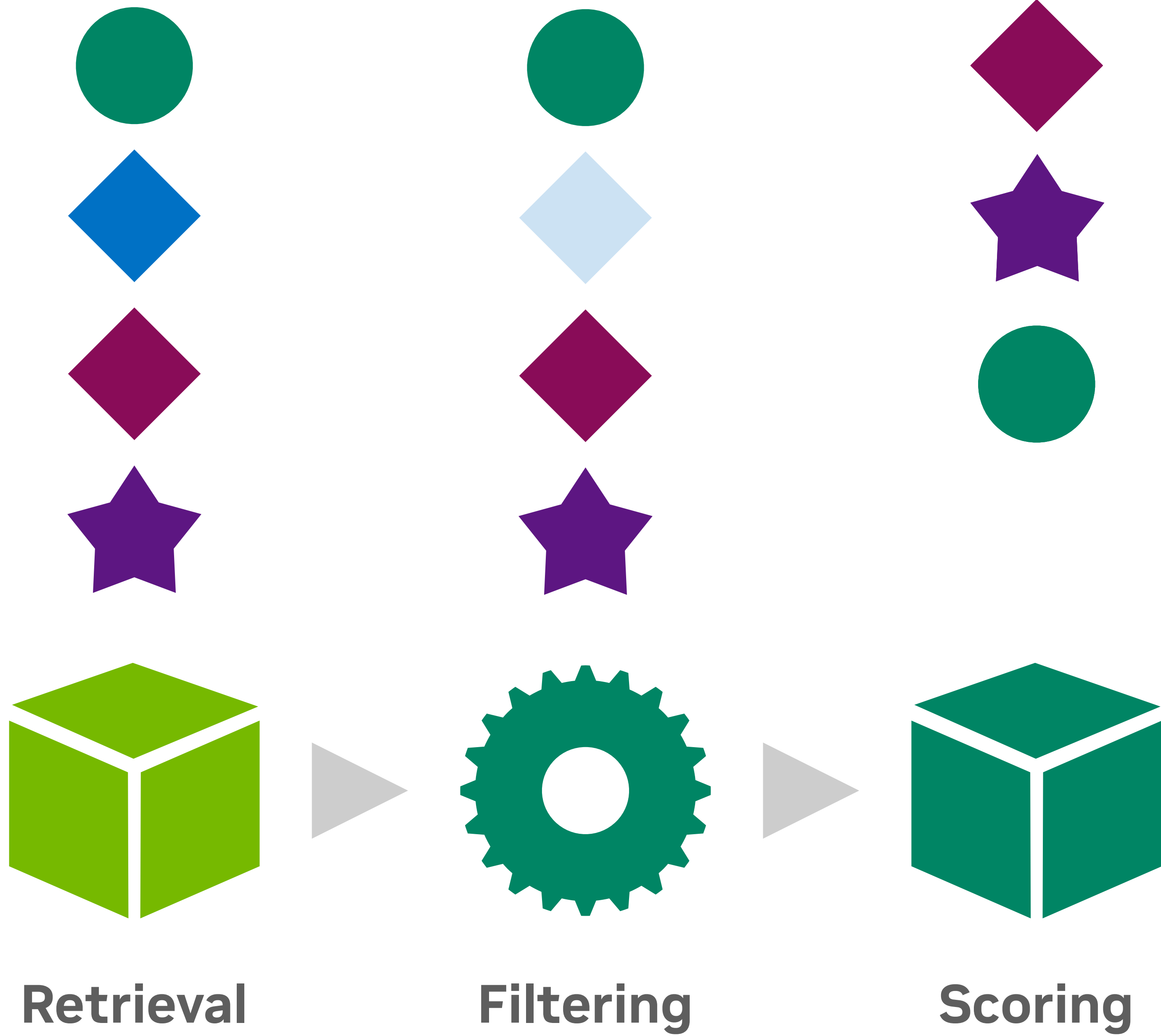


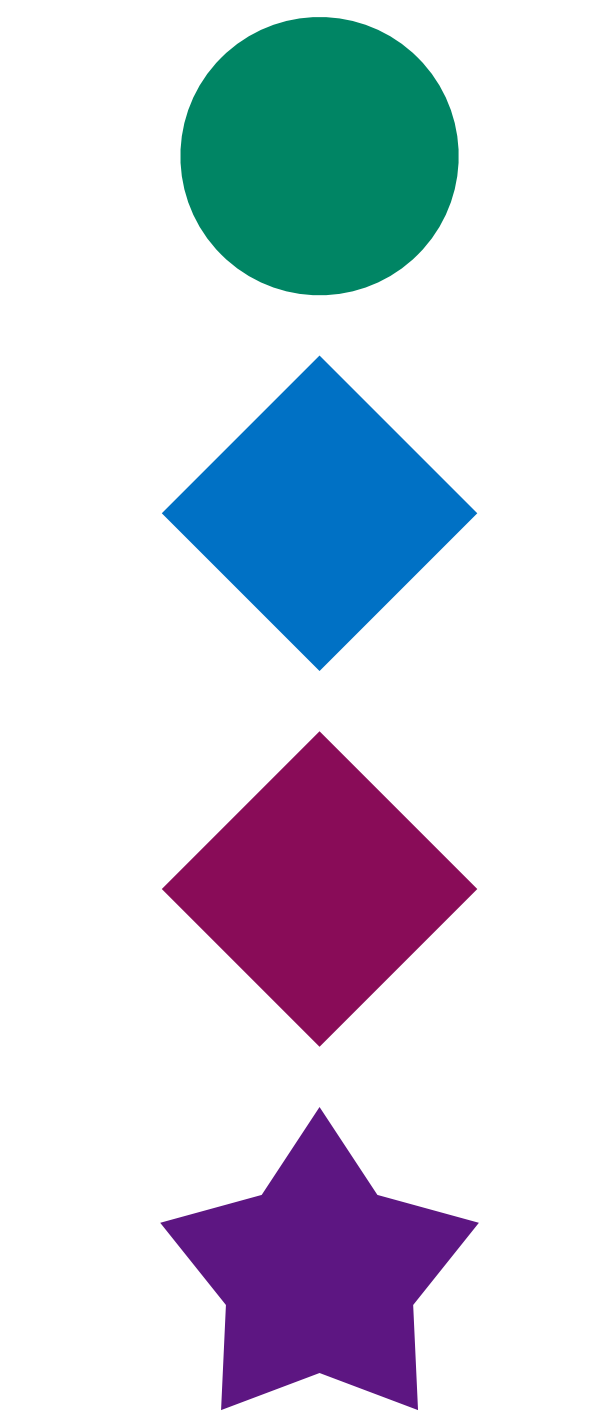
Retrieval



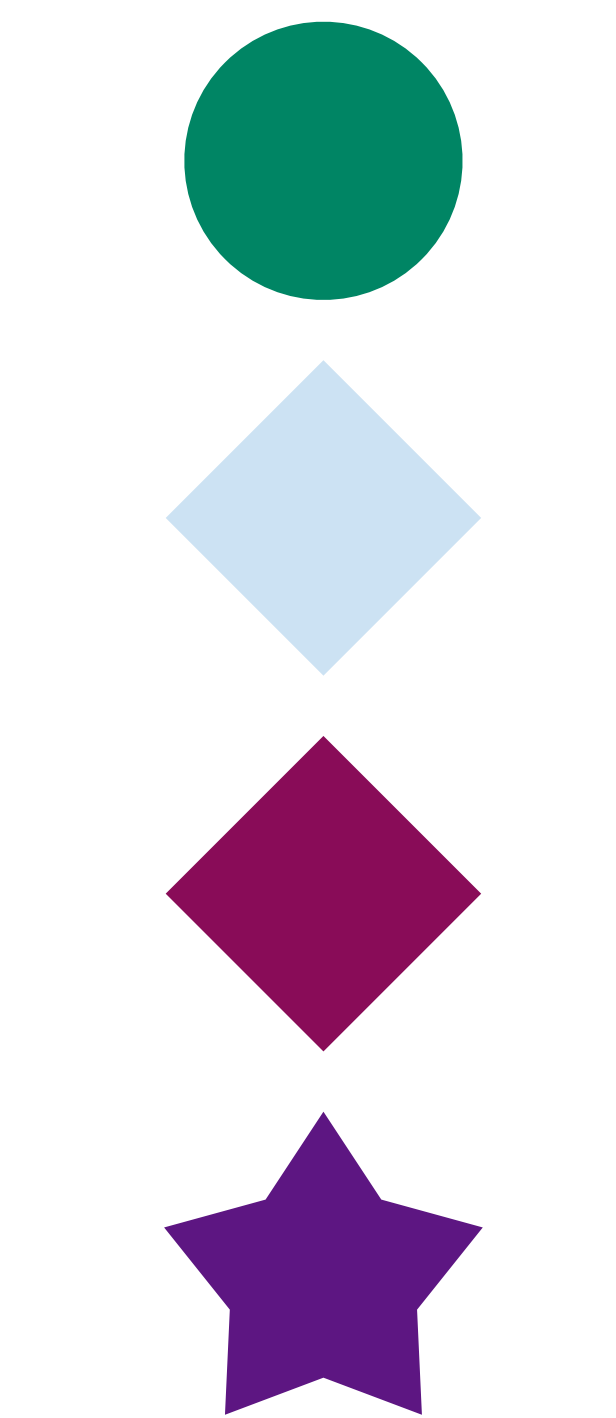
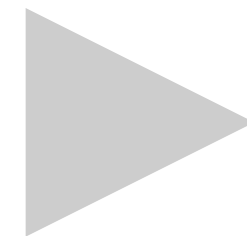
Retrieval

Filtering

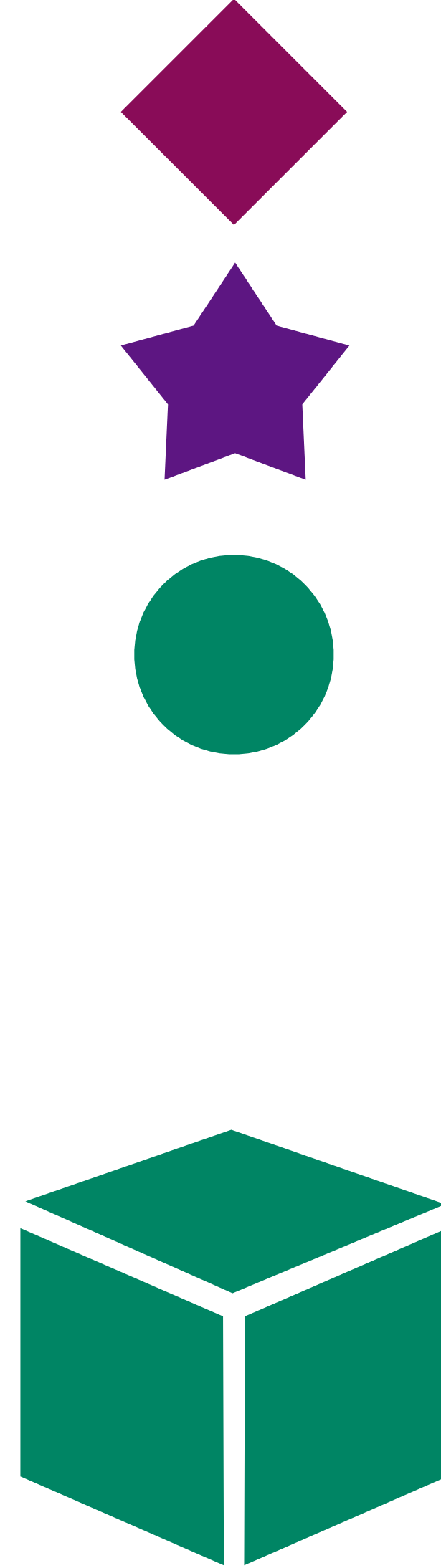
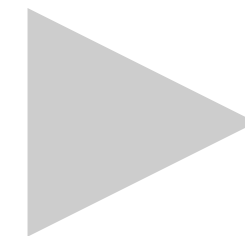




Retrieval



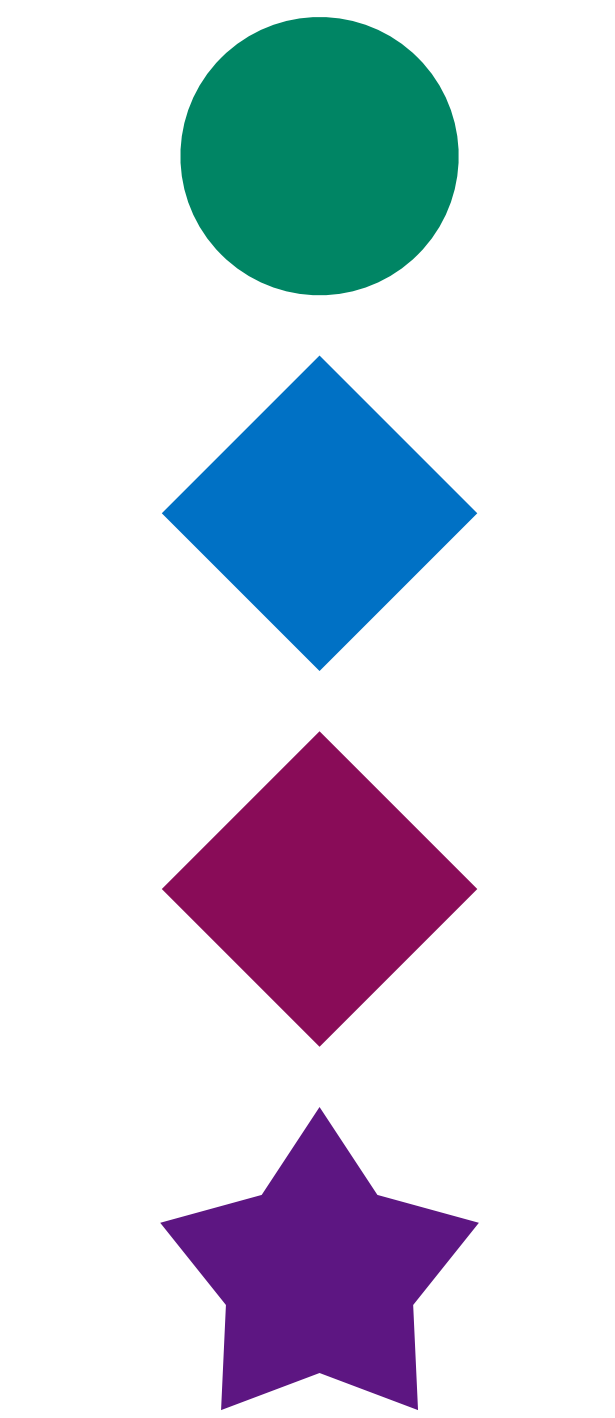
Filtering



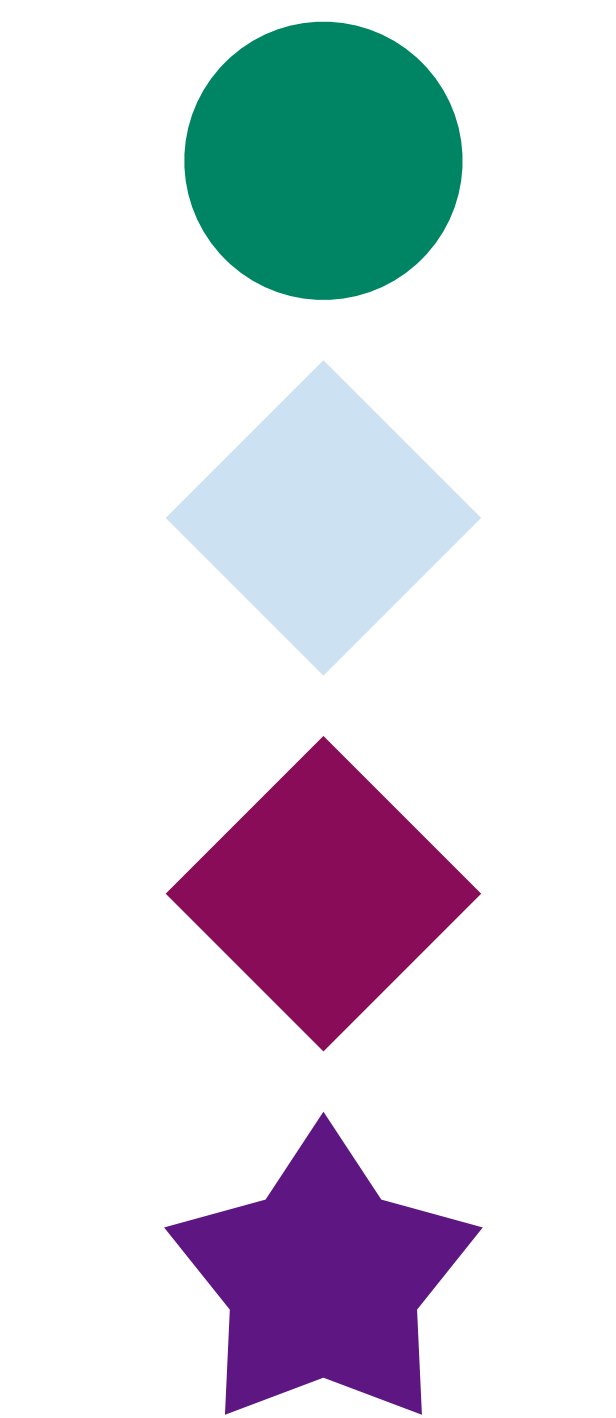
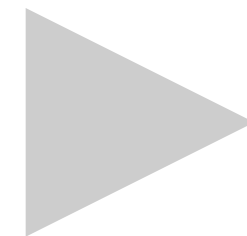
Scoring



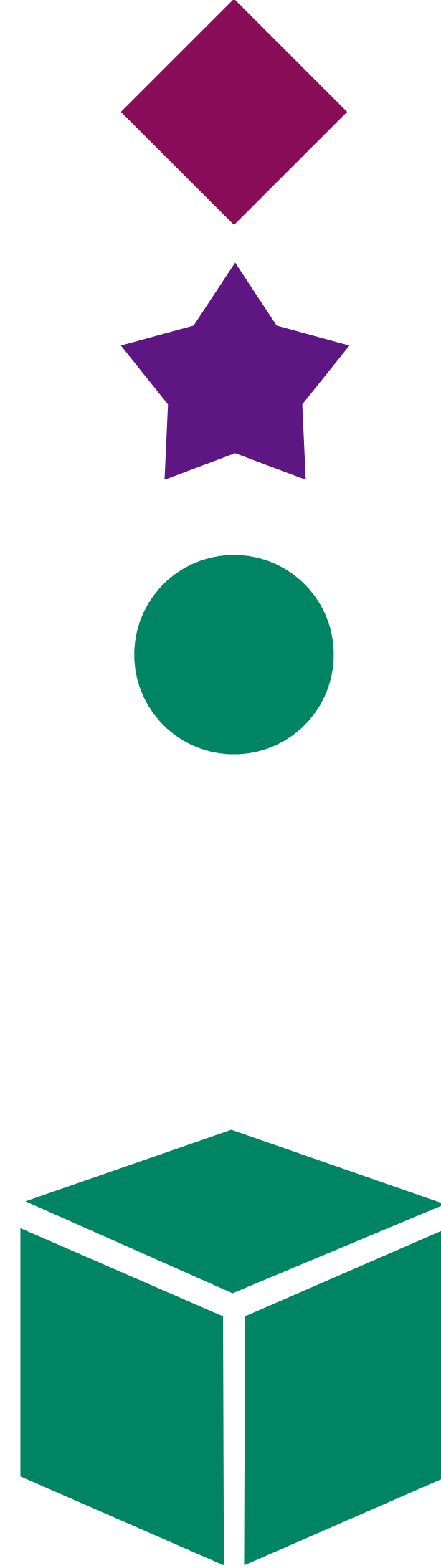
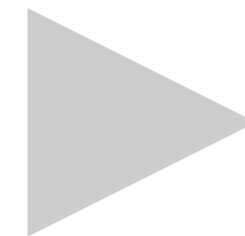
Ordering



Retrieval



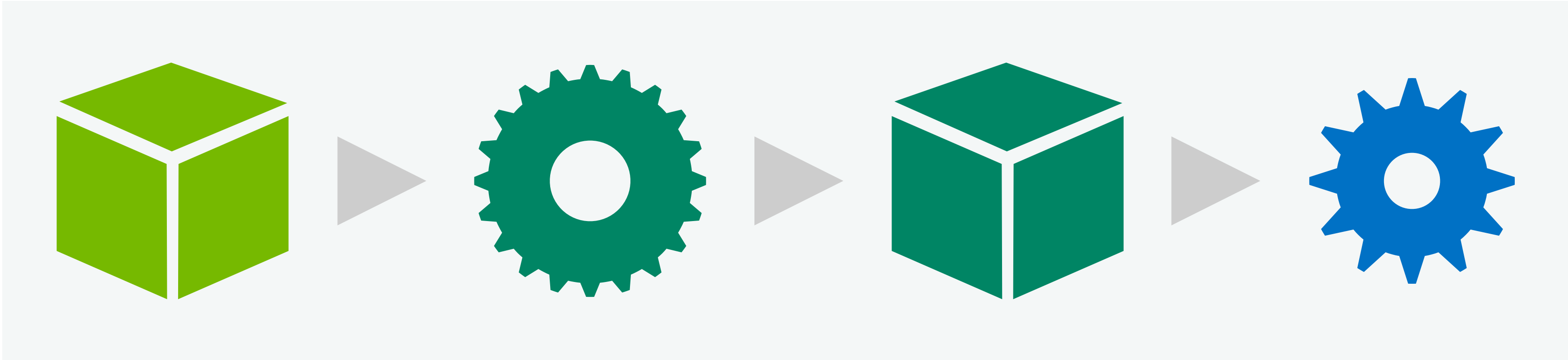
Filtering



Scoring

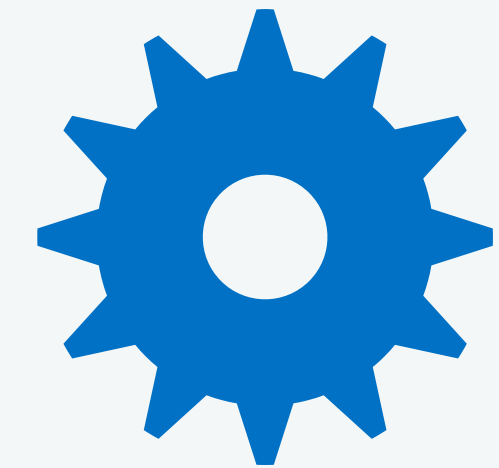
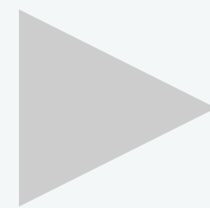
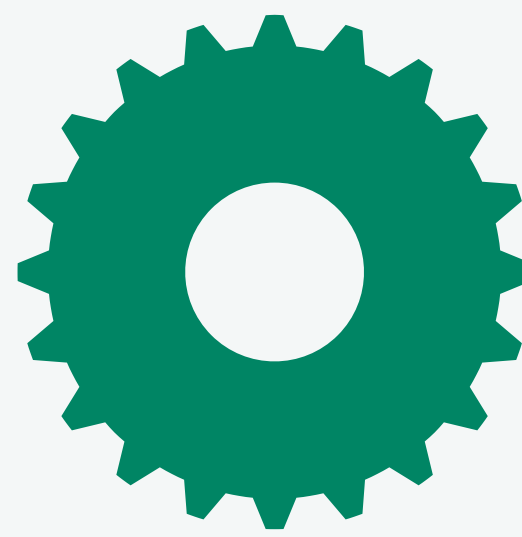
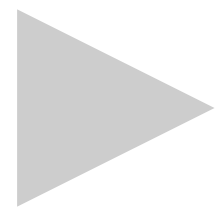


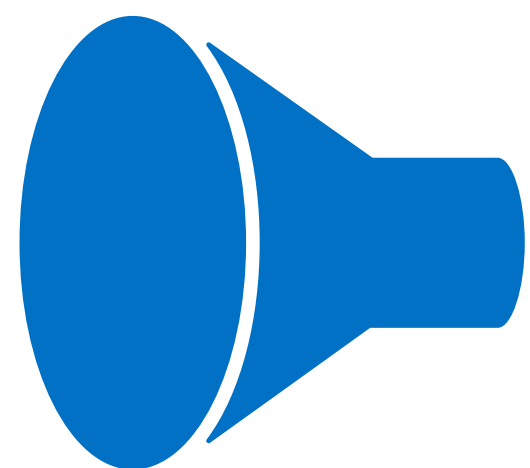
Ordering



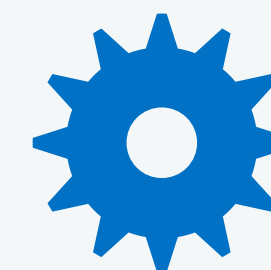


**Aggregation
and grouping**



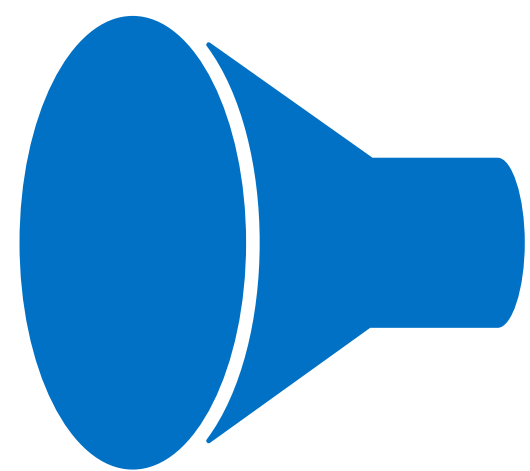


**Streaming
live data**

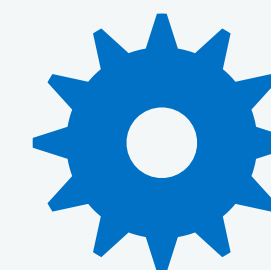
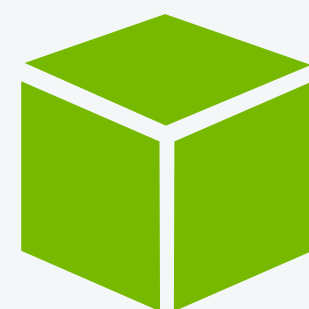


Predictions

Production inference



**Streaming
live data**

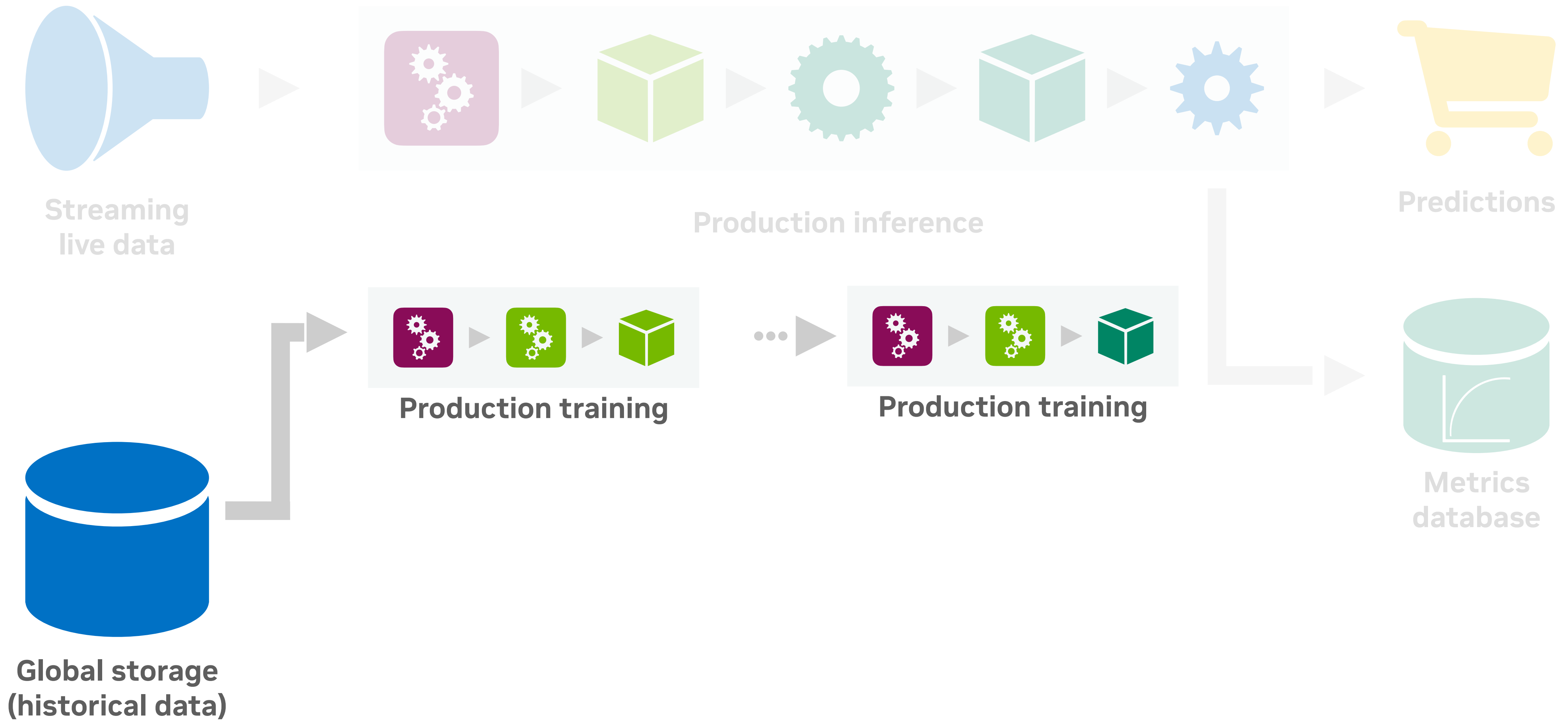


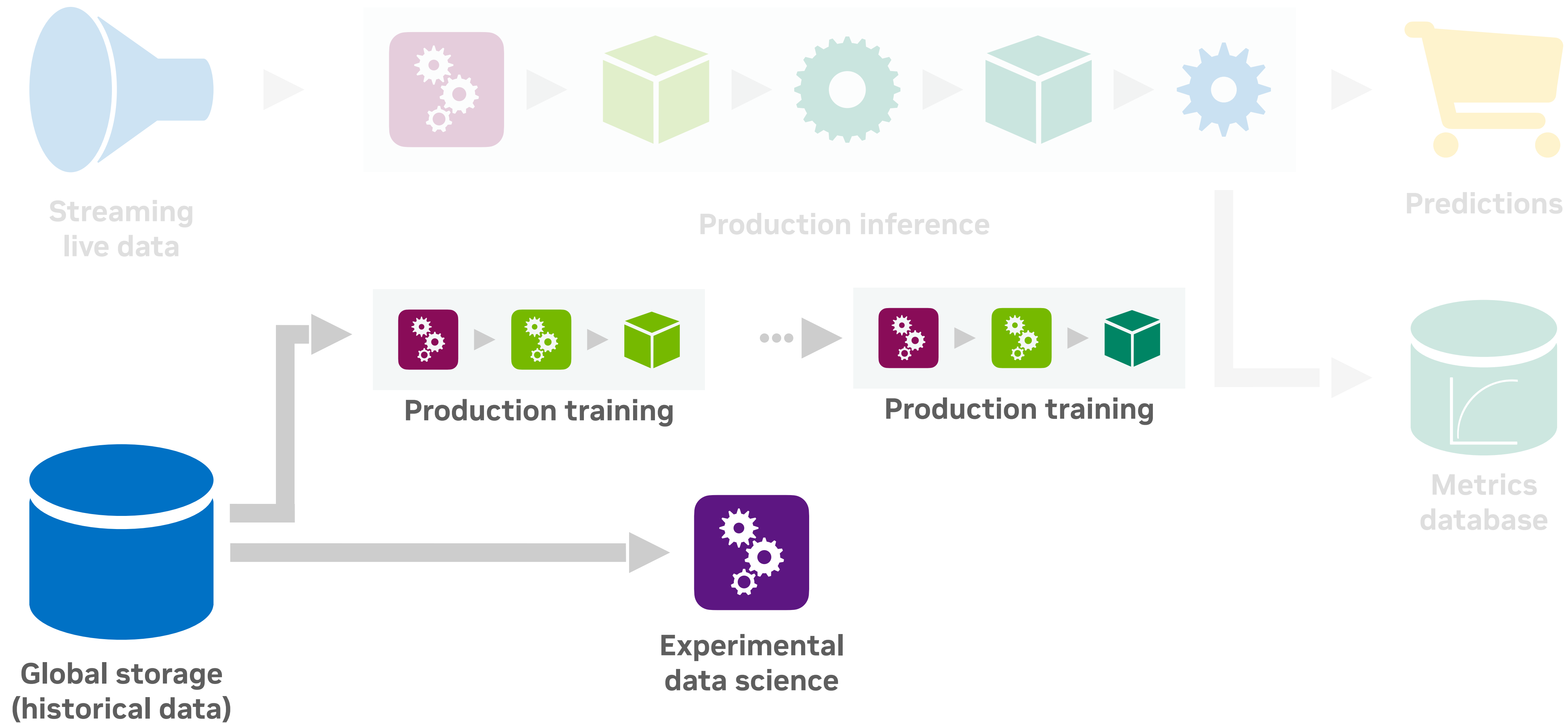
Predictions

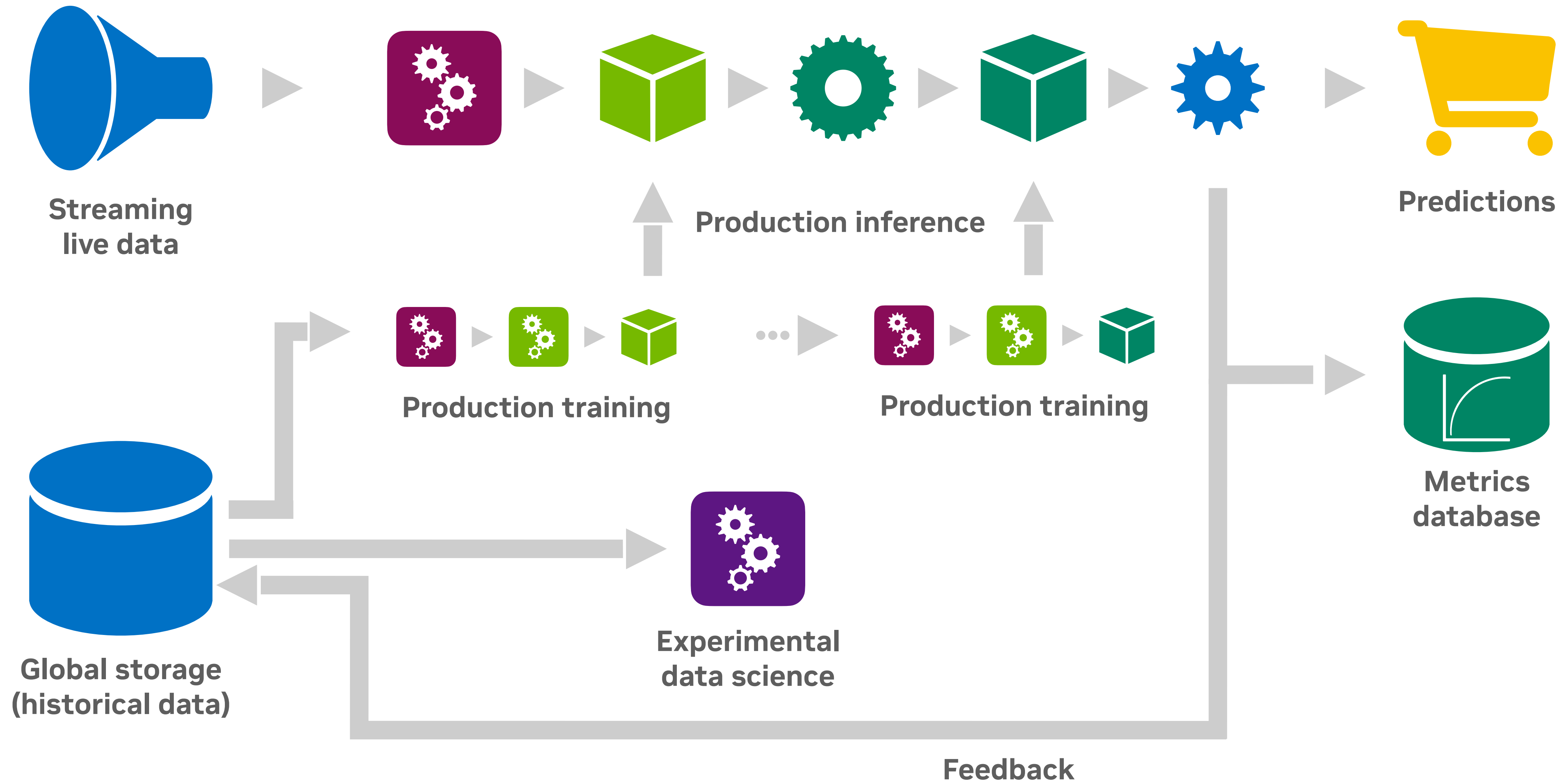


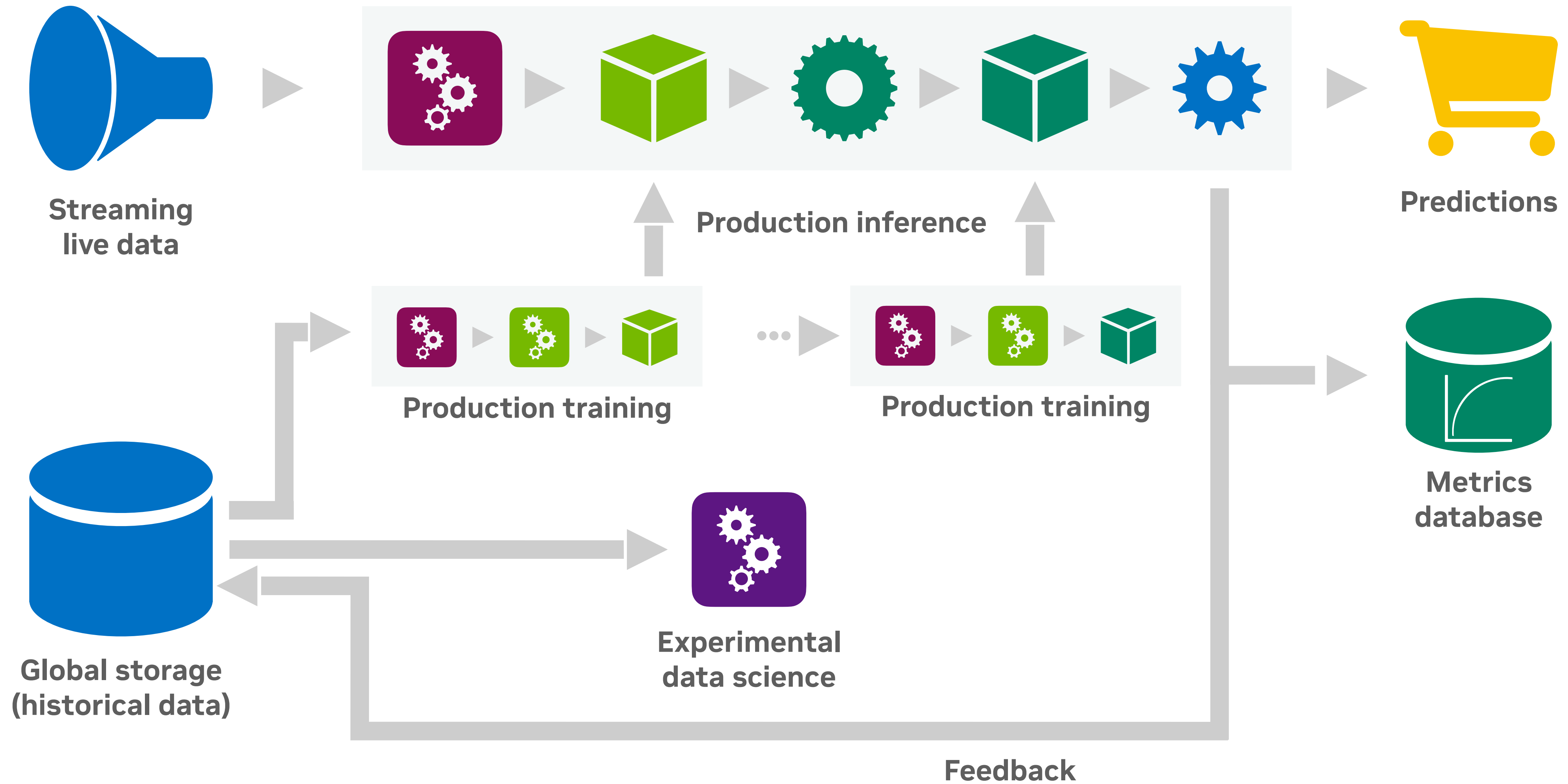
**Metrics
database**

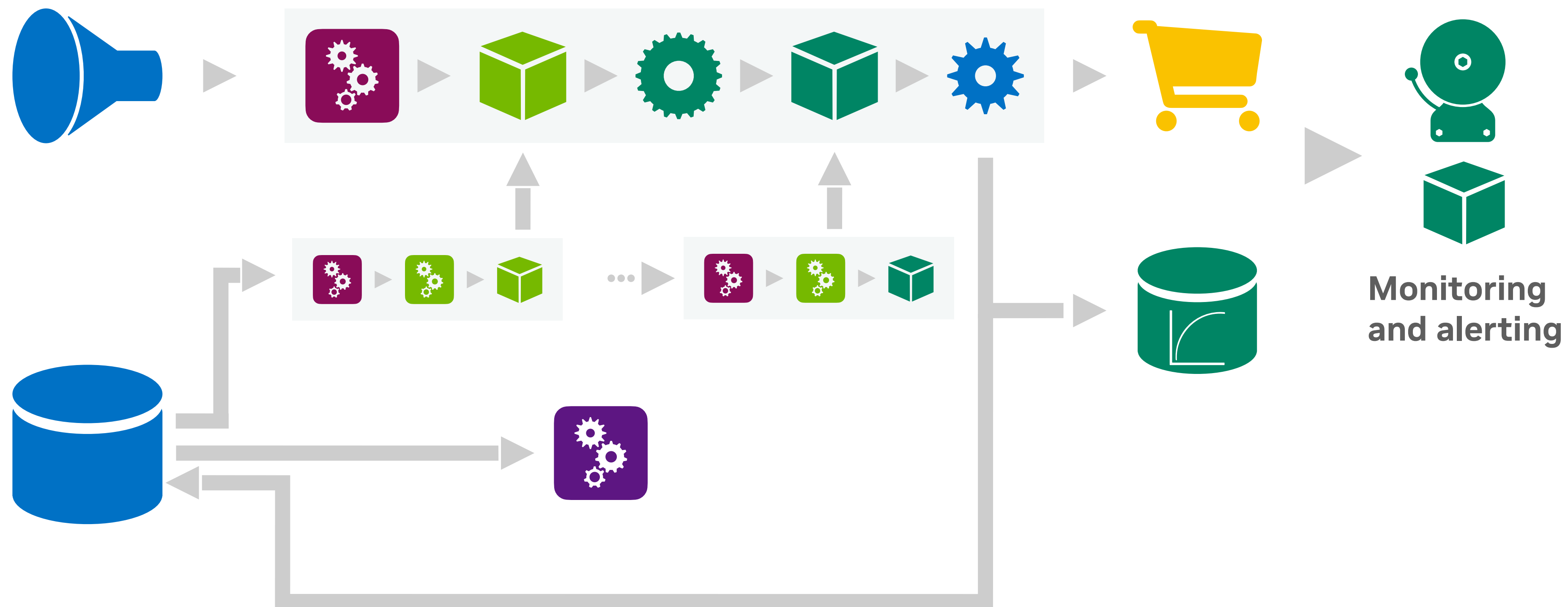
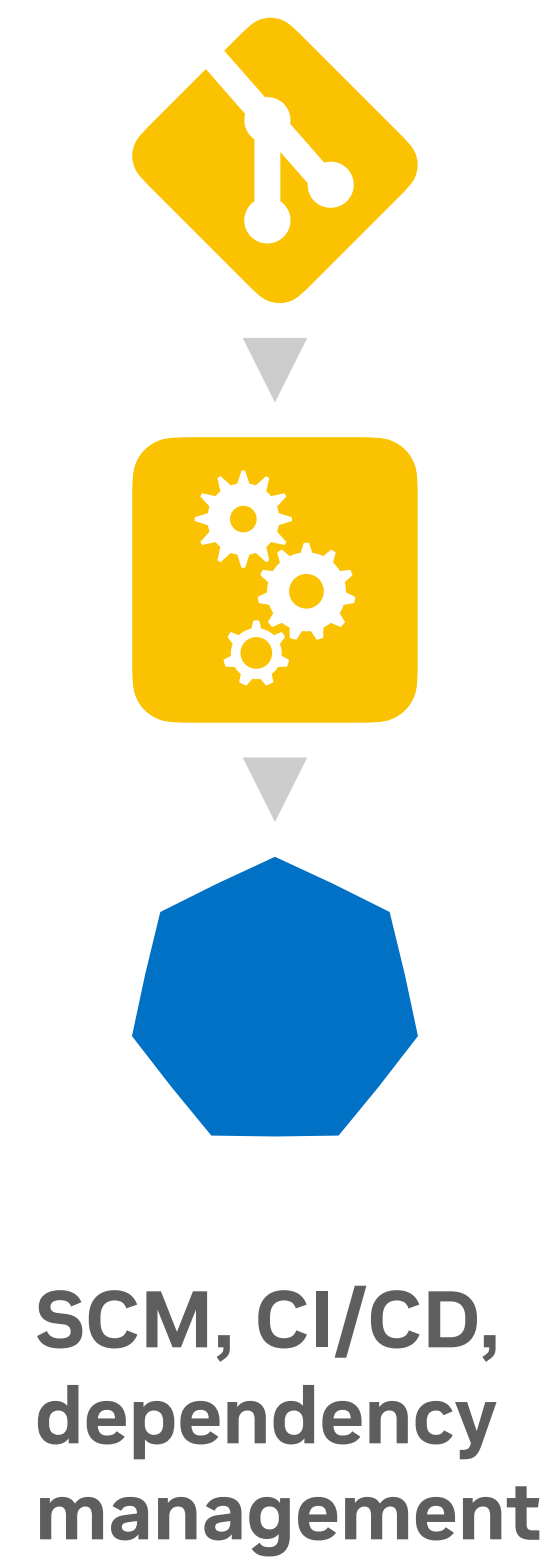
Production inference










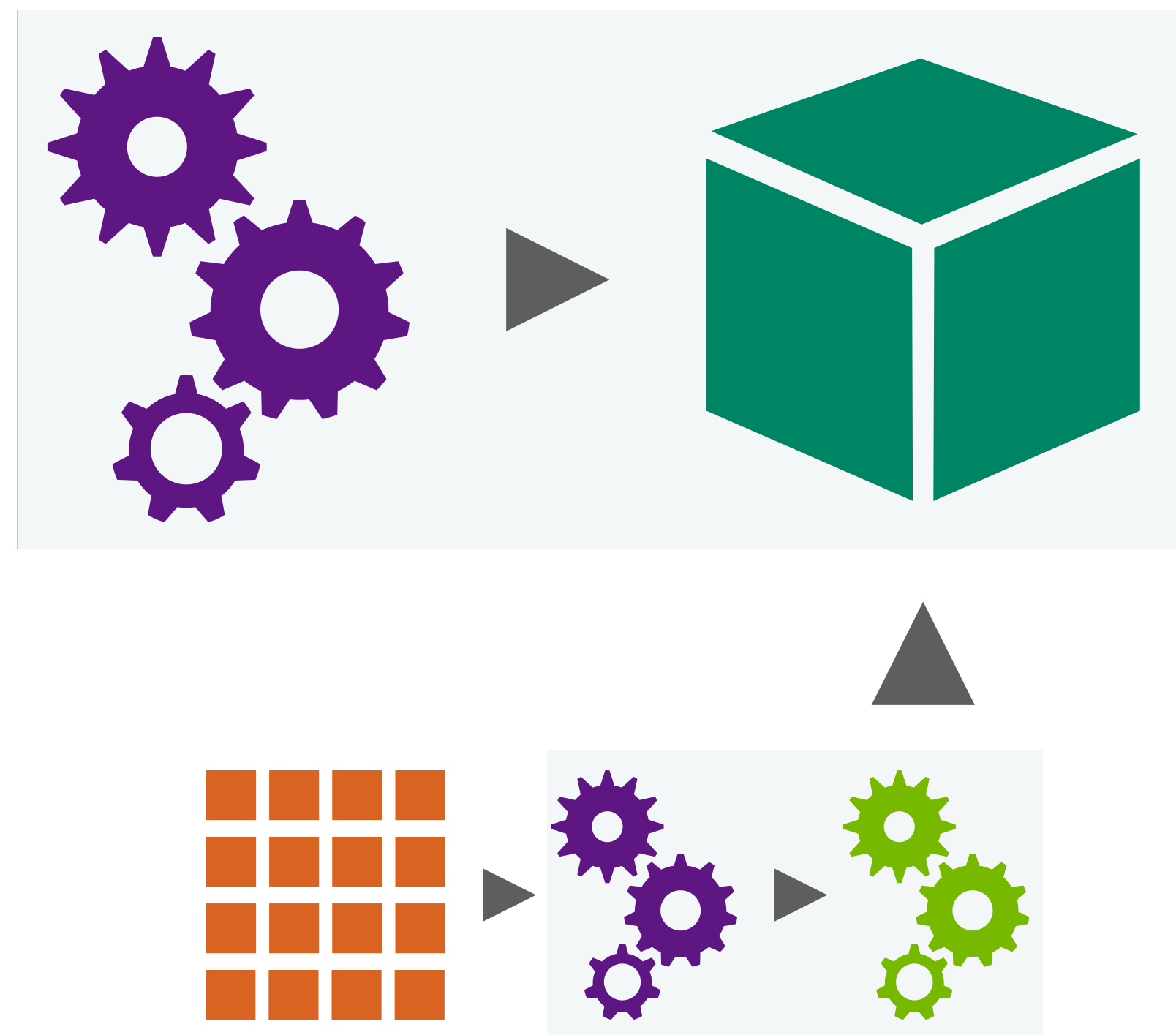


Infrastructure management

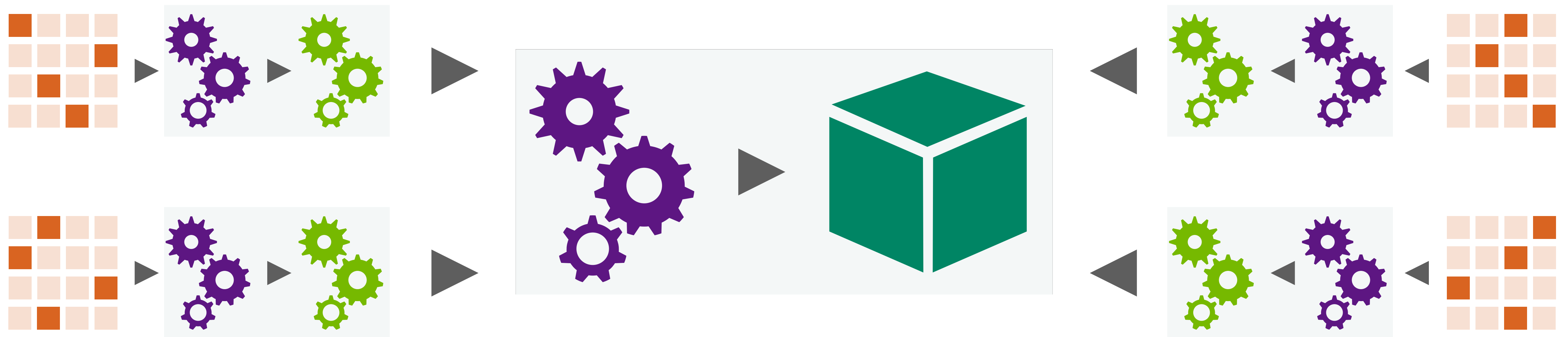


Other Special Considerations

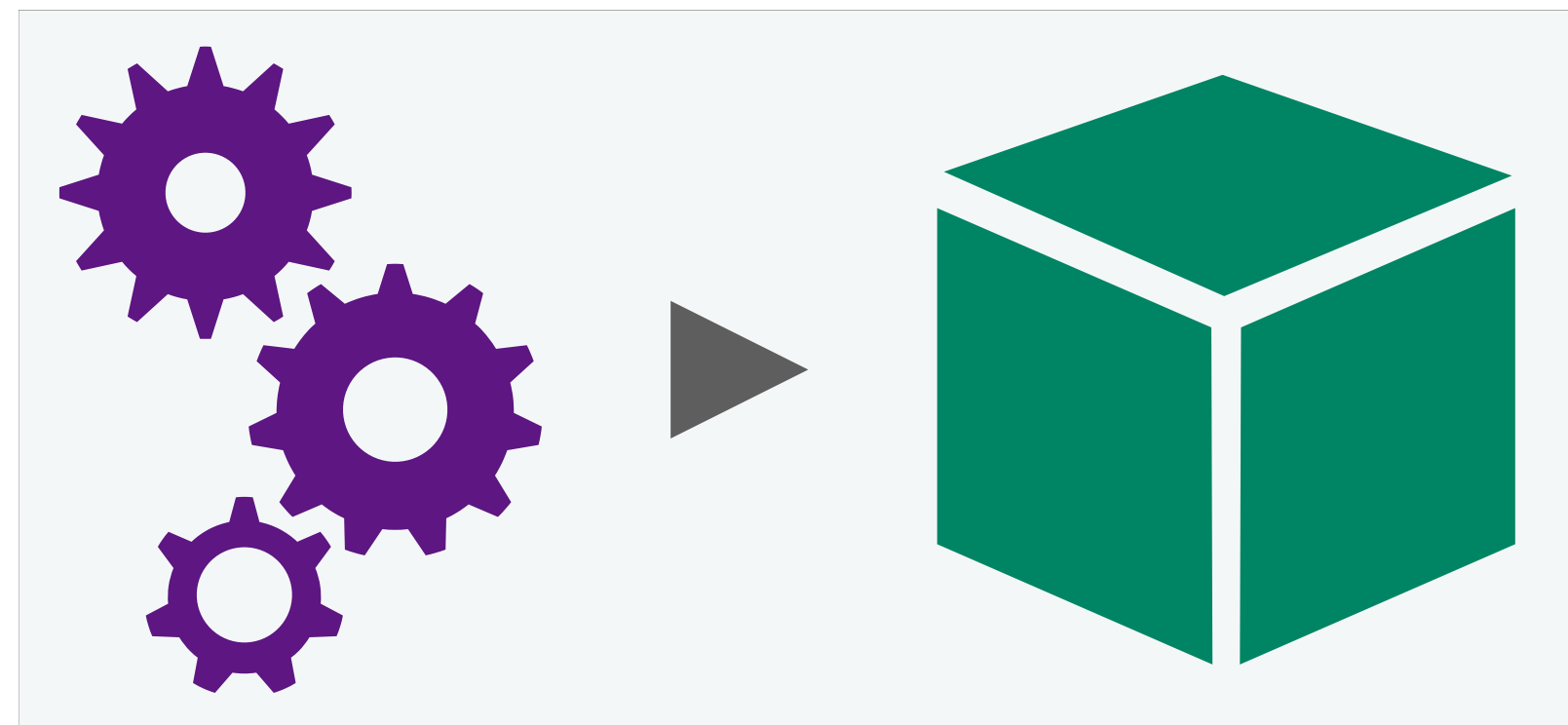
Privacy-preserving and distributed ML



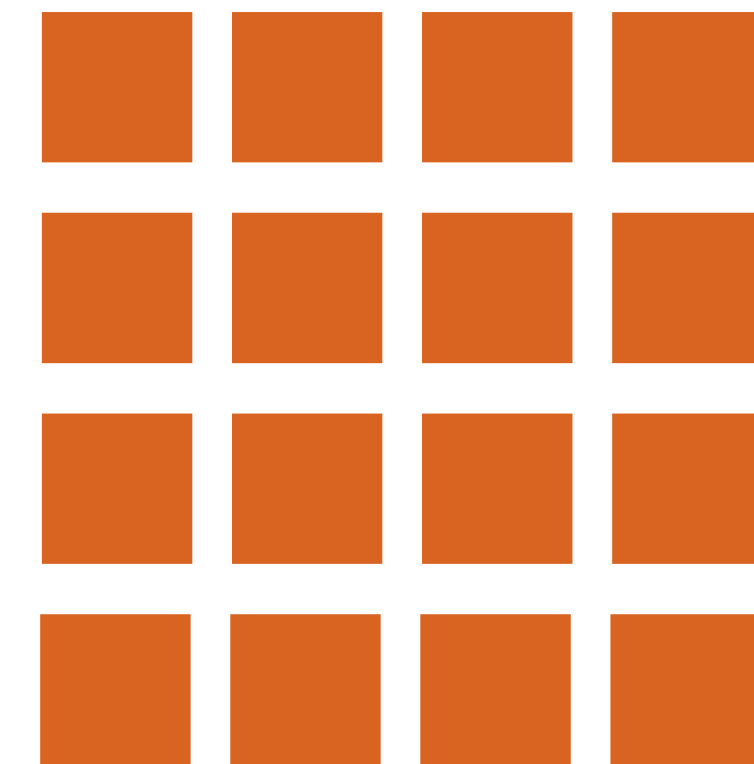
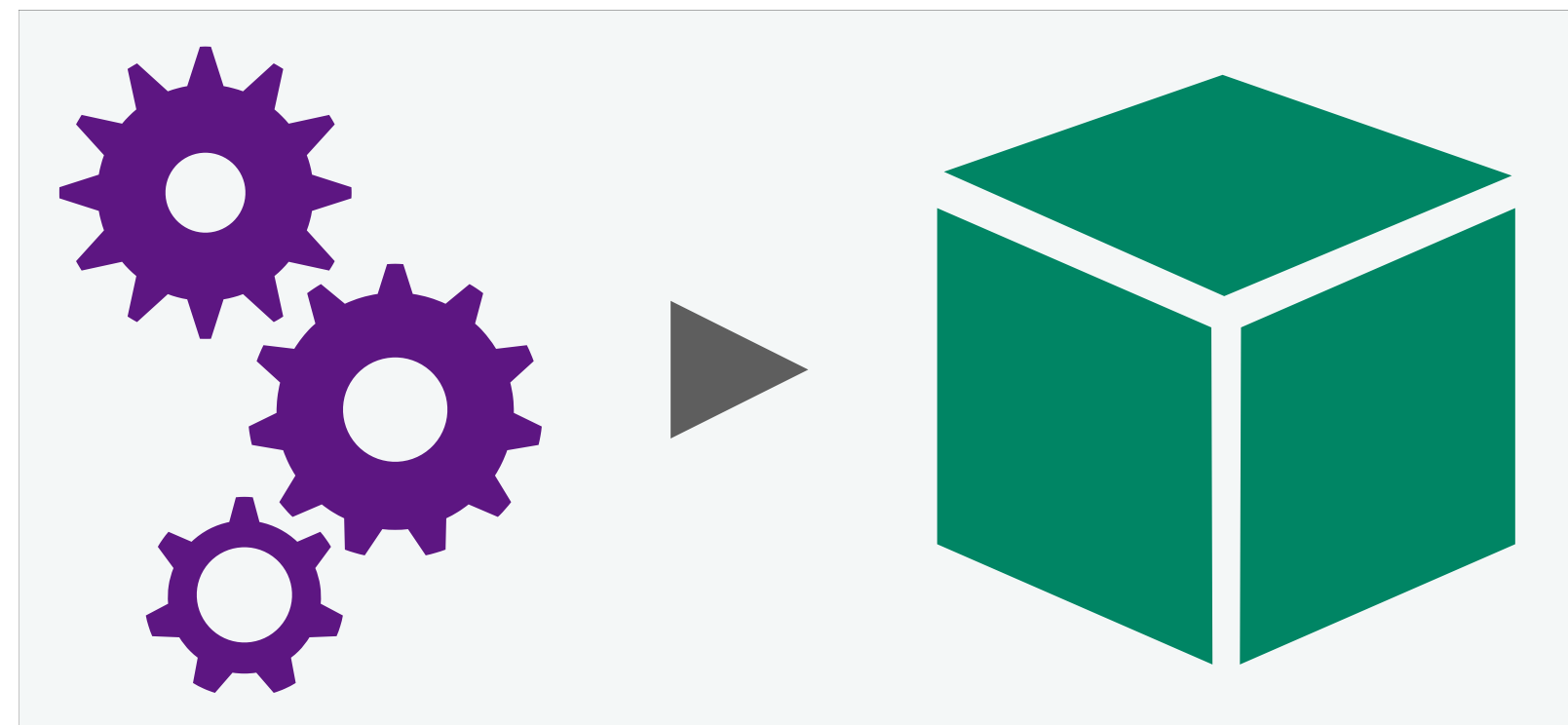
Privacy-preserving and distributed ML



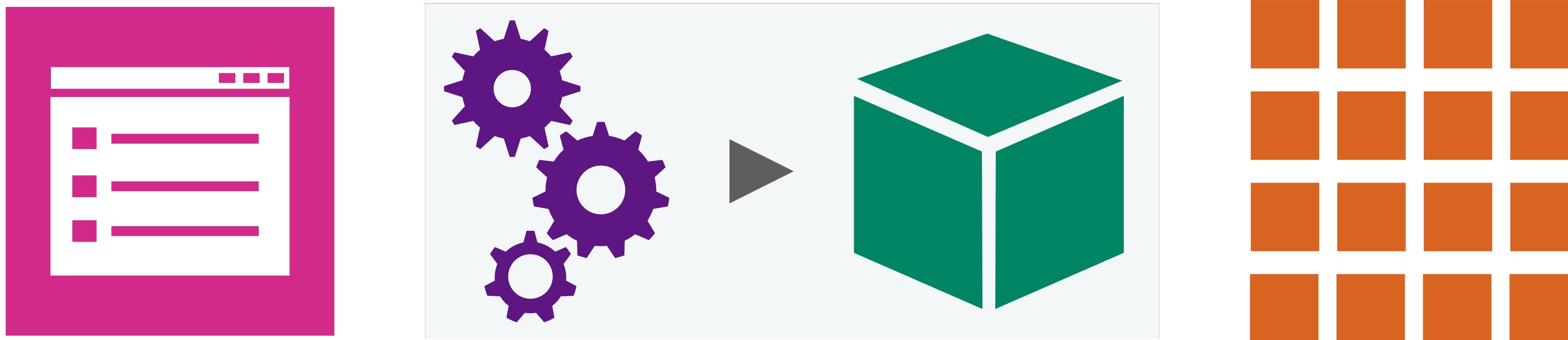
Privacy-preserving and distributed ML



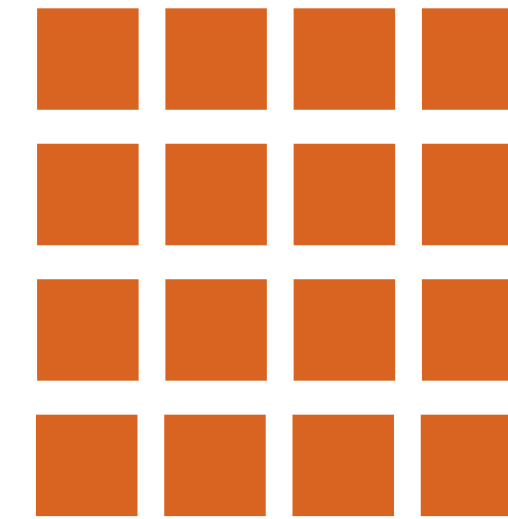
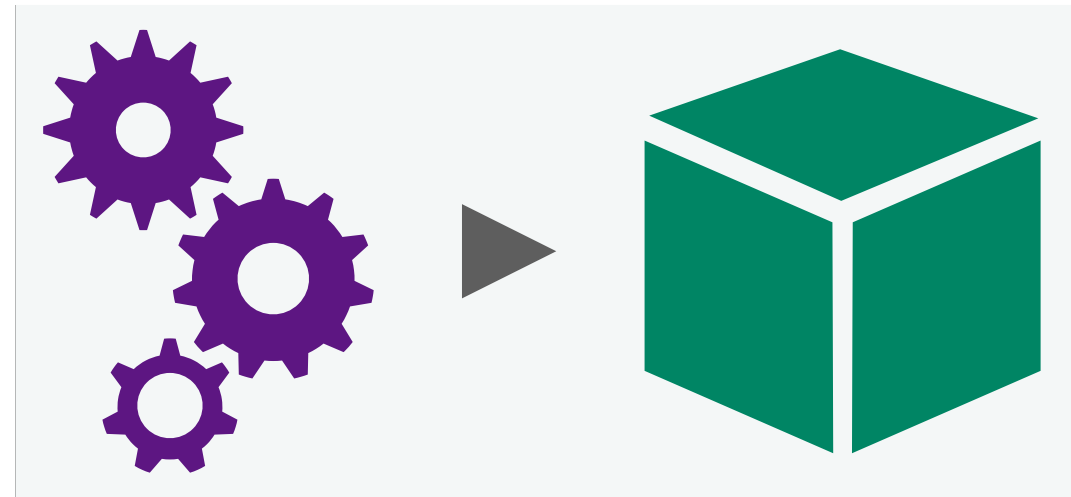
Privacy-preserving and distributed ML



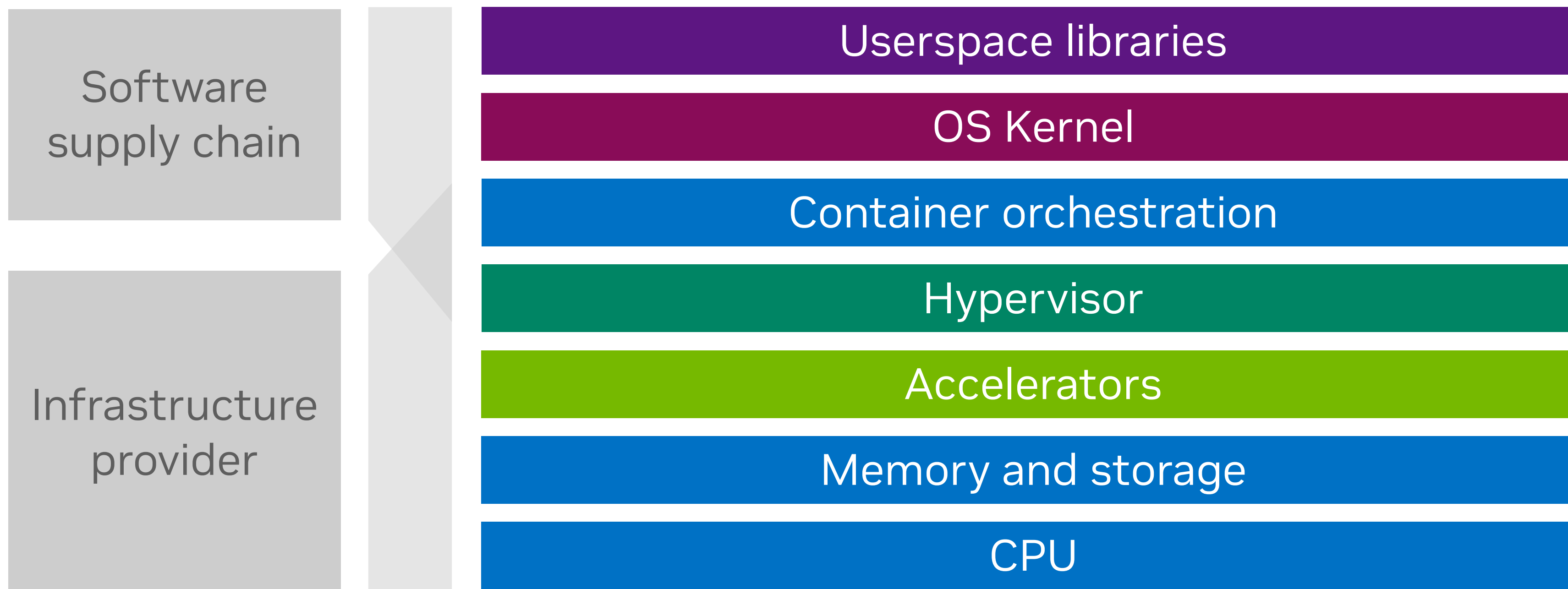
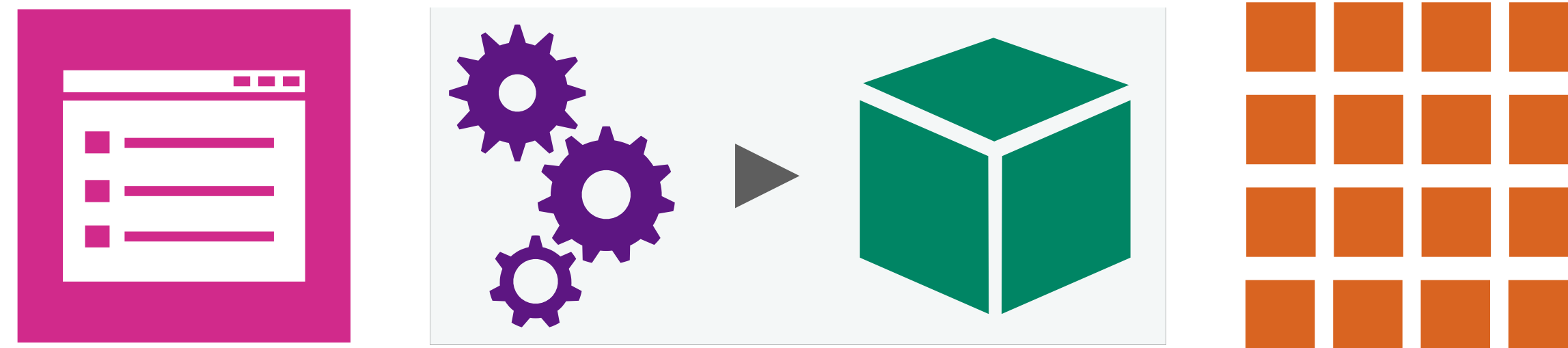
Privacy-preserving and distributed ML



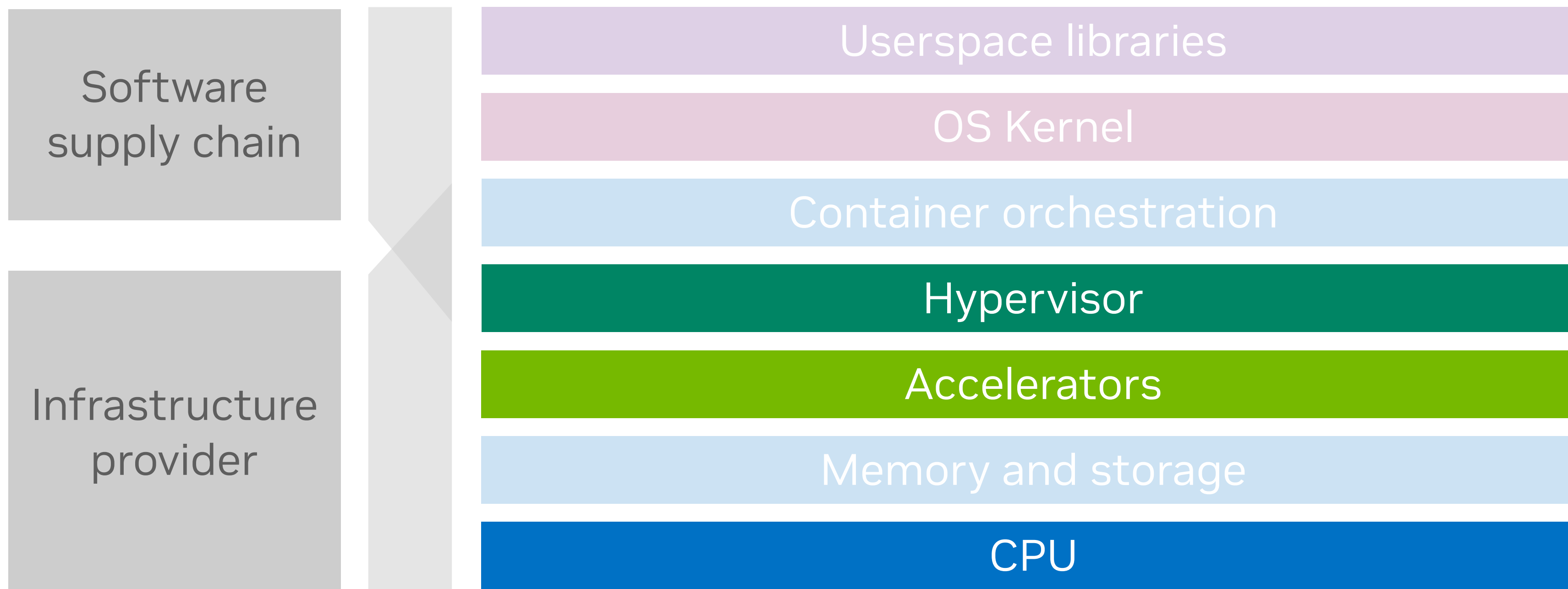
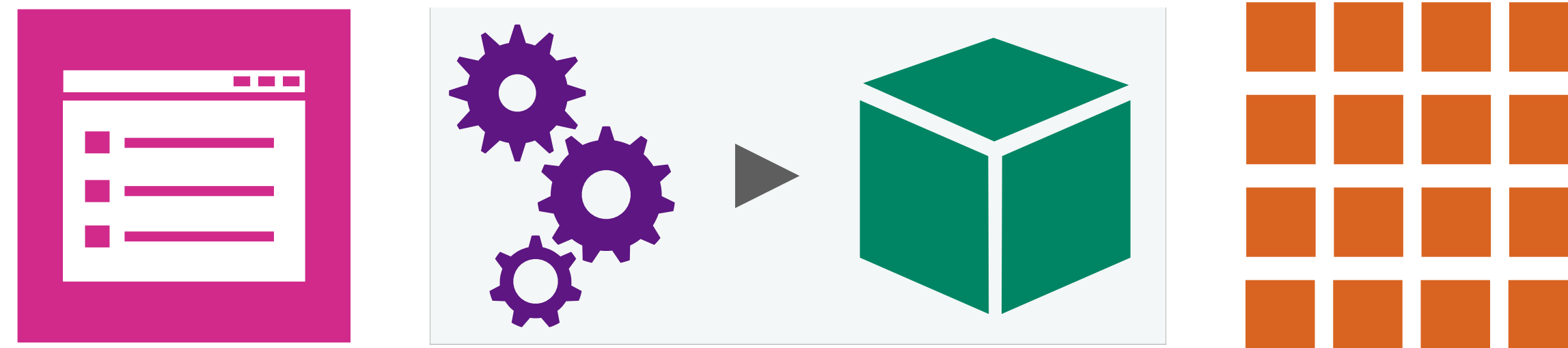
Privacy-preserving and distributed ML



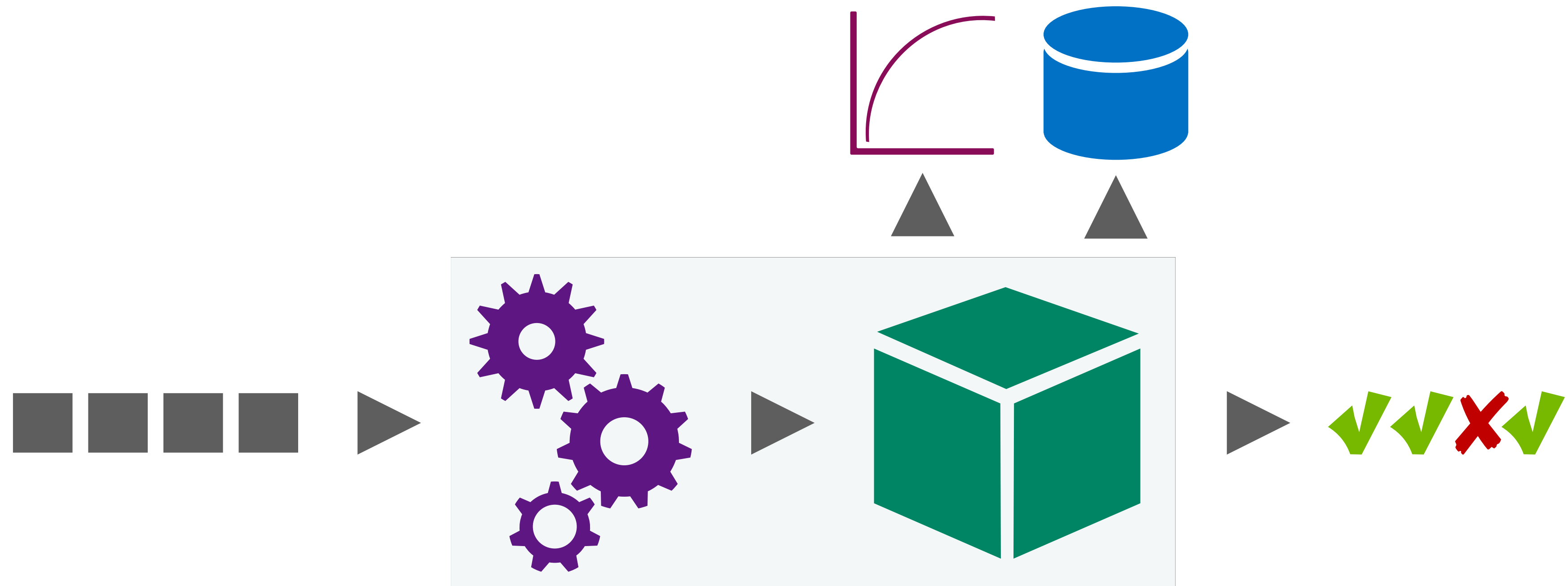
Privacy-preserving and distributed ML



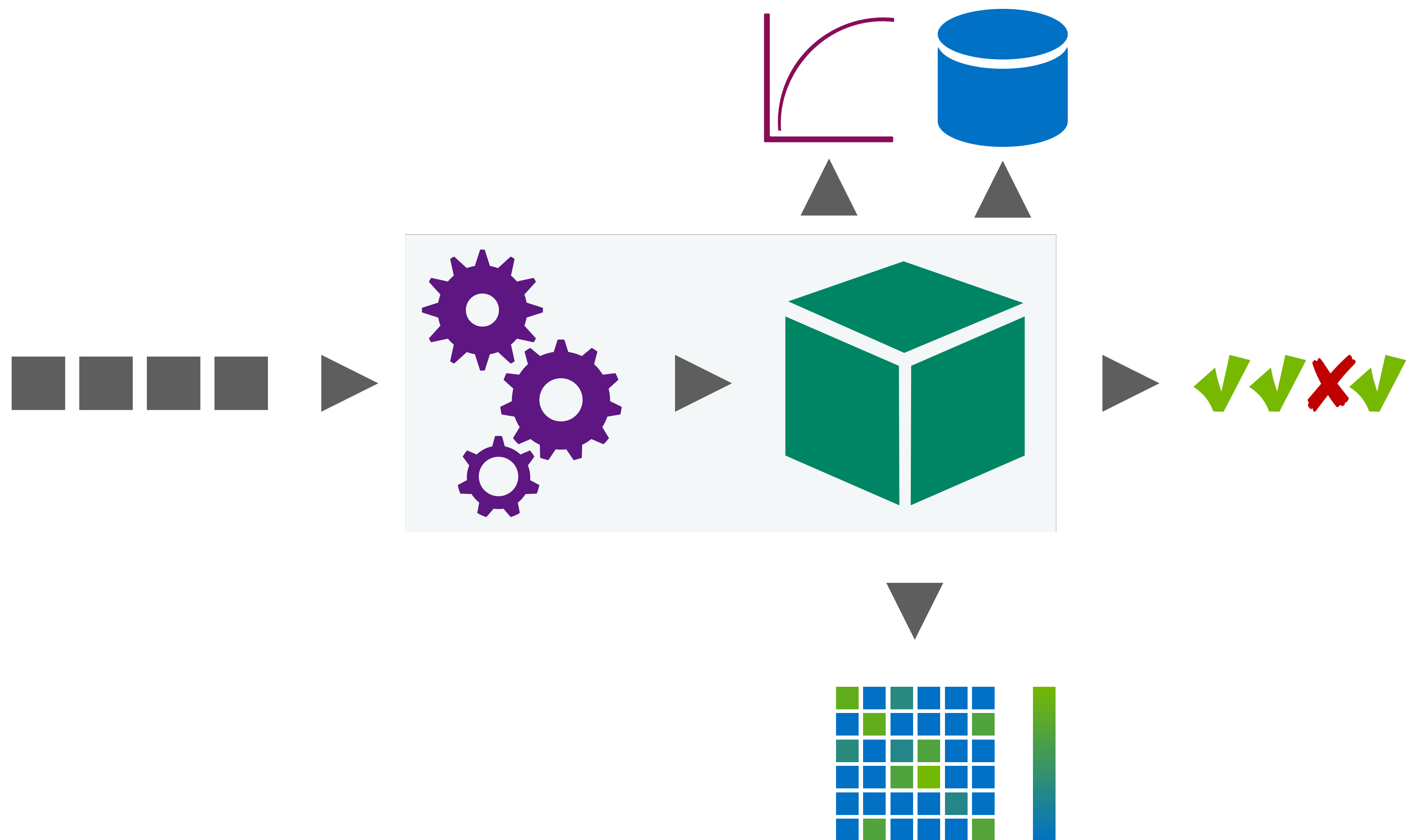
Privacy-preserving and distributed ML



Explaining predictions



Explaining predictions



Guardrails



Guardrails



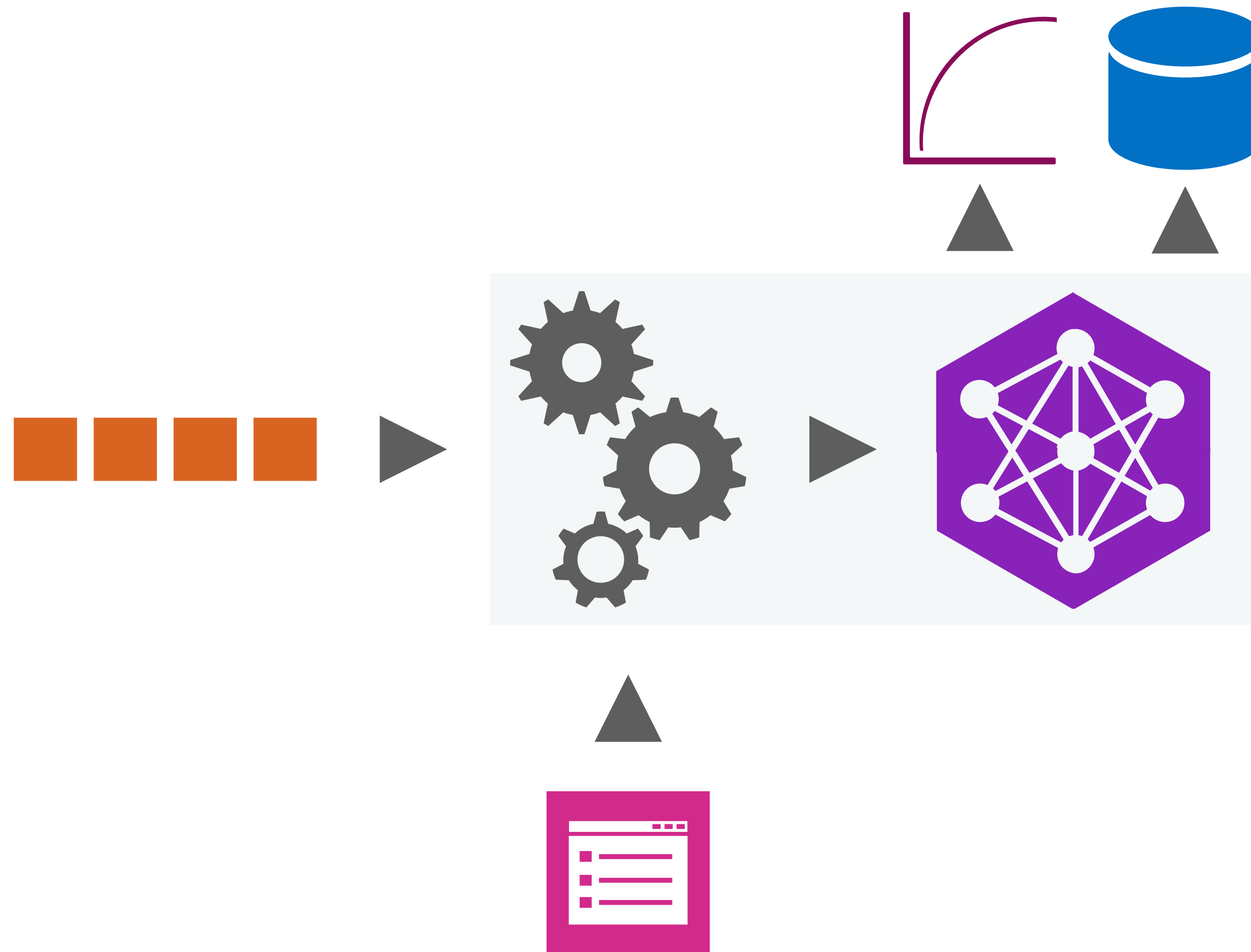
Guardrails



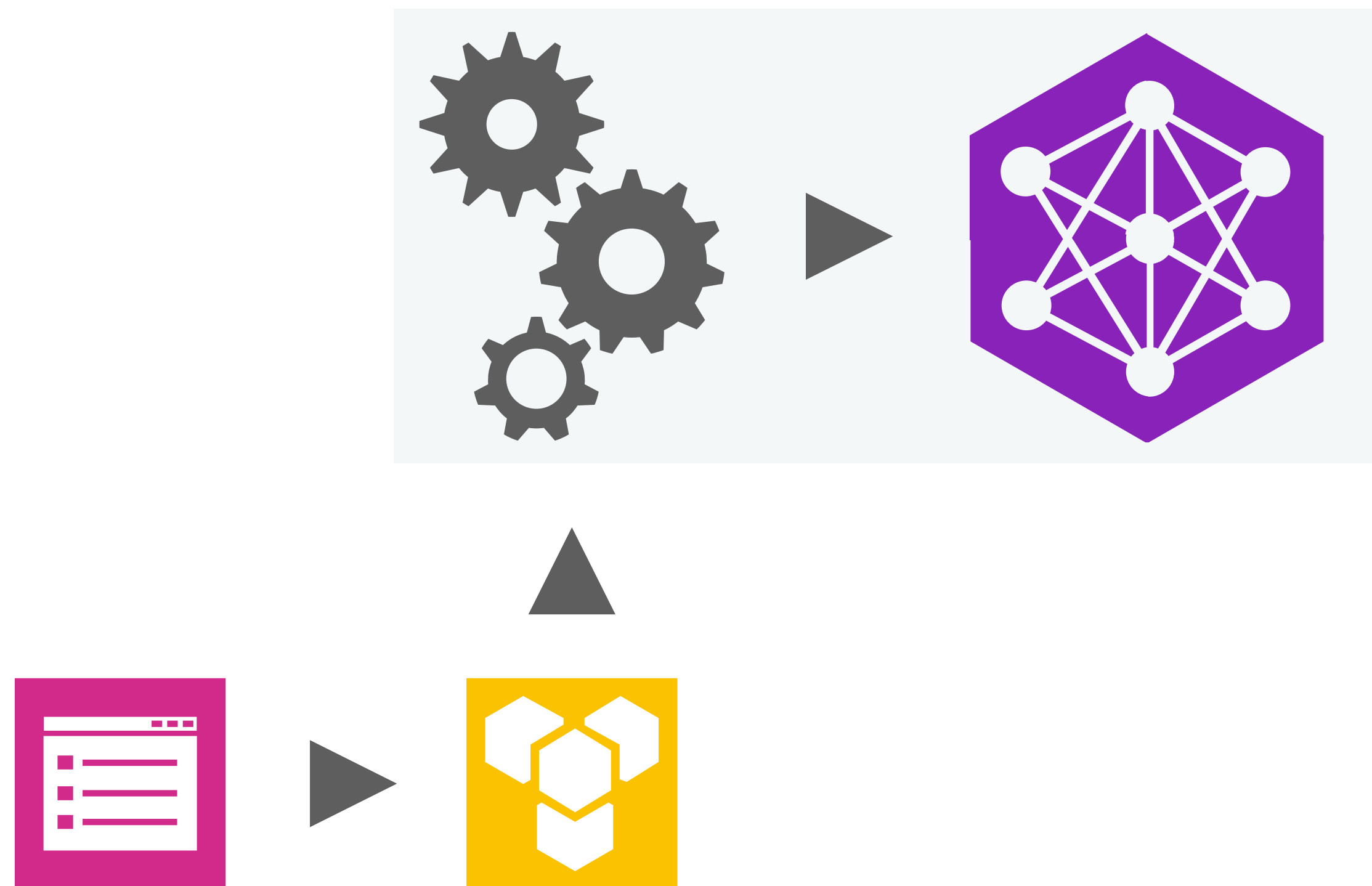
Guardrails



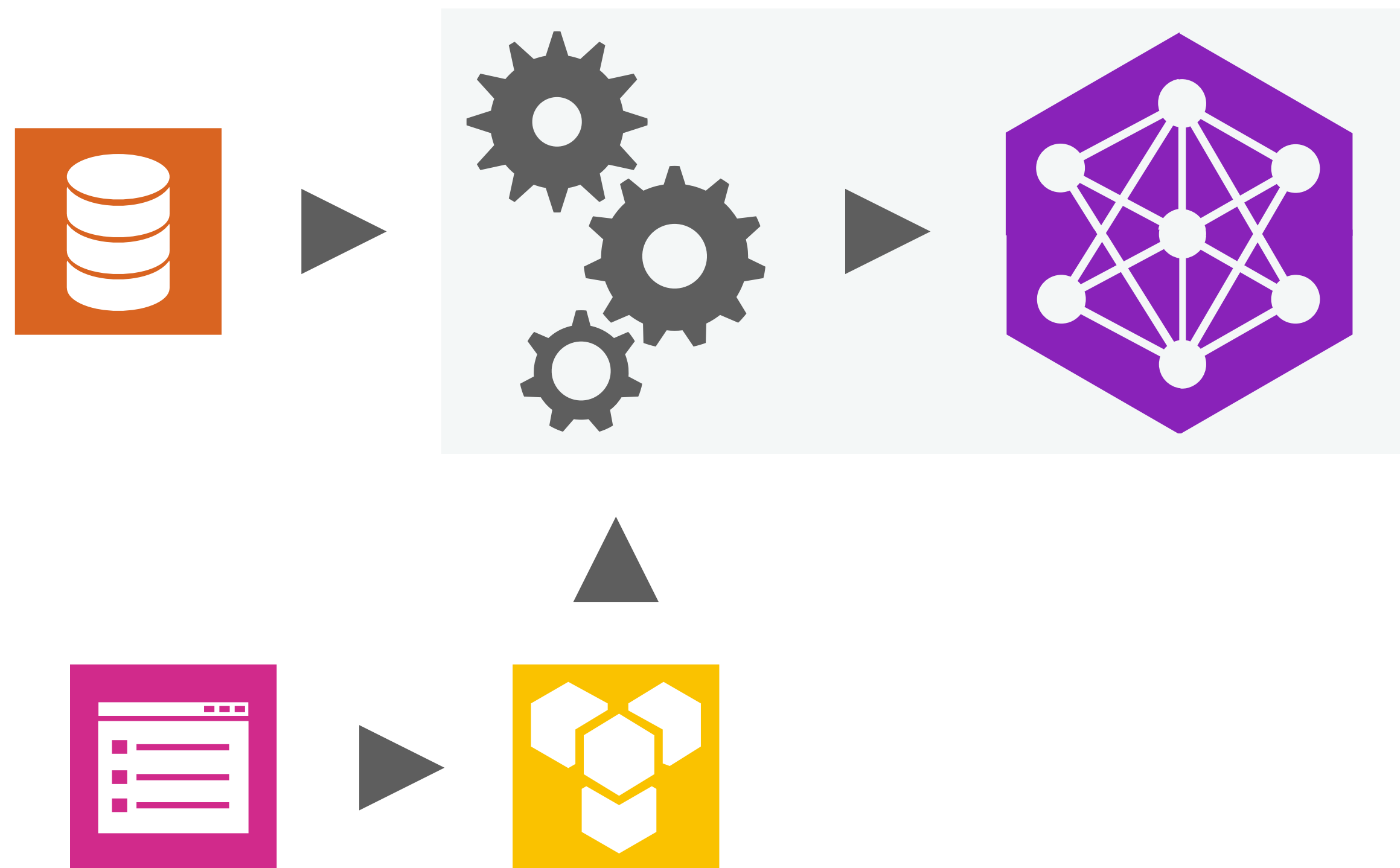
Special concerns for LLMs and generative AI



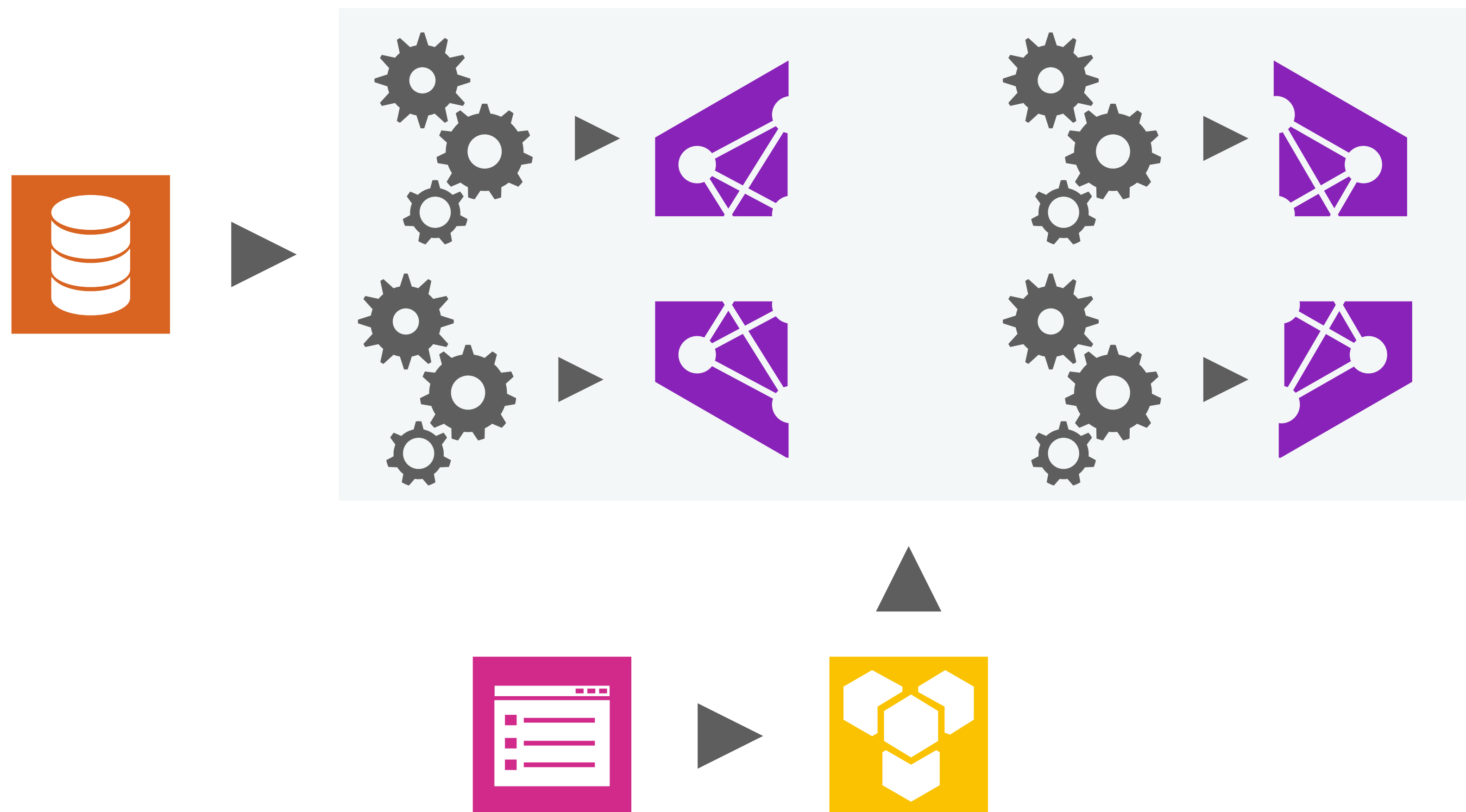
Special concerns for LLMs and generative AI



Special concerns for LLMs and generative AI




Special concerns for LLMs and generative AI




To learn more...

[S62427] Confidential Computing: New Features and NVIDIA Hardware Attestation (Michael O'Connor et al.)

 [S62149] Decentralized Collaborative AI With Federated Learning in Trustworthy Environments (Emily Sakata)

 [S62960] XGBoost is All You Need (Bojan Tunguz)

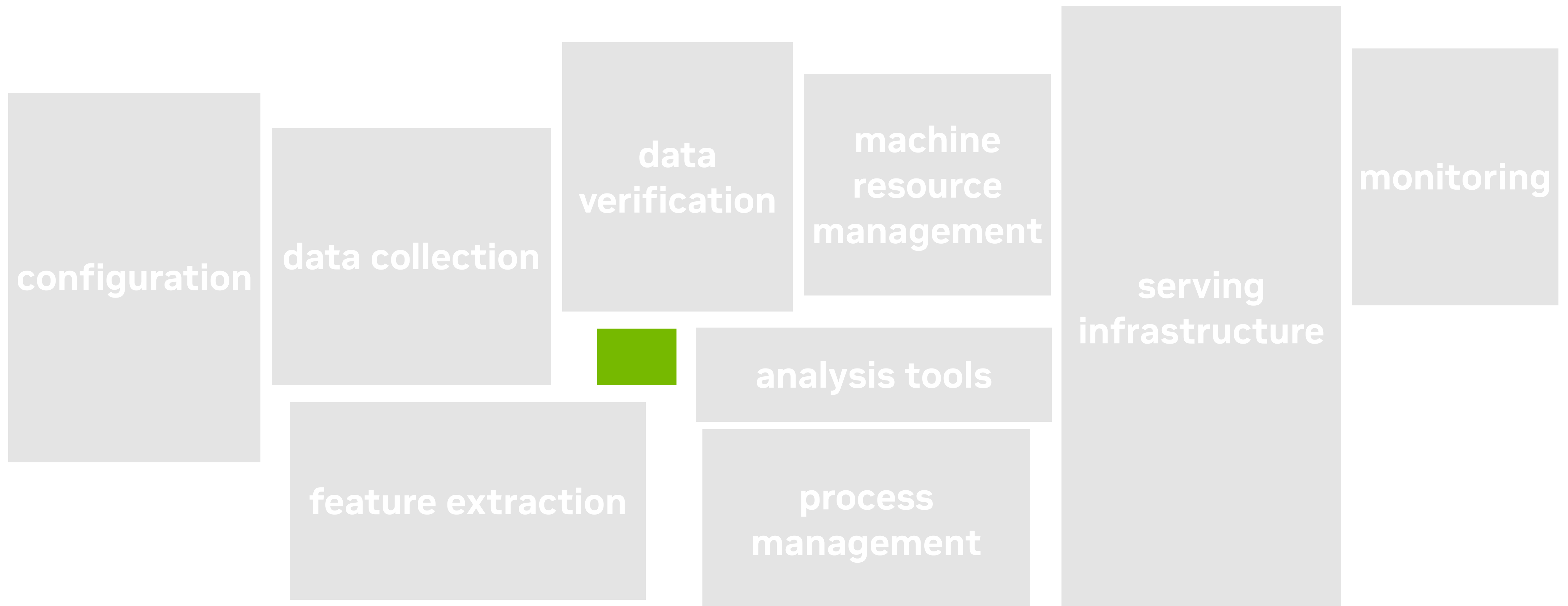
[S62458] LLMOps: The New Frontier of Machine Learning Operations (Nik Spirin and Michael Balint)



Defining MLOps, revisited

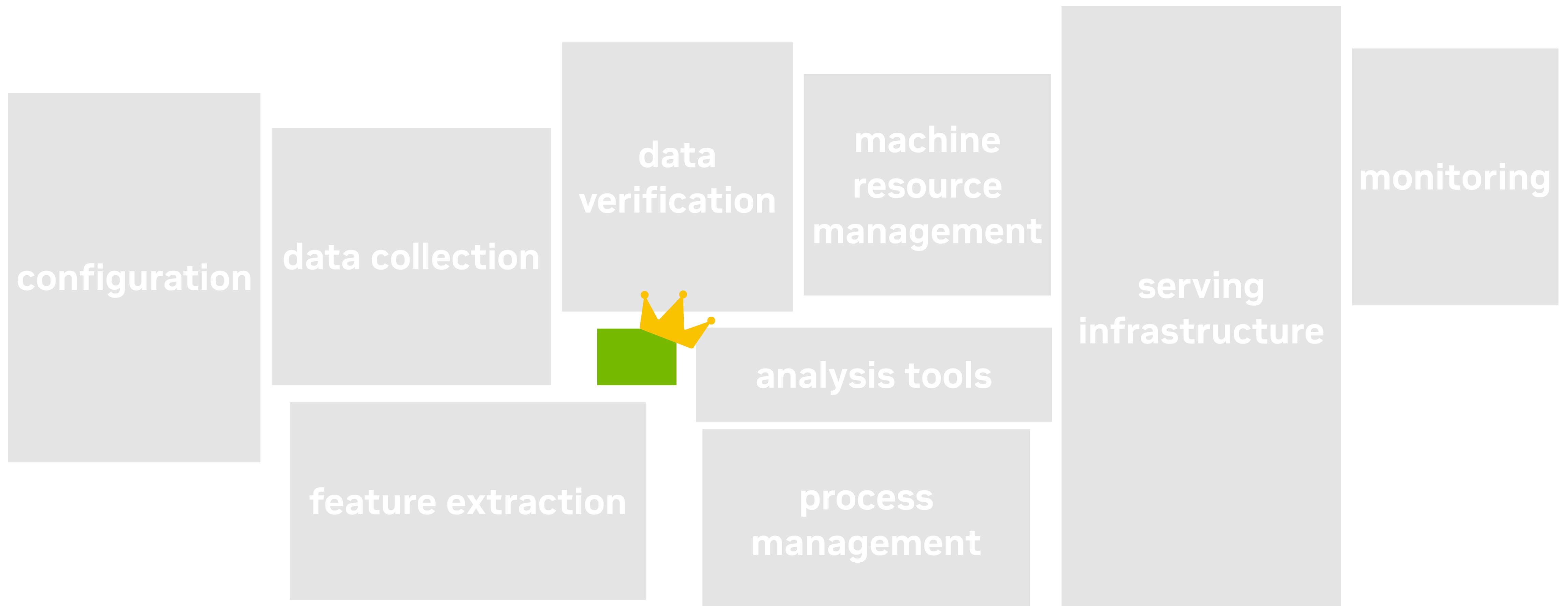
Machine learning systems

(source: Sculley et al., “Hidden Technical Debt in Machine Learning Systems,” in *NIPS 2015*.)



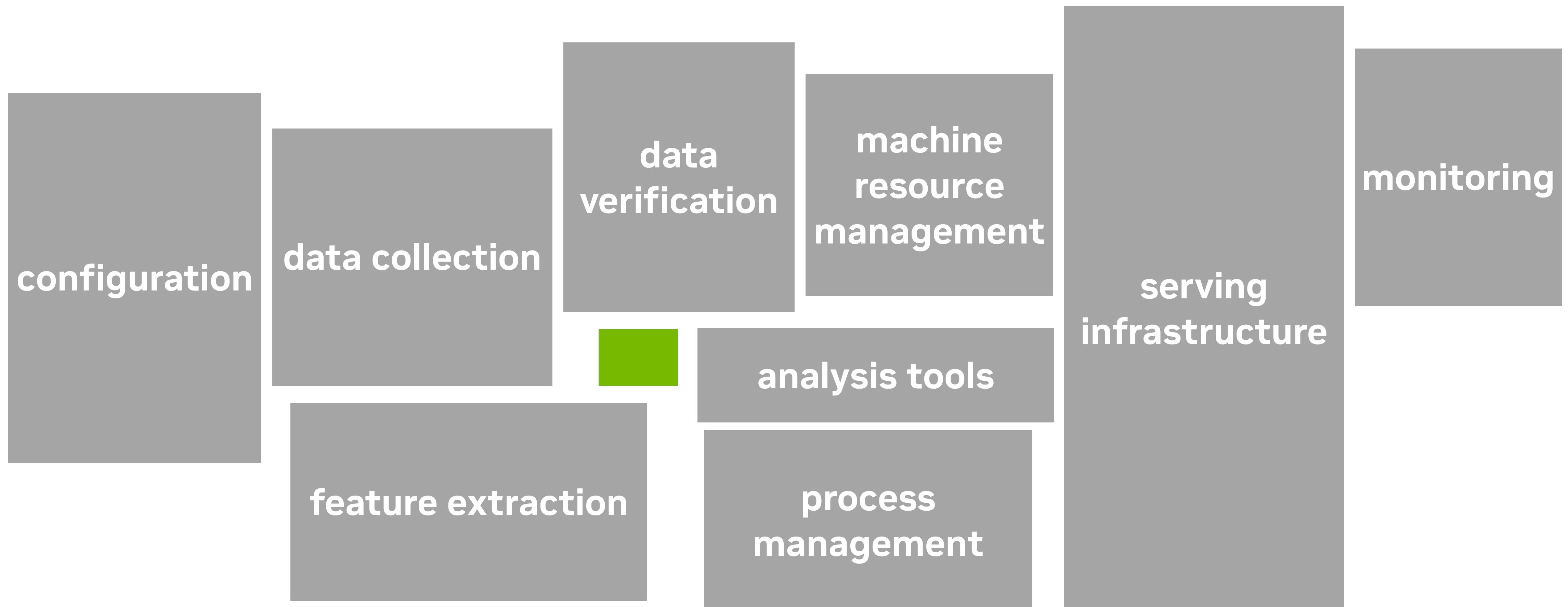
Machine learning systems

(source: Sculley et al., “Hidden Technical Debt in Machine Learning Systems,” in *NIPS 2015*.)



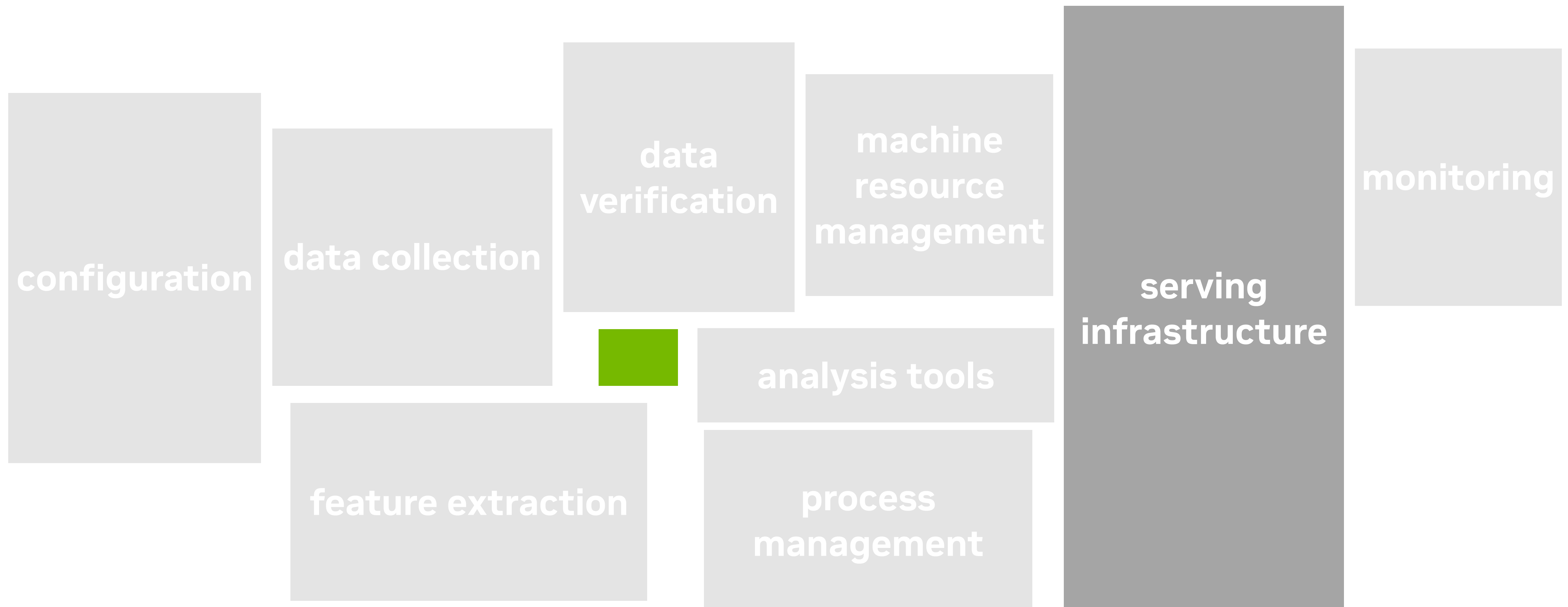
Machine learning systems

(source: Sculley et al., “Hidden Technical Debt in Machine Learning Systems,” in *NIPS 2015*.)



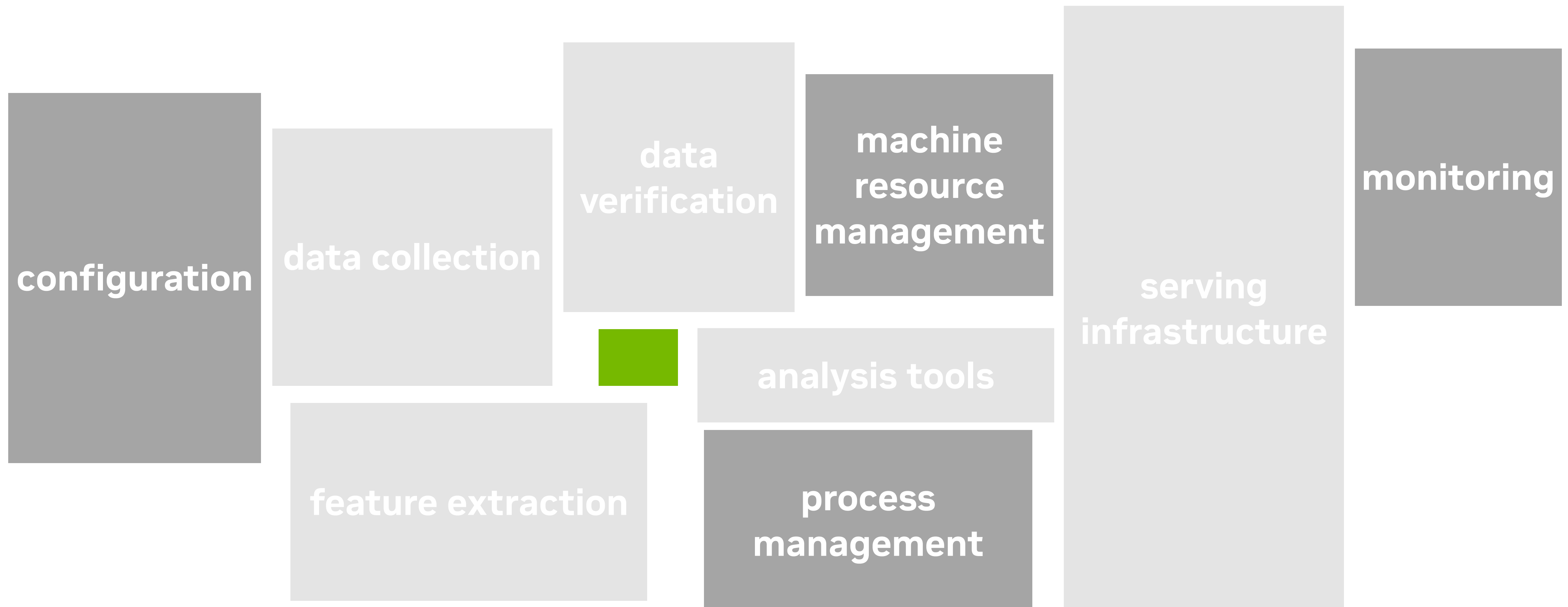
Machine learning systems

(source: Sculley et al., “Hidden Technical Debt in Machine Learning Systems,” in *NIPS 2015*.)



Machine learning systems

(source: Sculley et al., “Hidden Technical Debt in Machine Learning Systems,” in *NIPS 2015*.)



“End-to-end” / ML platforms

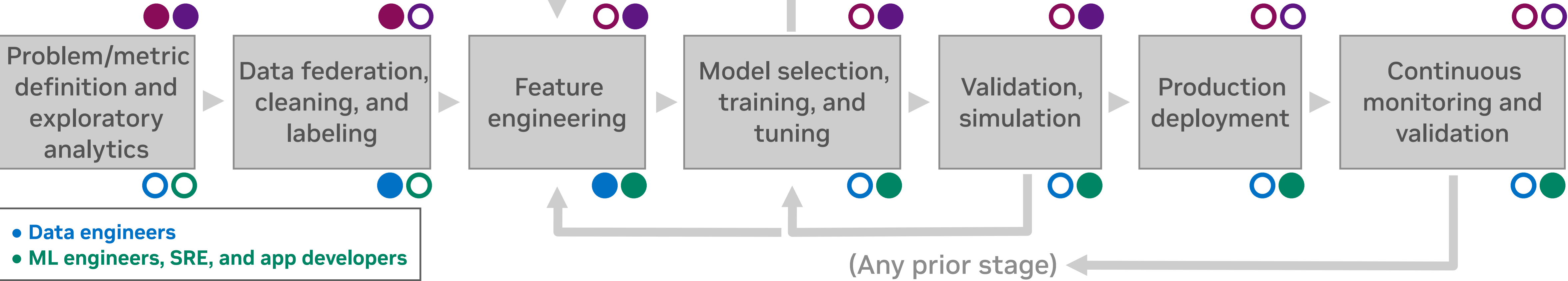
infrastructure agnostic

Data wrangling, analytics, and labeling

Interactive development

Experiment management and AutoML

- Data scientists
- Business analysts



- Data engineers
- ML engineers, SRE, and app developers

infrastructure aware

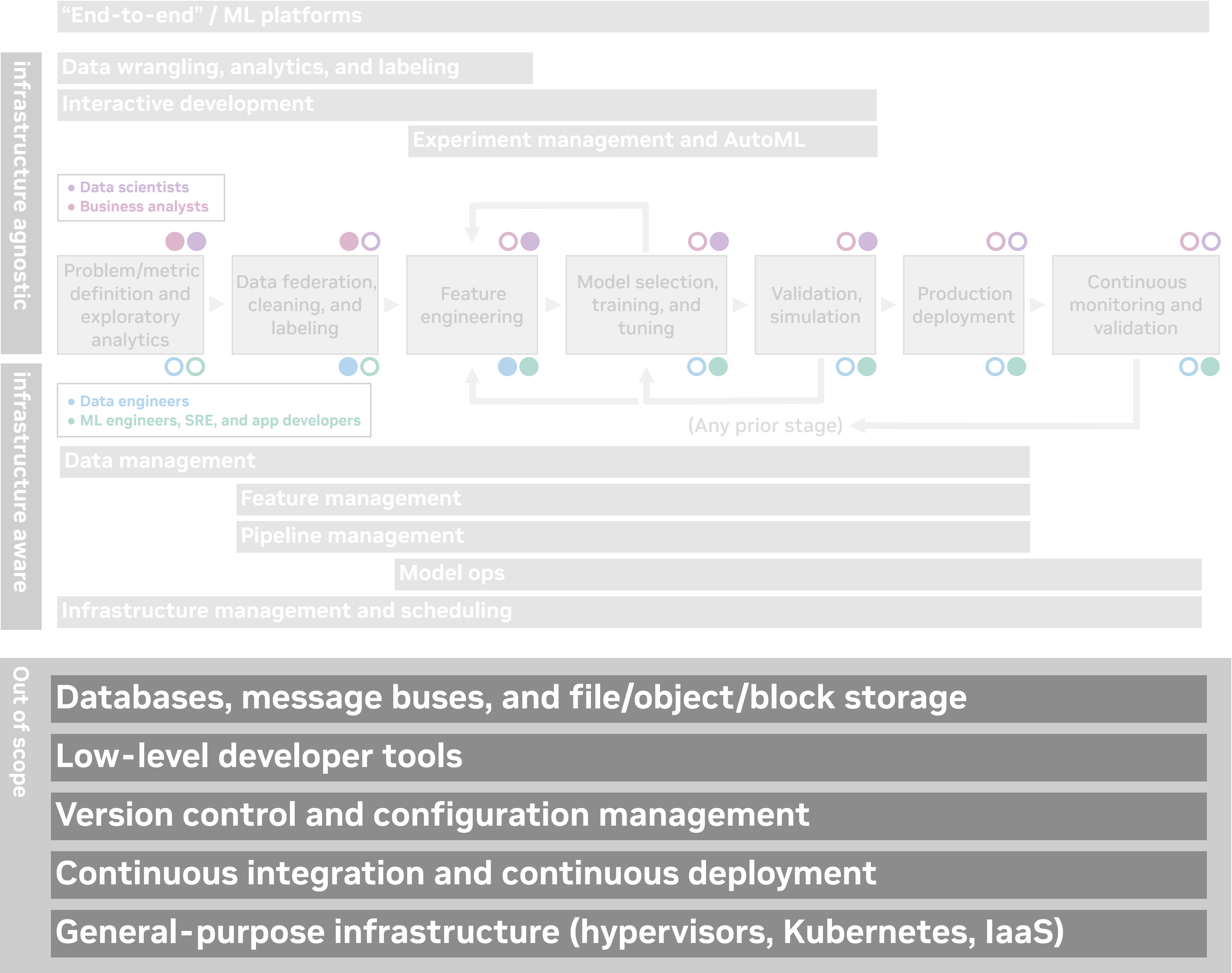
Data management

Feature management

Pipeline management

Model ops

Infrastructure management and scheduling





**Evaluating components, stacks,
and solutions for your use case**

“End-to-end” / ML platforms

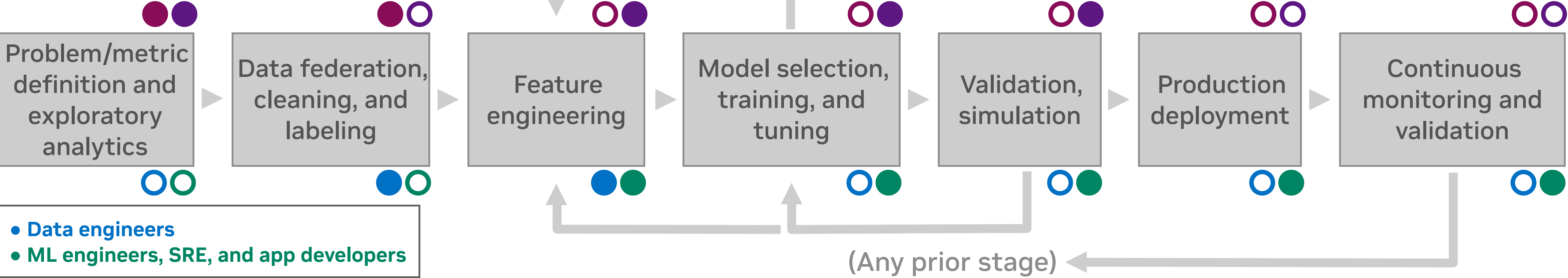
infrastructure agnostic

Data wrangling, analytics, and labeling

Interactive development

Experiment management and AutoML

- Data scientists
- Business analysts



- Data engineers
- ML engineers, SRE, and app developers

infrastructure aware

Data management

Feature management

Pipeline management

Model ops

Infrastructure management and scheduling

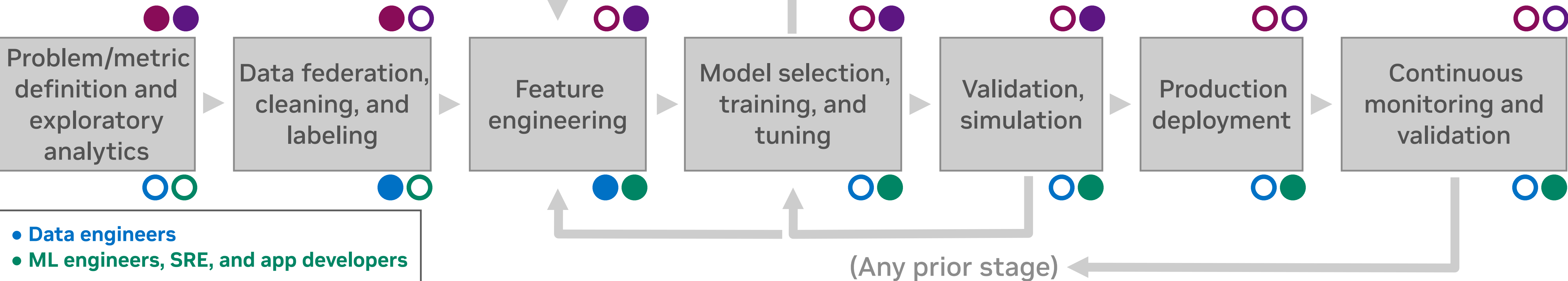
“End-to-end” / ML platforms

Data wrangling, analytics, and labeling

Interactive development

Experiment management and AutoML

- Data scientists
- Business analysts



Data management

Feature management

Pipeline management

Model ops

Infrastructure management and scheduling

infrastructure agnostic

infrastructure aware

infrastructure agnostic

infrastructure aware

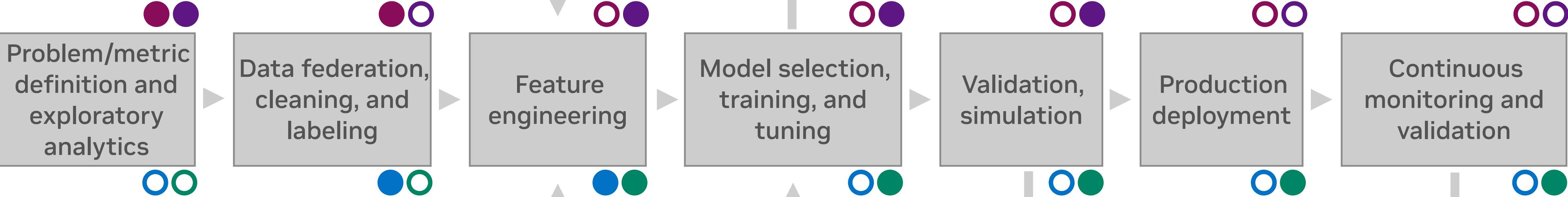
“End-to-end” / ML platforms

Data wrangling, analytics, and labeling

Interactive development

Experiment management and AutoML

- Data scientists
- Business analysts



- Data engineers
- ML engineers, SRE, and app developers

Data management

Feature management

Pipeline management

Model ops

Infrastructure management and scheduling

“End-to-end” / ML platforms

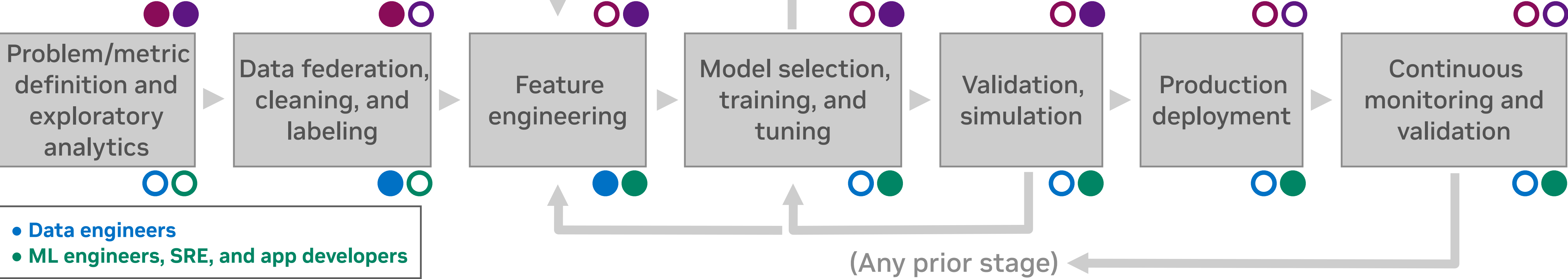
infrastructure agnostic

Data wrangling, analytics, and labeling

Interactive development

Experiment management and AutoML

- Data scientists
- Business analysts



- Data engineers
- ML engineers, SRE, and app developers

infrastructure aware

Data management

Feature management

Pipeline management

Model ops

Infrastructure management and scheduling



“End-to-end” / ML platforms

infrastructure agnostic

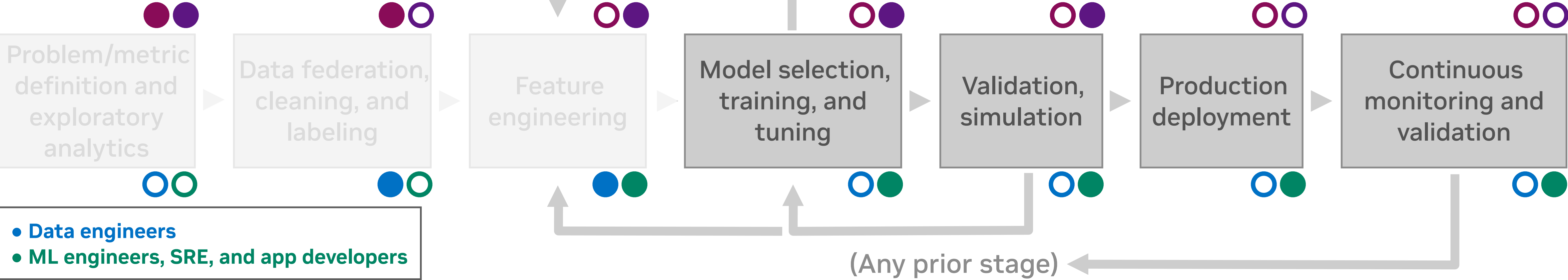
infrastructure aware

Data wrangling, analytics, and labeling

Interactive development

Experiment management and AutoML

- Data scientists
- Business analysts



Data management

Feature management

Pipeline management

Model ops

Infrastructure management and scheduling



“End-to-end” / ML platforms

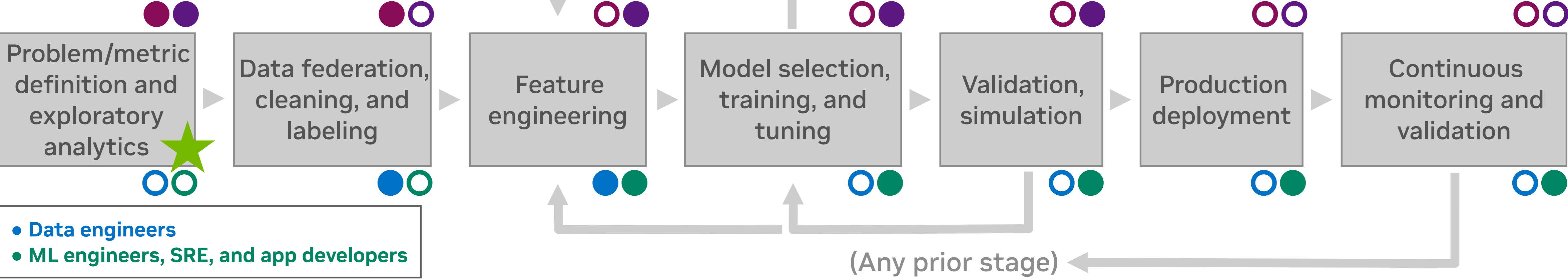
infrastructure agnostic

Data wrangling, analytics, and labeling

Interactive development

Experiment management and AutoML

- Data scientists
- Business analysts



infrastructure aware



“End-to-end” / ML platforms

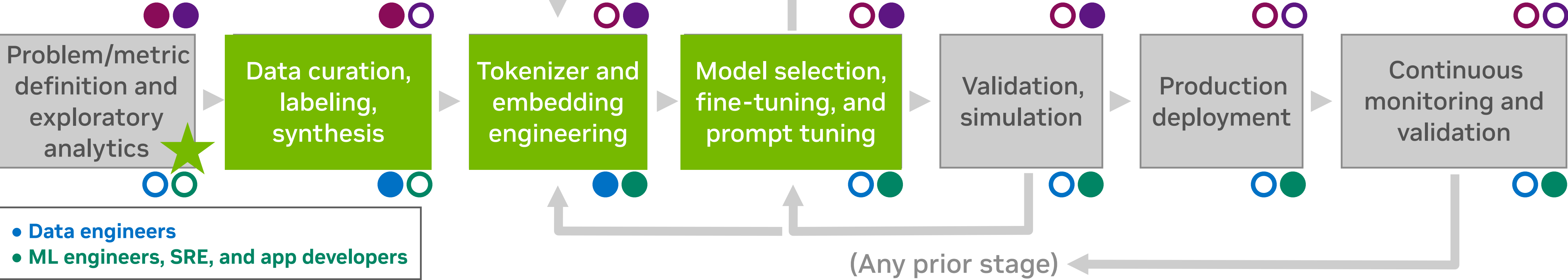
infrastructure agnostic

Data wrangling, analytics, and labeling

Interactive development

Experiment management and AutoML

- Data scientists
- Business analysts



infrastructure aware

Data management

Feature management

Pipeline management

Model ops

Infrastructure management and scheduling



“End-to-end” / ML platforms and AI application platforms

infrastructure agnostic

Data wrangling, analytics, and labeling

Interactive development

Experiment and prompt management

- Data scientists
- Business analysts

Problem/metric definition and exploratory analytics

●●

○●

Data curation, labeling, synthesis

●○

●○

Tokenizer and embedding engineering

○●

●●

Model selection, fine-tuning, and prompt tuning

○●

○●

Validation, simulation

○●

○●

Production deployment

○●

○●

Continuous monitoring and validation

○●

○●

- Data engineers
- ML engineers, SRE, and app developers

(Any prior stage)

infrastructure aware

Data management

Vector databases and search engines

Pipeline, agent, and chain management

Model ops

Guardrails

Infrastructure management and scheduling

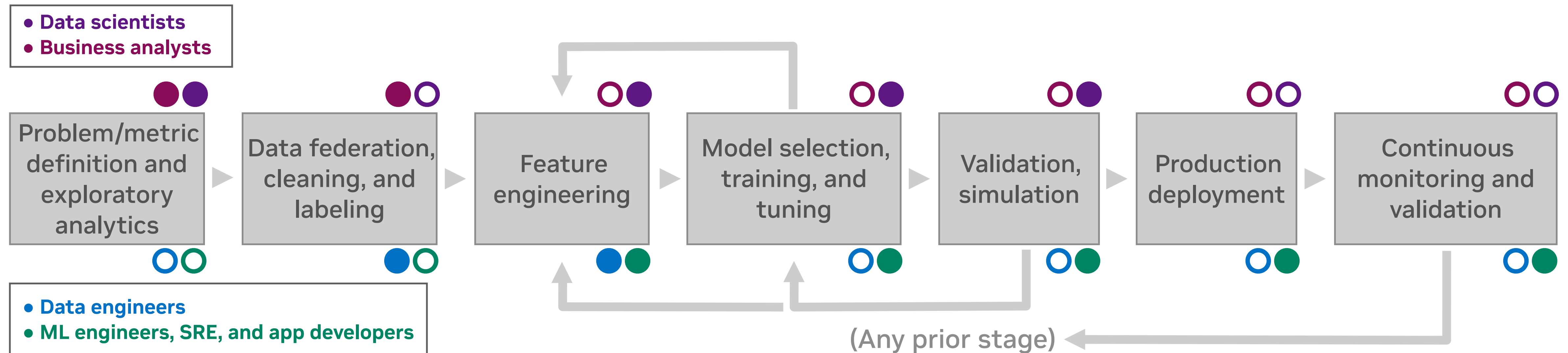


“End-to-end” / ML platforms

Consider CSP native offerings: Sagemaker, Vertex AI, AzureML; consider also hybrid-cloud offerings: Domino Data Lab, Dataiku, and OpenShift AI.

Many products in other categories are expanding in scope!

Kubeflow is a great starting point for an open-source ML platform.

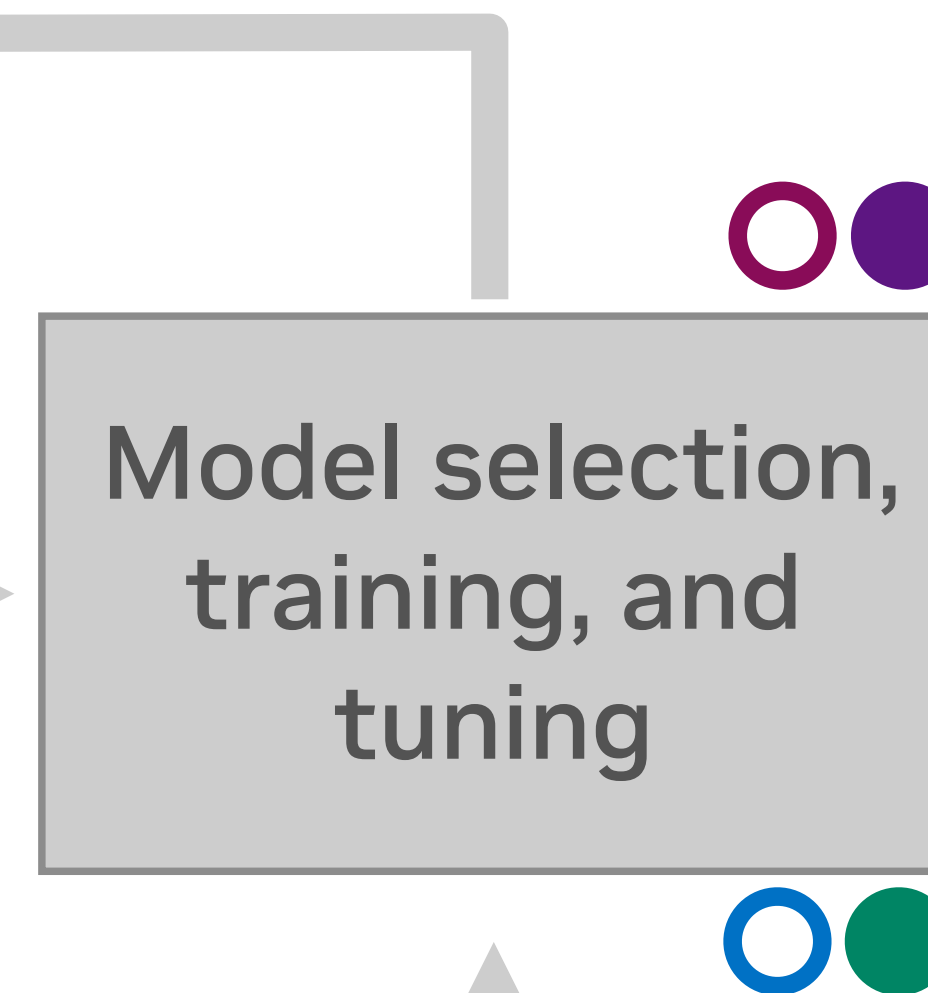
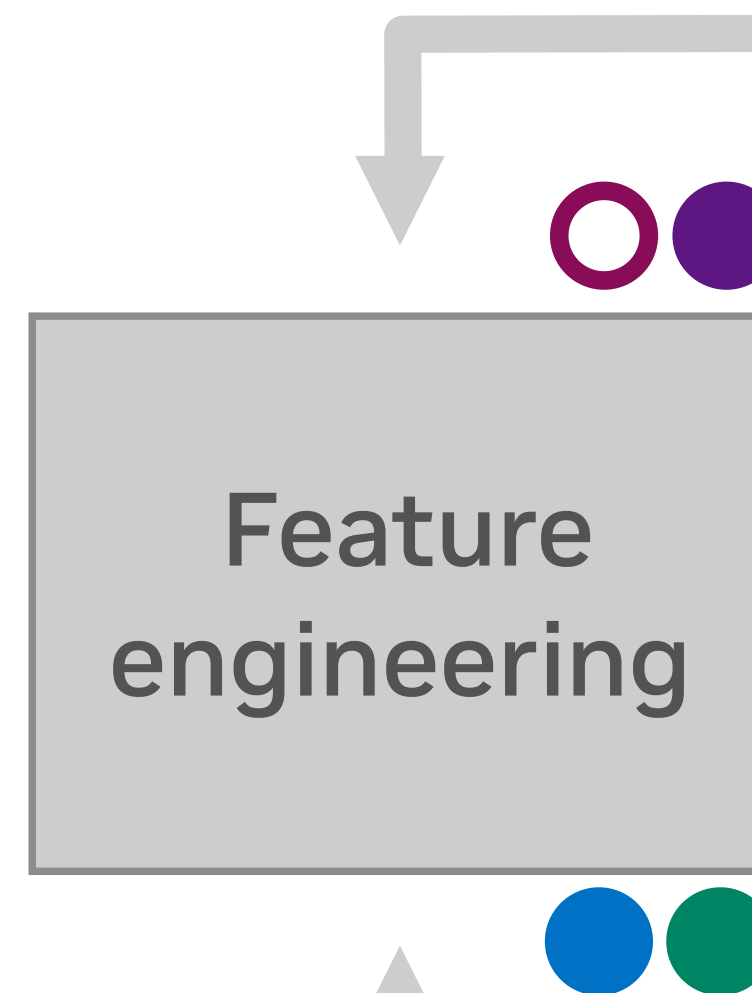


Interactive development

Jupyter or VSCode is a great place to start. GitHub Codespaces offers hosted development environments with version control.

NVIDIA AI Workbench makes it easy to package, share, and reproduce your ML experiments and techniques.

- Data scientists
- Business analysts



- Data engineers
- ML engineers, SRE, and app developers

(Any prior stage) ←

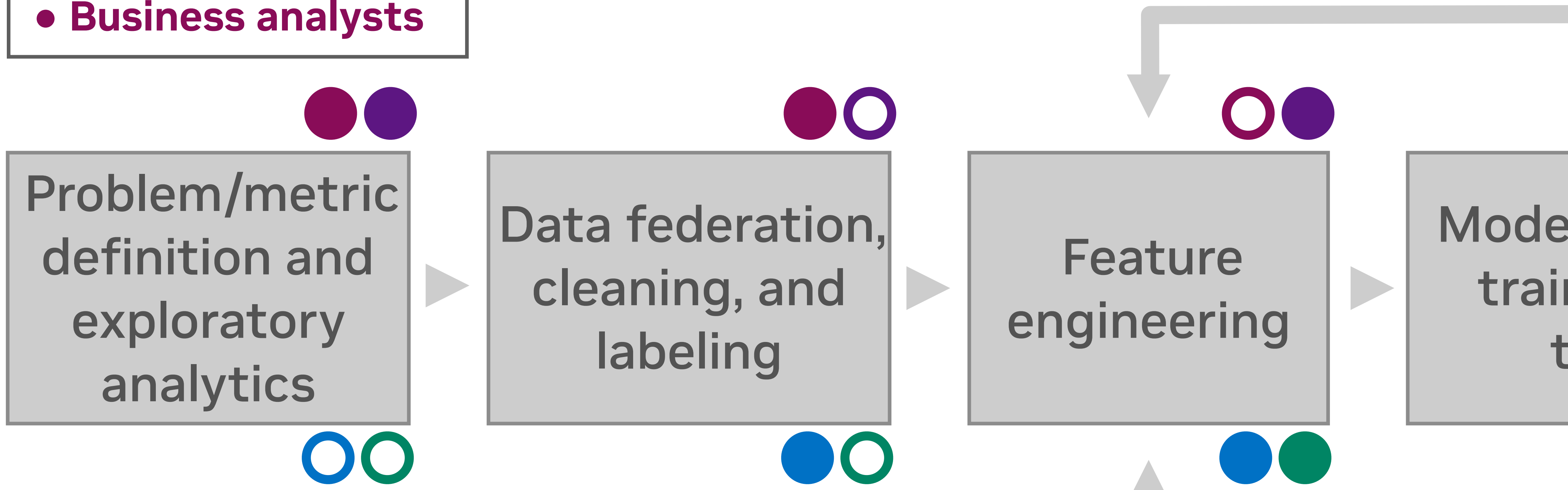
Data wrangling, analytics, and labeling

Consider platforms like Datarobot and Dataiku; consider also SQReam, Heavy.ai, Dataloop, Kinetica, and Alteryx.

Labeling leaders include LabelBox, LabelStudio, Scale, and Appen.

The Python data ecosystem (including RAPIDS) is useful here!

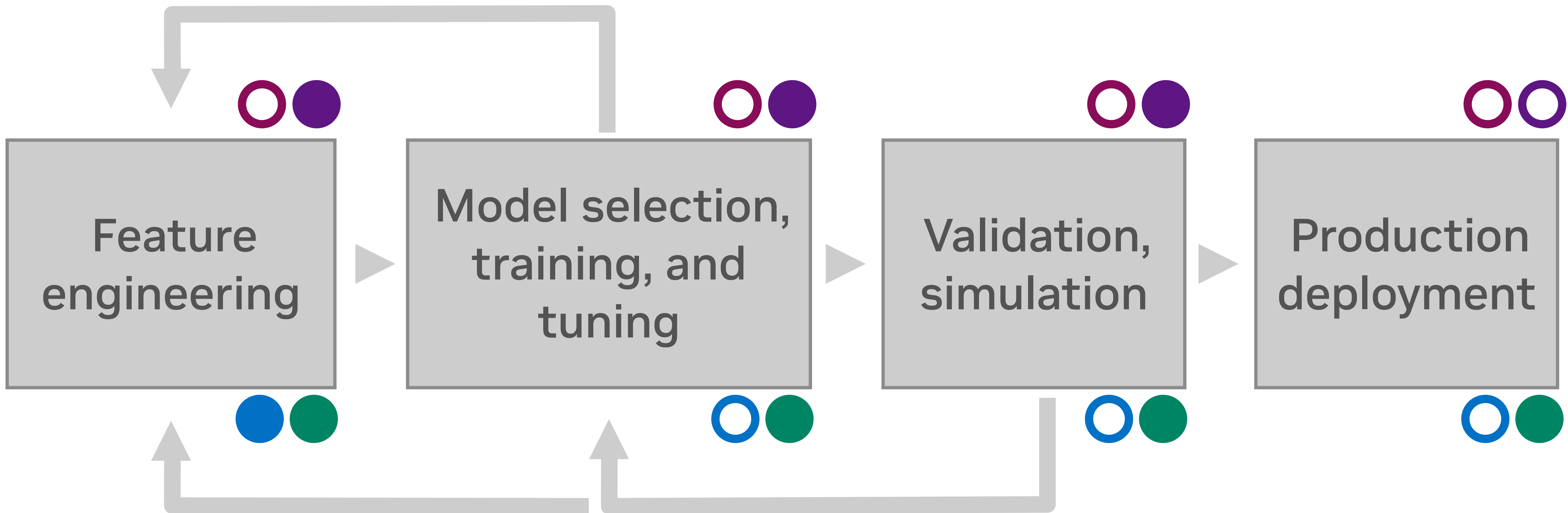
- Data scientists
- Business analysts

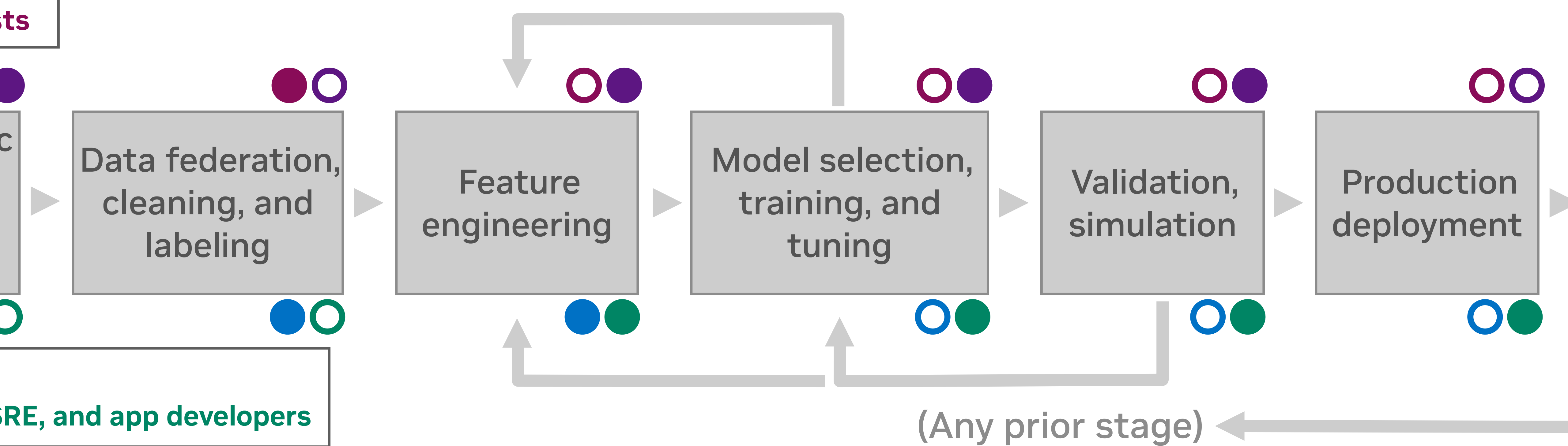


Experiment management

Consider Weights and Biases and Comet ML.

MLFlow and Katib are relevant open-source projects.



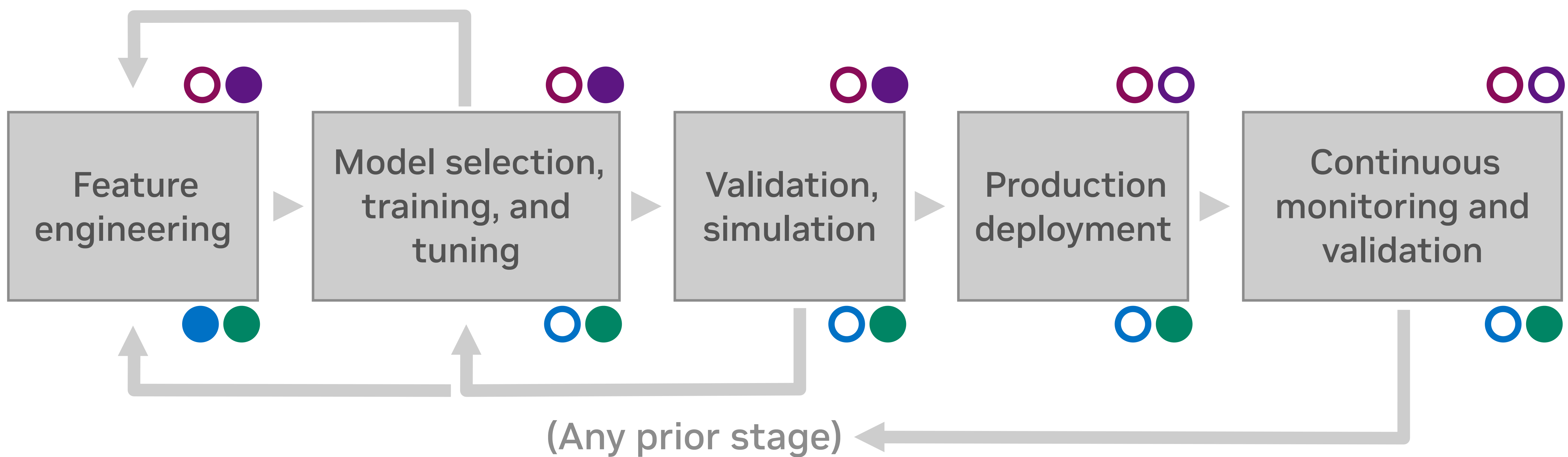


Feature management

Consider Hopsworks and Tecton.

LLM applications will often benefit from vector databases like Milvus.

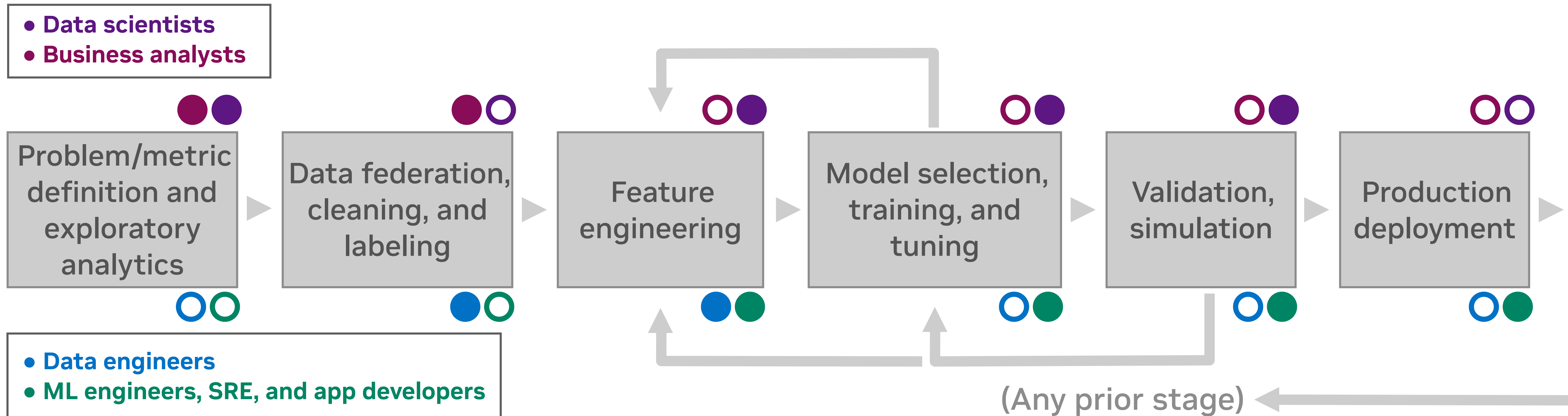
Feast is the most notable open-source solution. Fast datastores like Redis often figure in production feature-serving use cases.



Model ops

For model serving, consider NVIDIA solutions: Triton, TRT, Seldon. For monitoring, consider Fiddler, Arize, and BentoML.

Weights and Biases and MLFlow both offer serving and model monitoring.

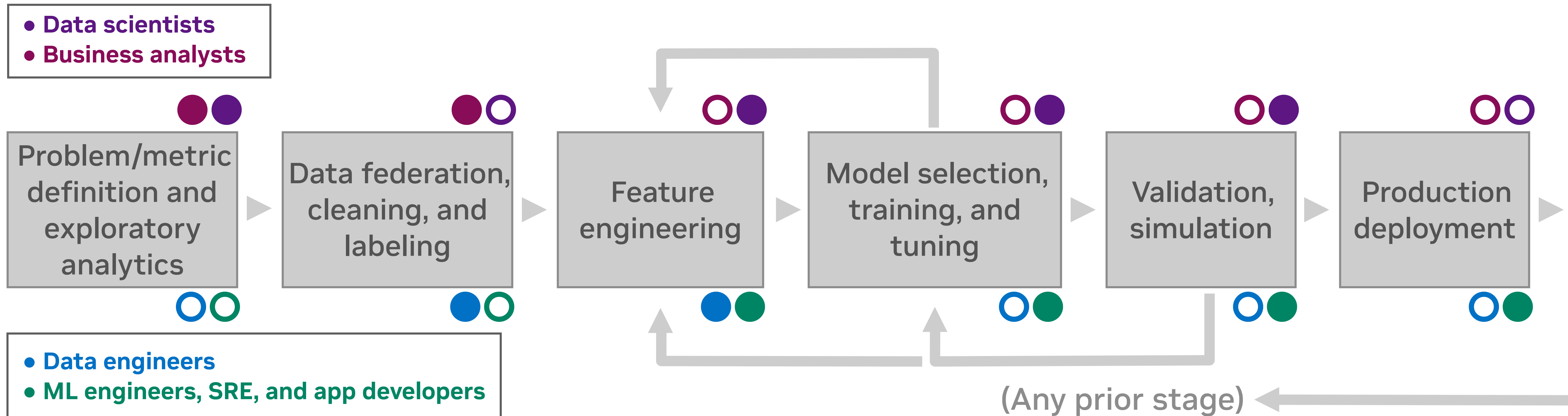


Data management

Consider HPE MLDM, Snowflake, Databricks, and DataLoop.

Many of NVIDIA's storage partners have mature offerings in this space.

This space marks the interface between data at rest and production AI/ML systems; the boundaries are necessarily a bit fuzzy.

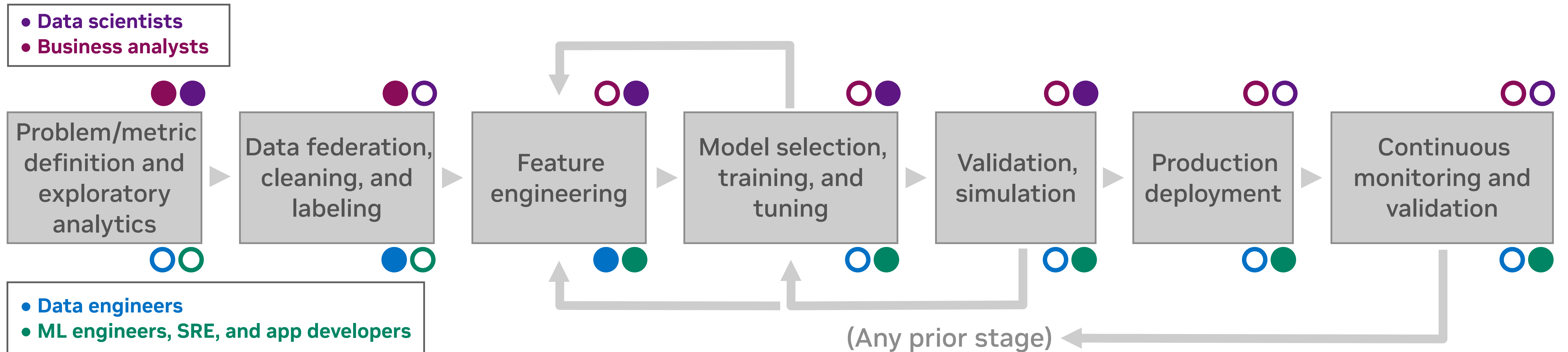


Pipeline management

Apache Airflow and Flyte are important open-source projects with commercial offerings from Astronomer and Union.ai, respectively.

Consider also Orkes and Temporal.

Other relevant projects include Luigi, Prefect, Argo, and Tekton.



Infrastructure management and scheduling

Base Command Manager (BCM) can provision & deploy a cluster. Run.ai has a really compelling offering for batch scheduling on K8s.

Platform partners cover many important integration points with enterprise IT systems *and* open-source ecosystems.

Open-source offerings include Volcano, Armada, Covalent, and SkyPilot.

“End-to-end” / ML platforms

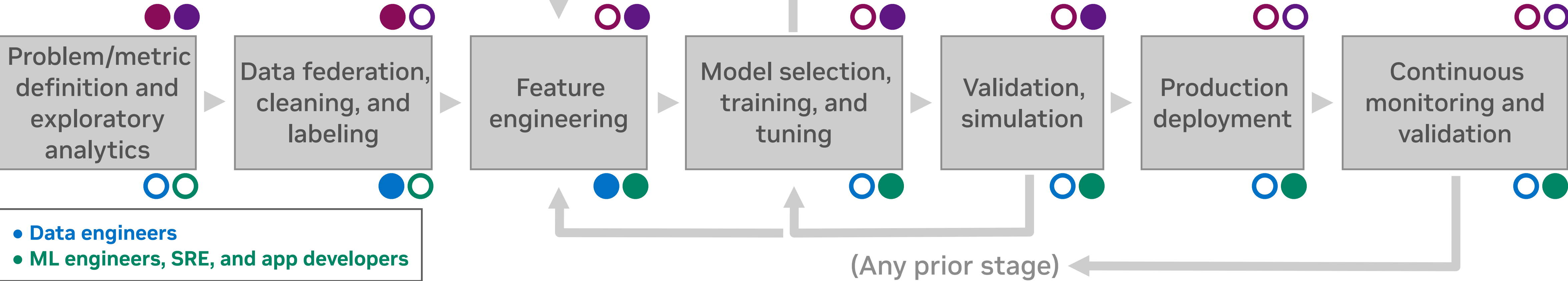
infrastructure agnostic

Data wrangling, analytics, and labeling

Interactive development

Experiment management and AutoML

- Data scientists
- Business analysts



- Data engineers
- ML engineers, SRE, and app developers

infrastructure aware

Data management

Feature management

Pipeline management

Model ops


Infrastructure management and scheduling



Conclusions and next steps

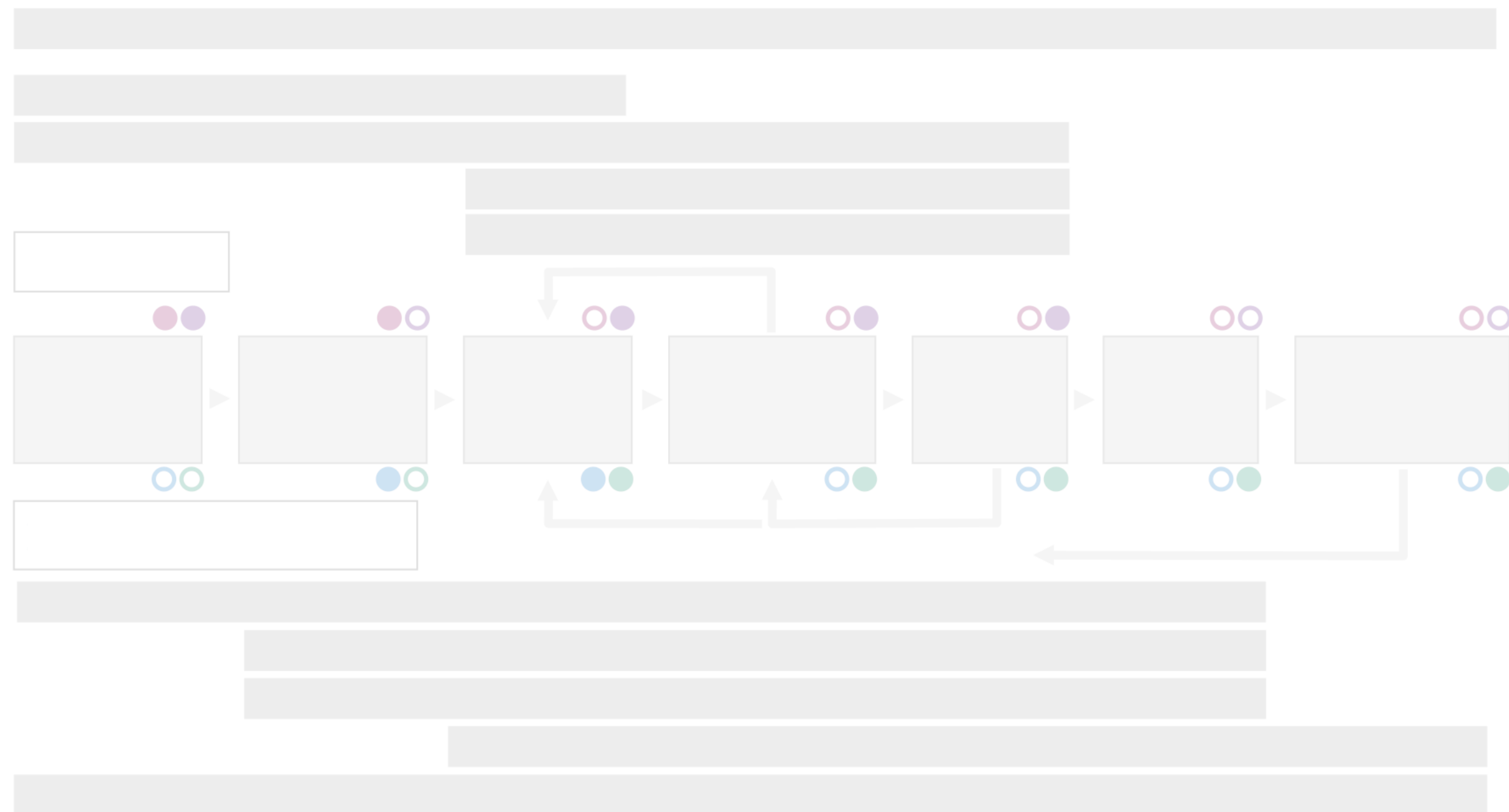
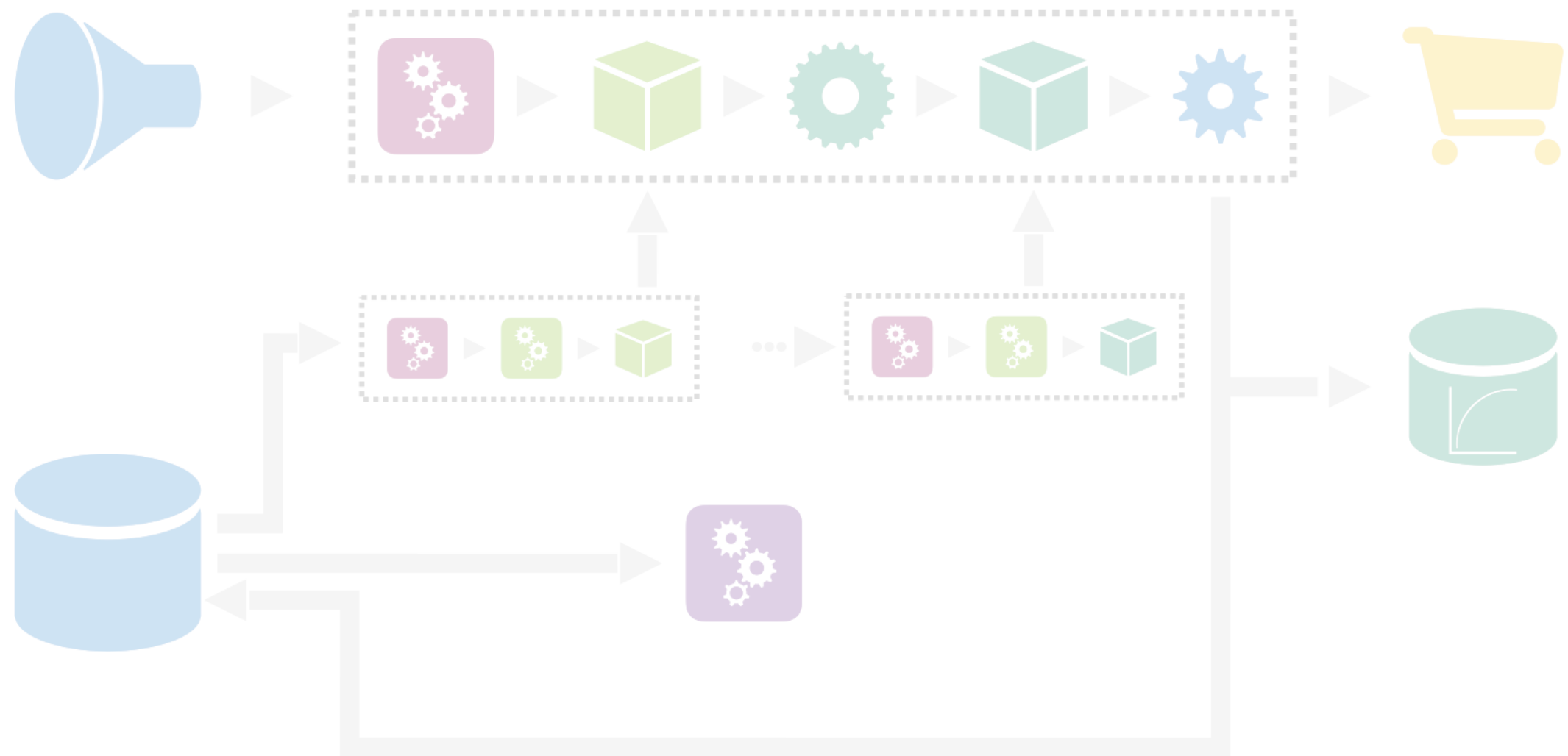
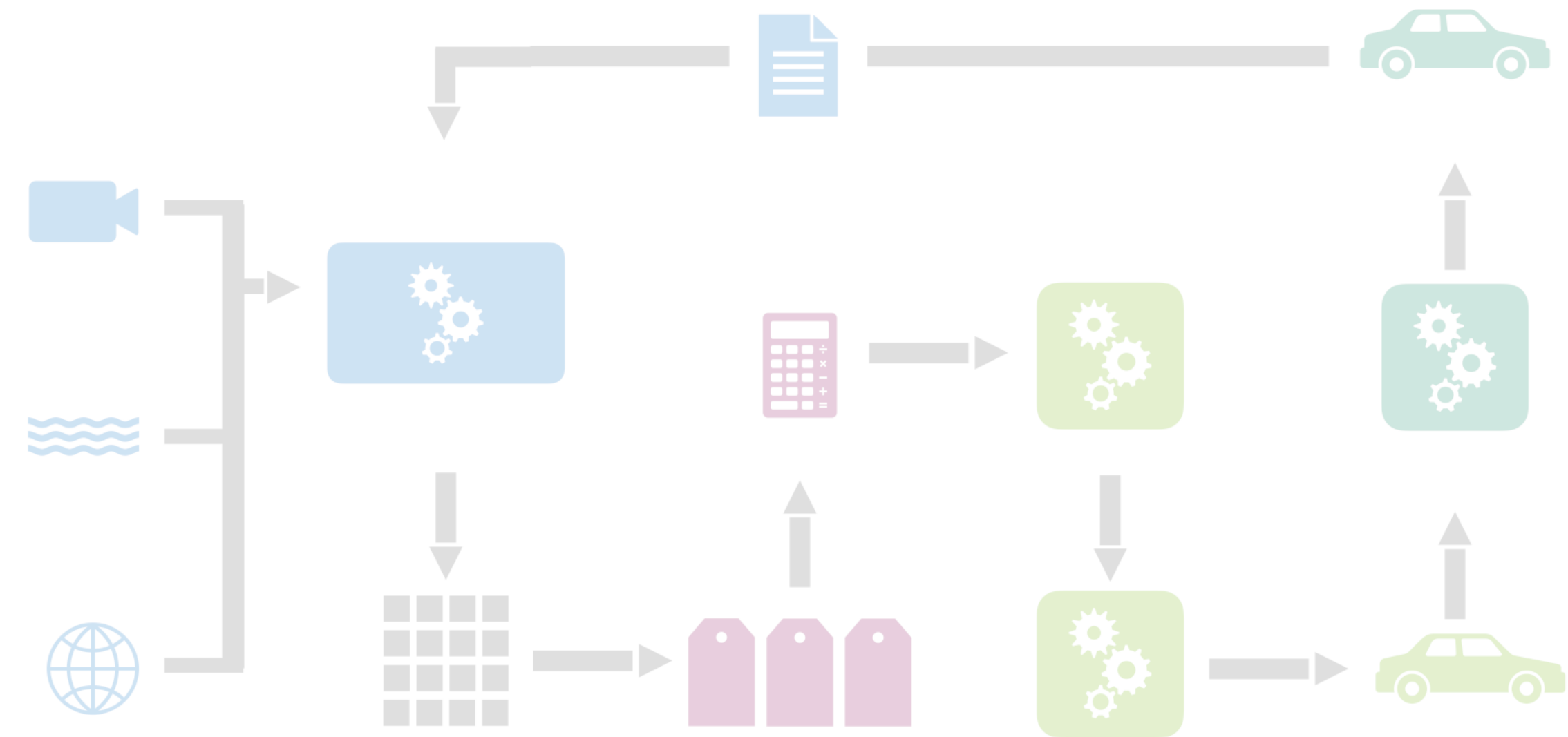
Check out these other great sessions

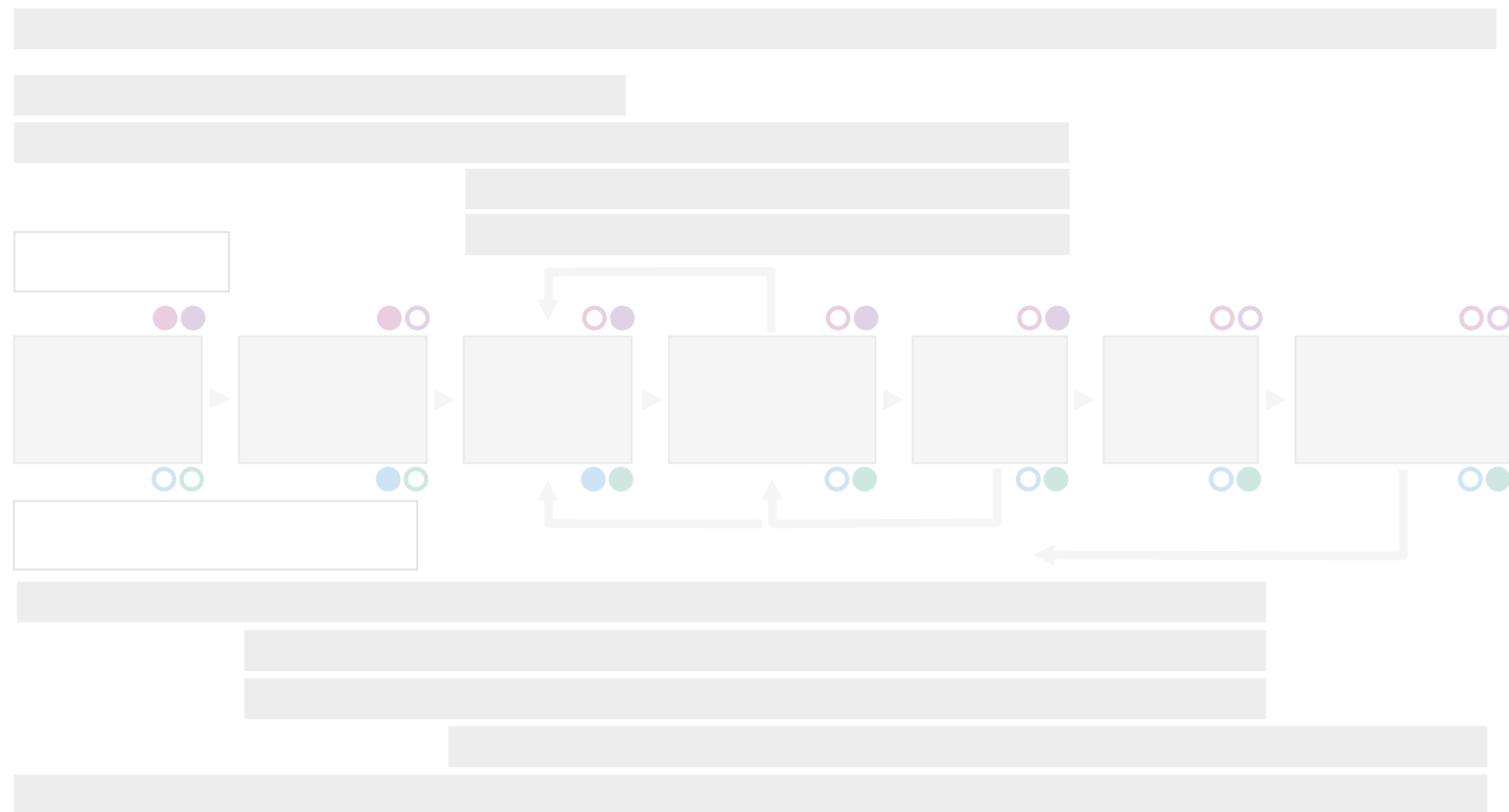
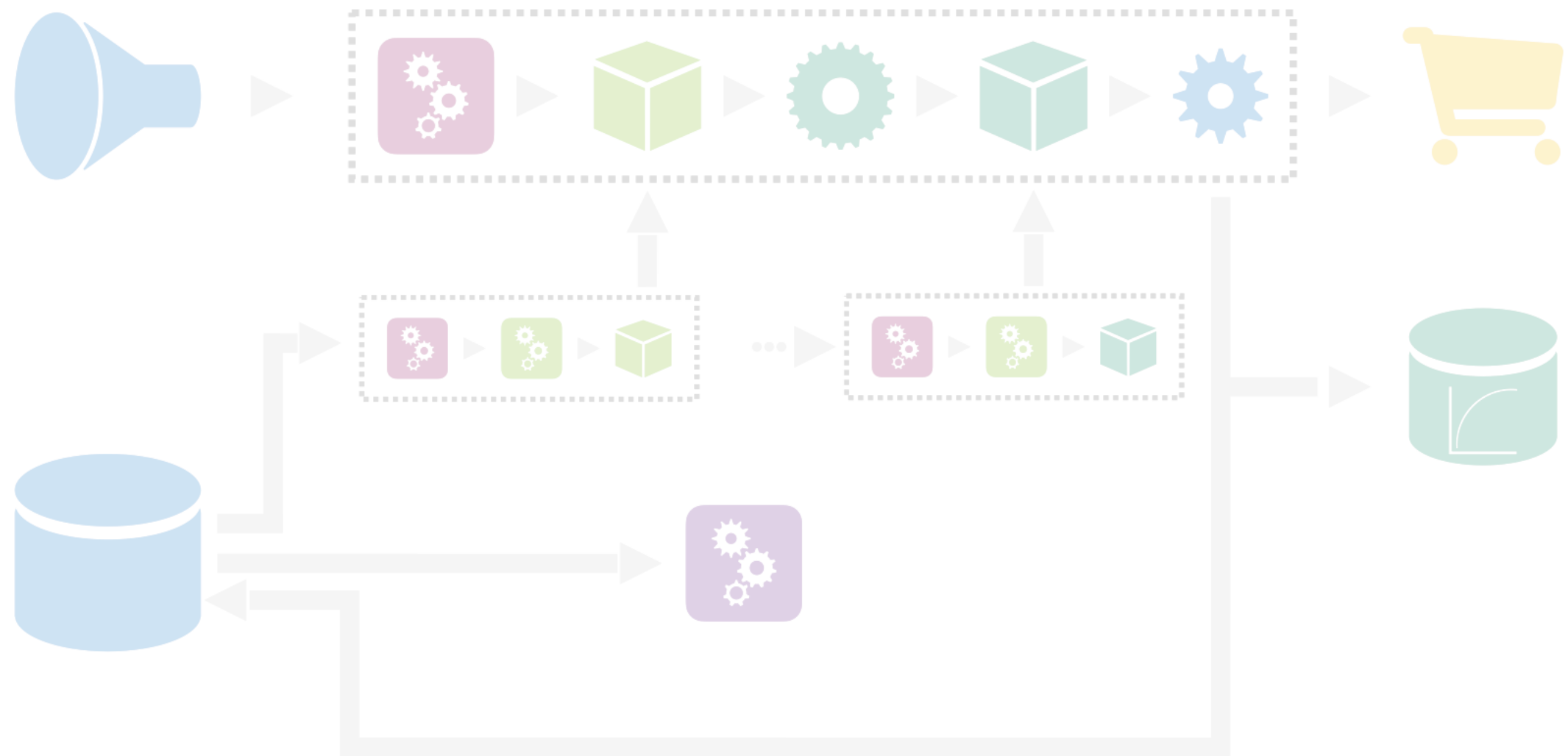
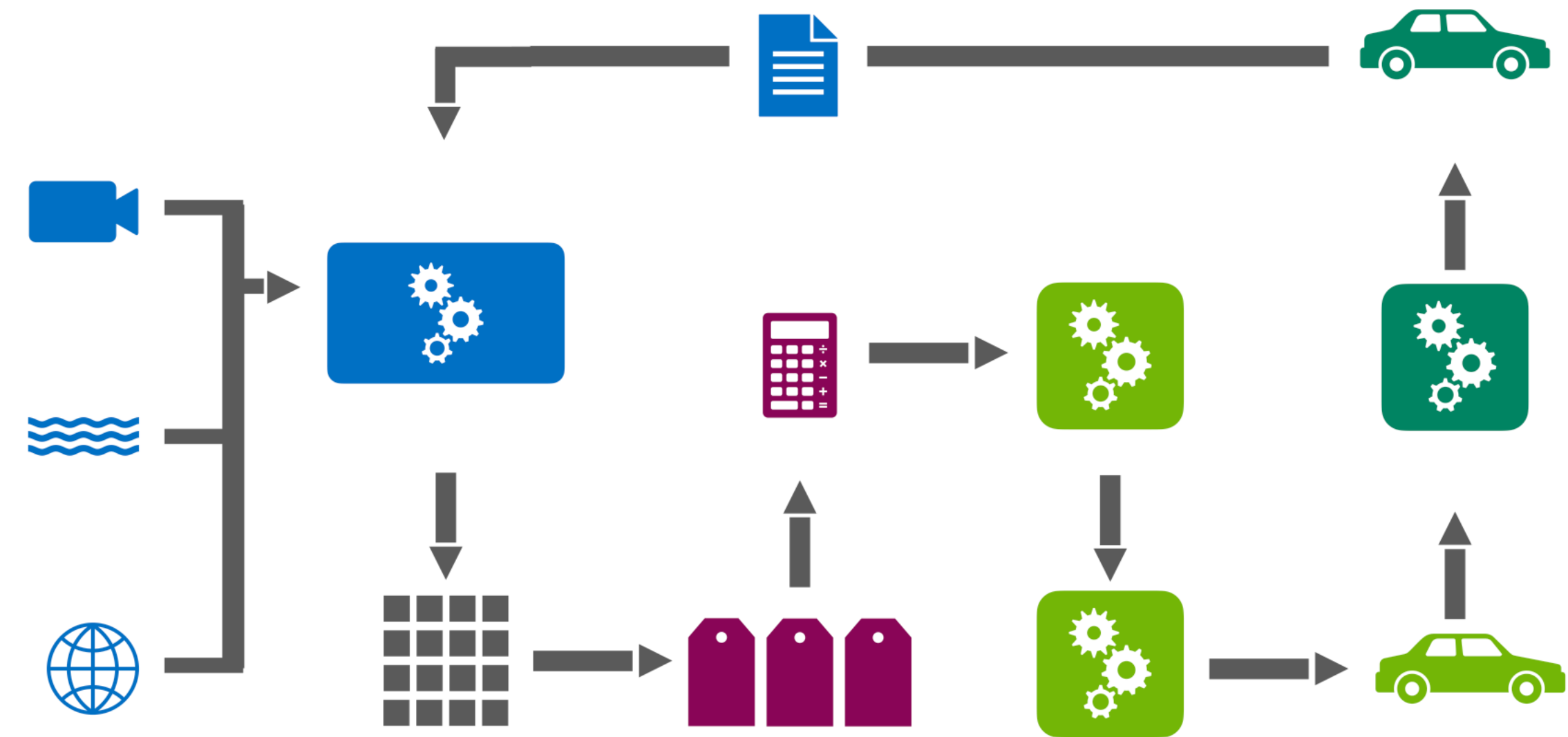
[S62427] Confidential Computing: New Features and NVIDIA Hardware Attestation (Michael O'Connor et al.)

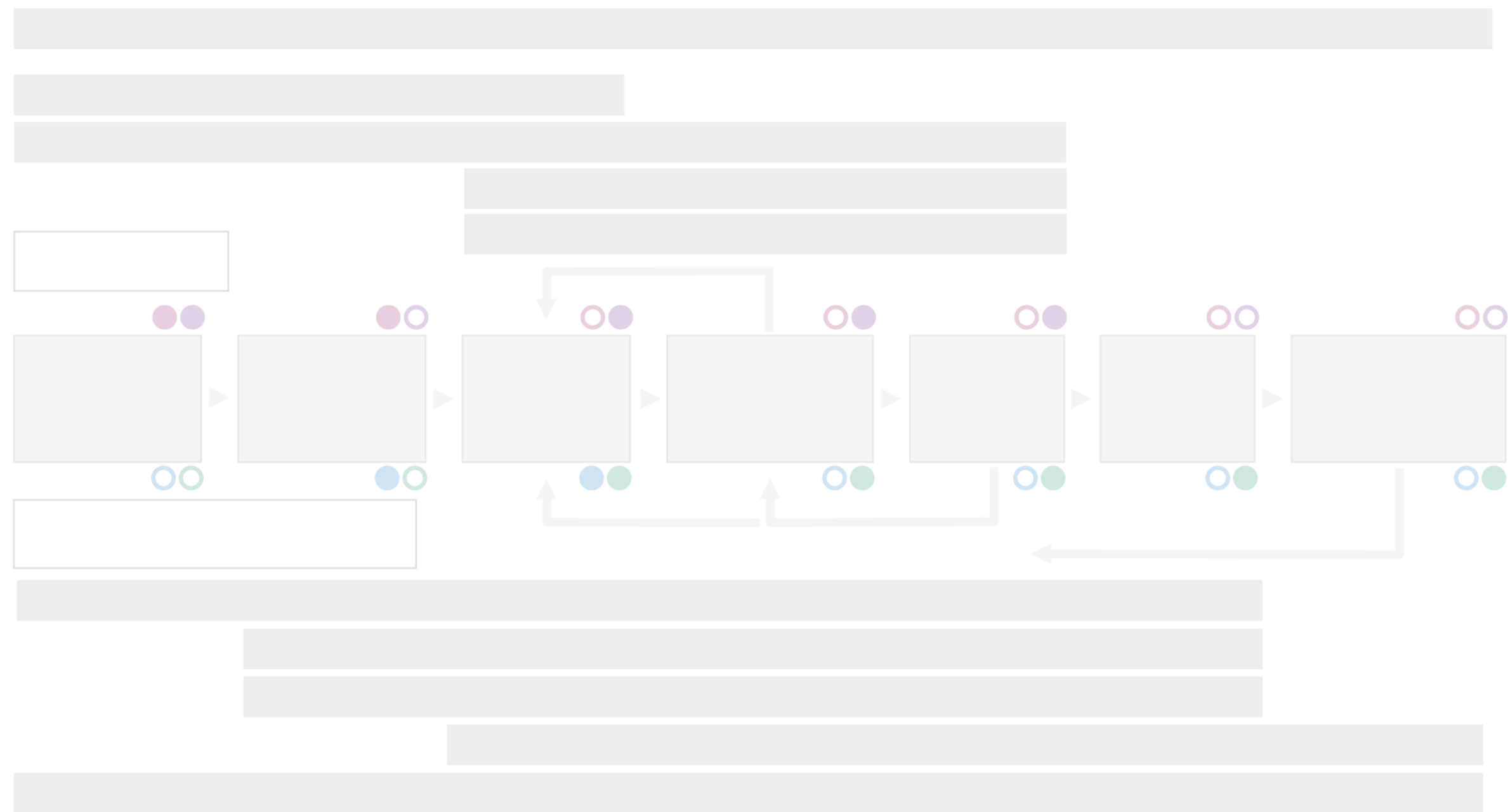
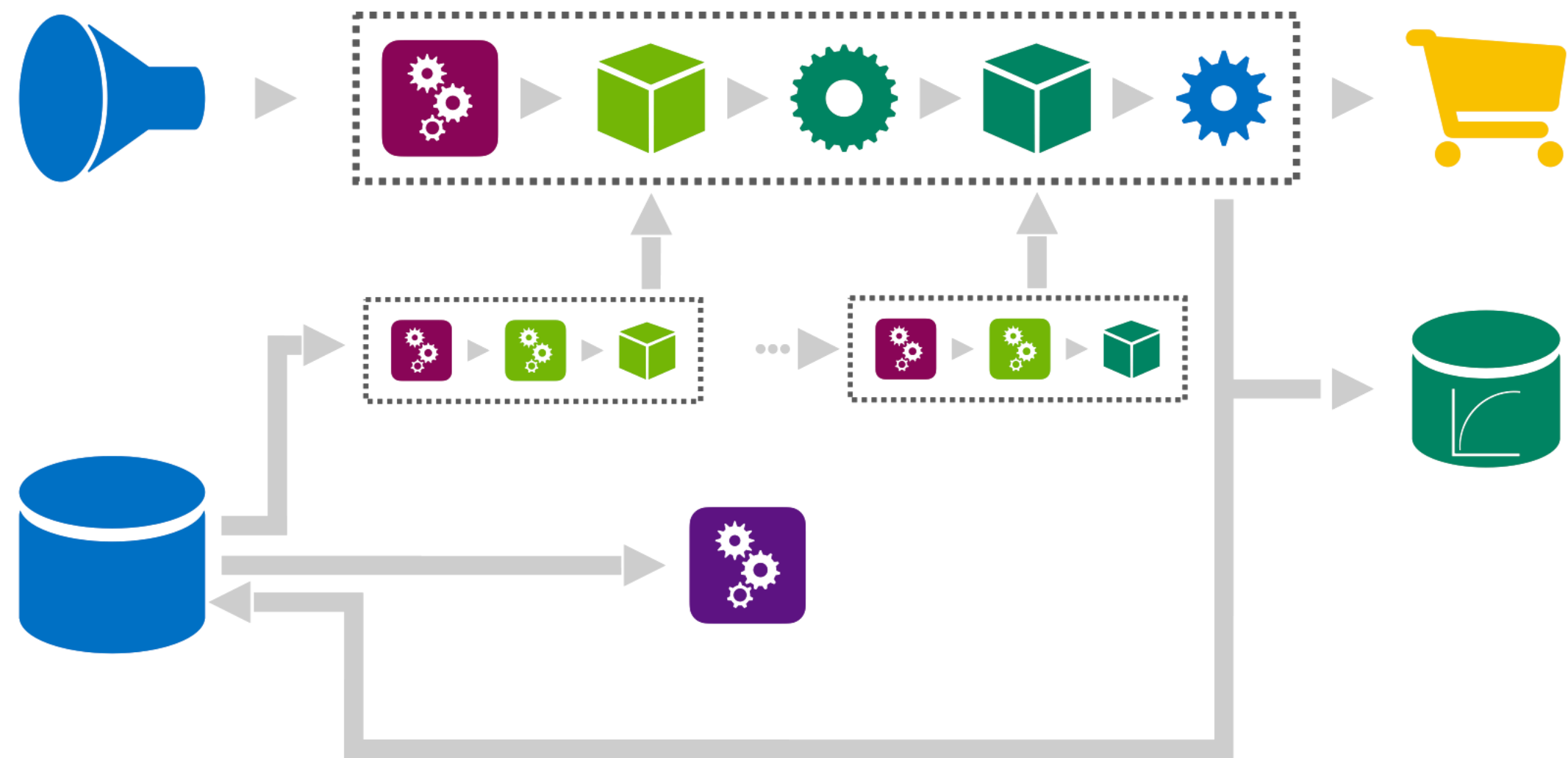
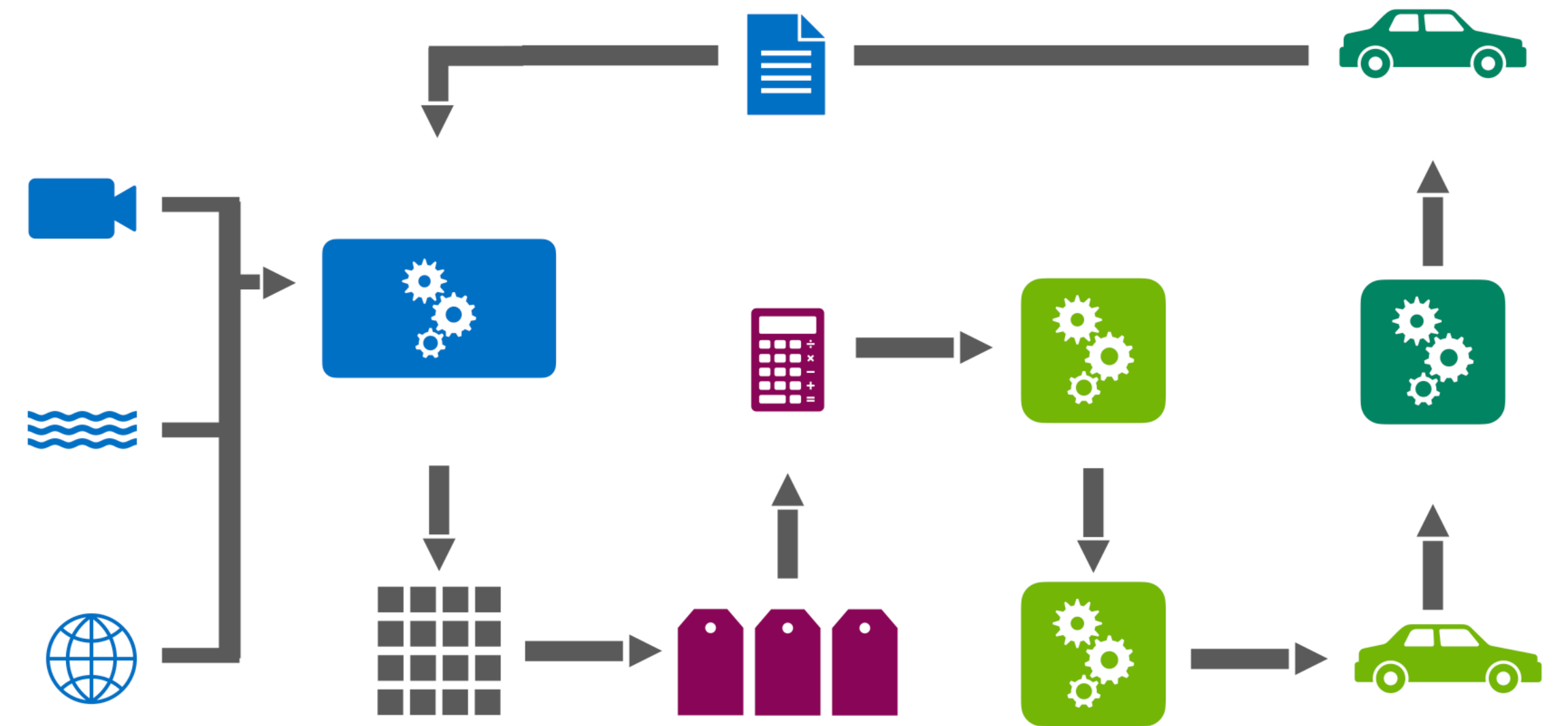
 [S62149] Decentralized Collaborative AI With Federated Learning in Trustworthy Environments (Emily Sakata)

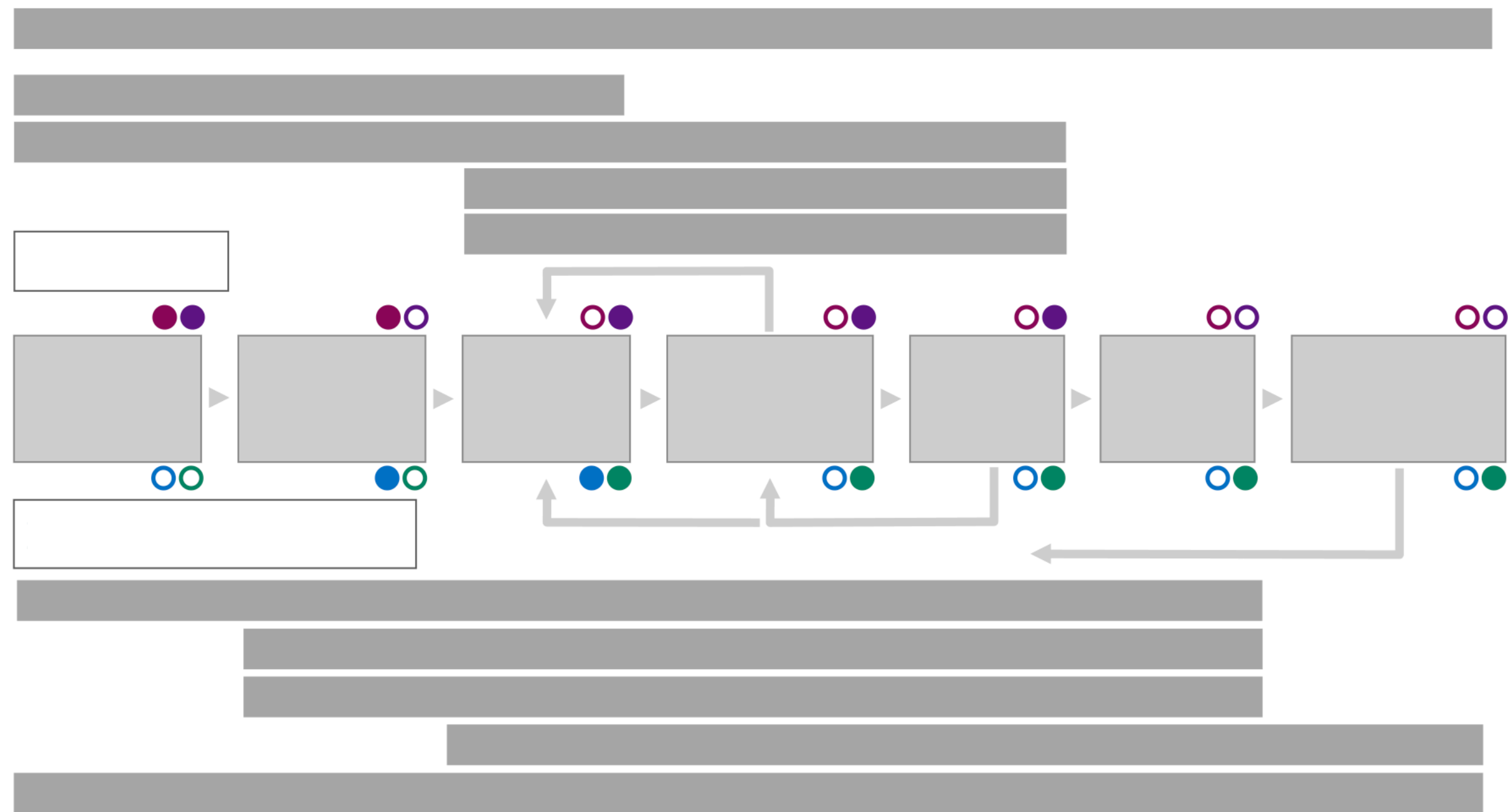
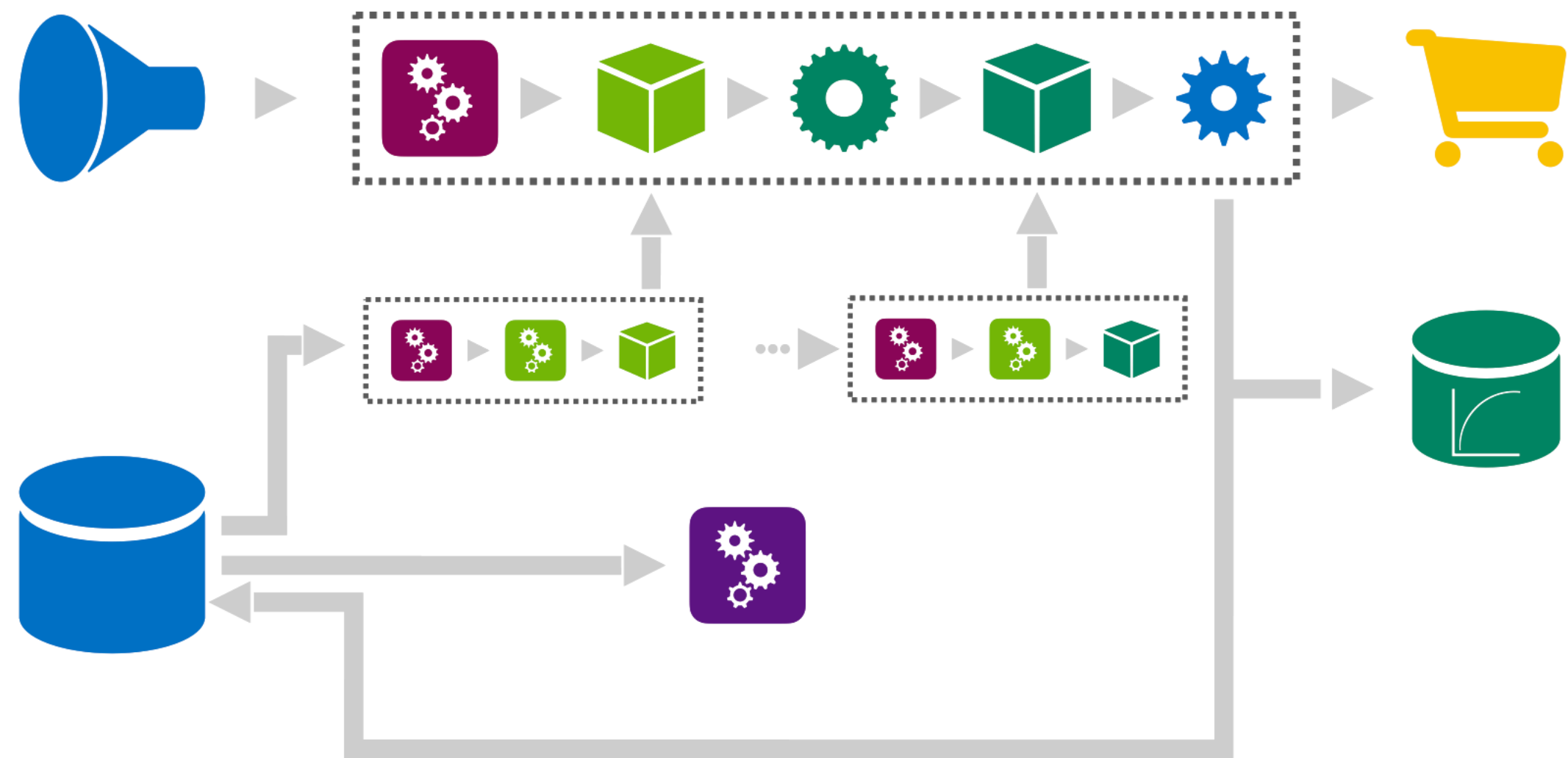
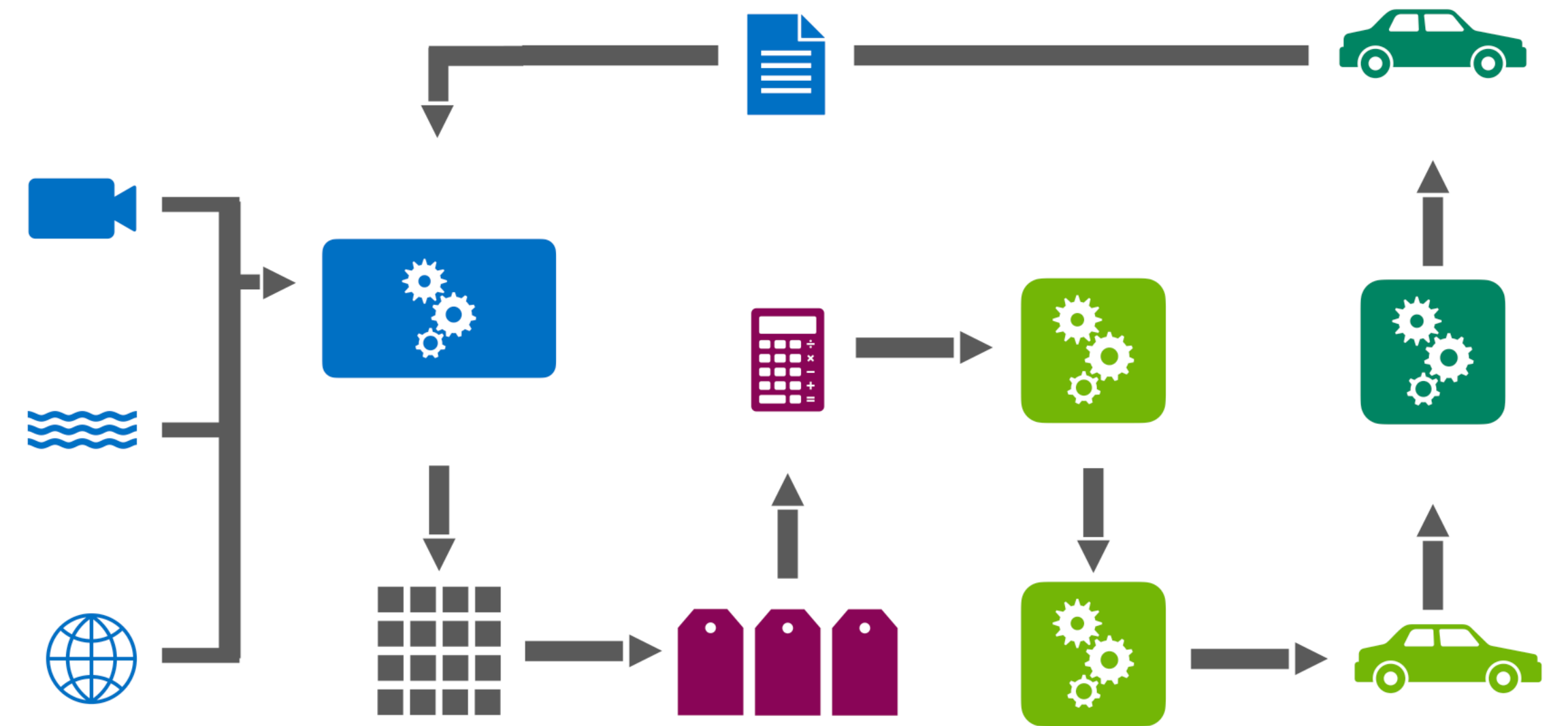
 [S62960] XGBoost is All You Need (Bojan Tunguz)

[S62458] LLMOps: The New Frontier of Machine Learning Operations (Nik Spirin and Michael Balint)



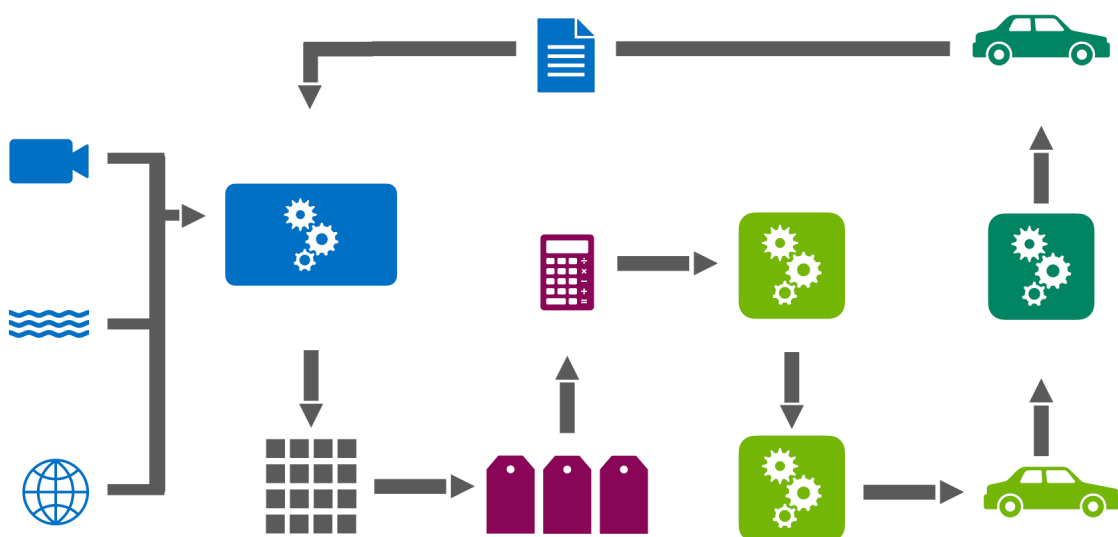







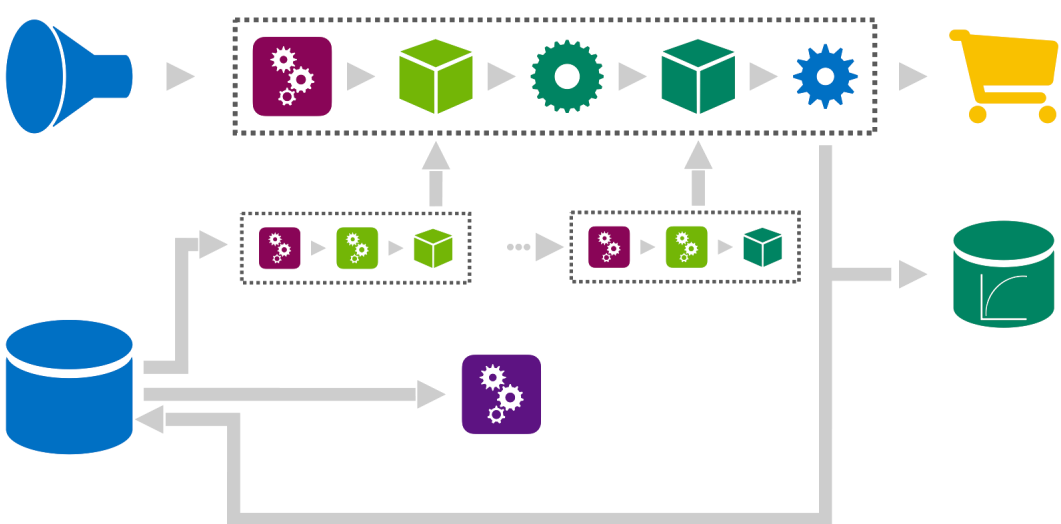


See these sessions next



[S62427] Confidential Computing: New Features and NVIDIA Hardware Attestation (Michael O'Connor et al.)

 [S62149] Decentralized Collaborative AI With Federated Learning in Trustworthy Environments (Emily Sakata)



 [S62960] XGBoost is All You Need (Bojan Tunguz)



[S62458] LLMOps: The New Frontier of Machine Learning Operations (Nik Spirin and Michael Balint)