

Databricks Demo: Transfer Learning with MLflow, the #1 Tool to Learn Today

Feifei Wang

Senior Data Scientist @ Databricks

feifei.wang@databricks.com

Feifei Wang

Ph.D in Applied Math & Computer Science @



Former Senior Decision Scientist @



Current Senior Data Scientist @



Feifei Wang

Hobbies: Ballet and Piano



Pet: Simba



Databricks

Unified Data Analytics Platform



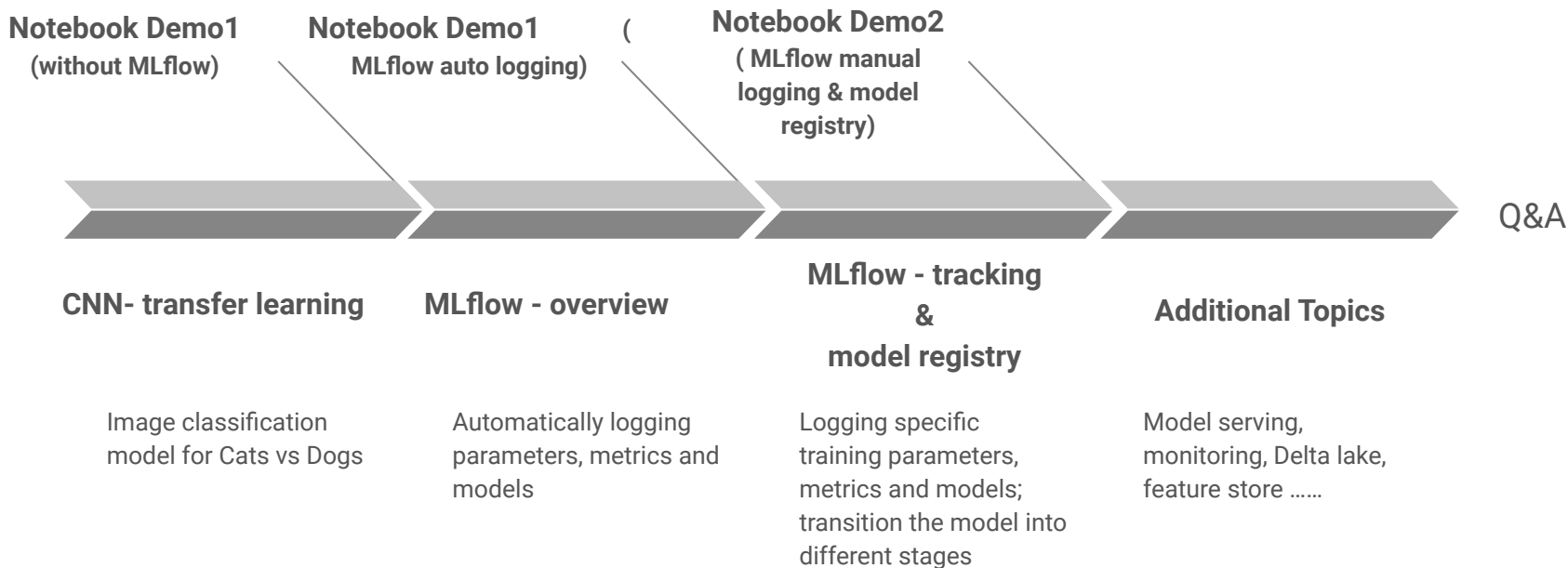
**Collaborative
Workspace**

**Data Quality +
Consistency**

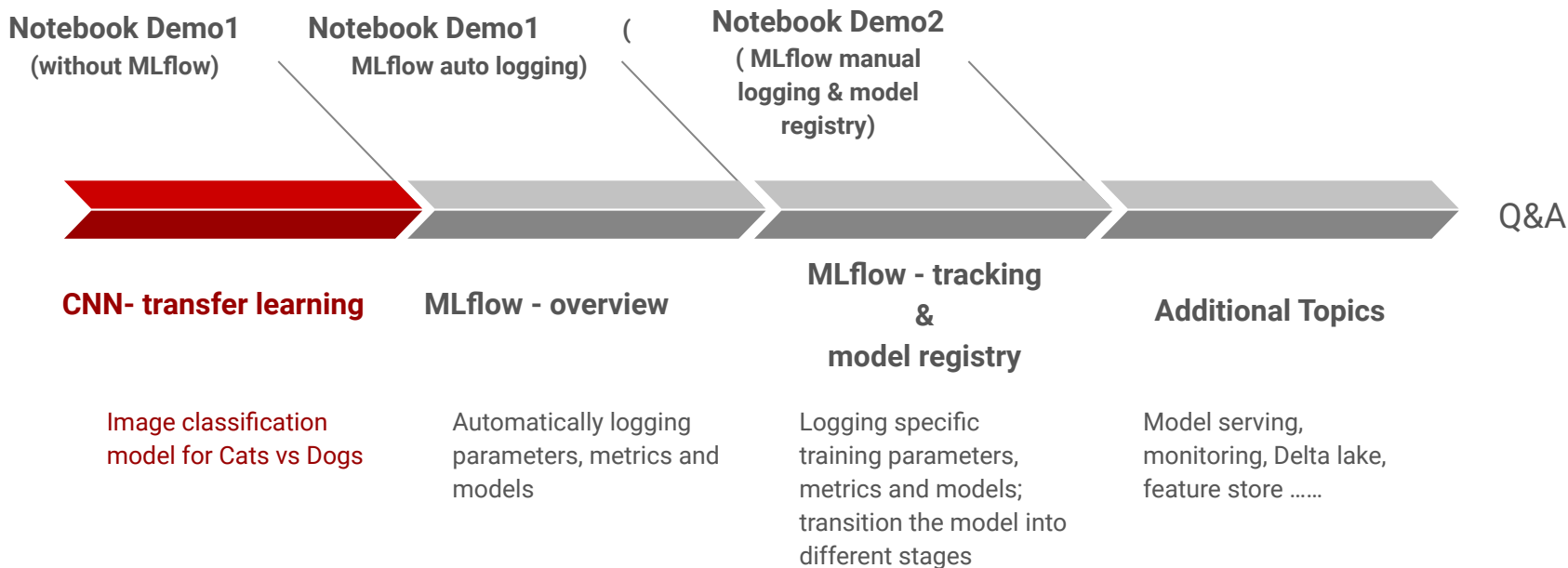
**Reproduce +
Productionize
Models**



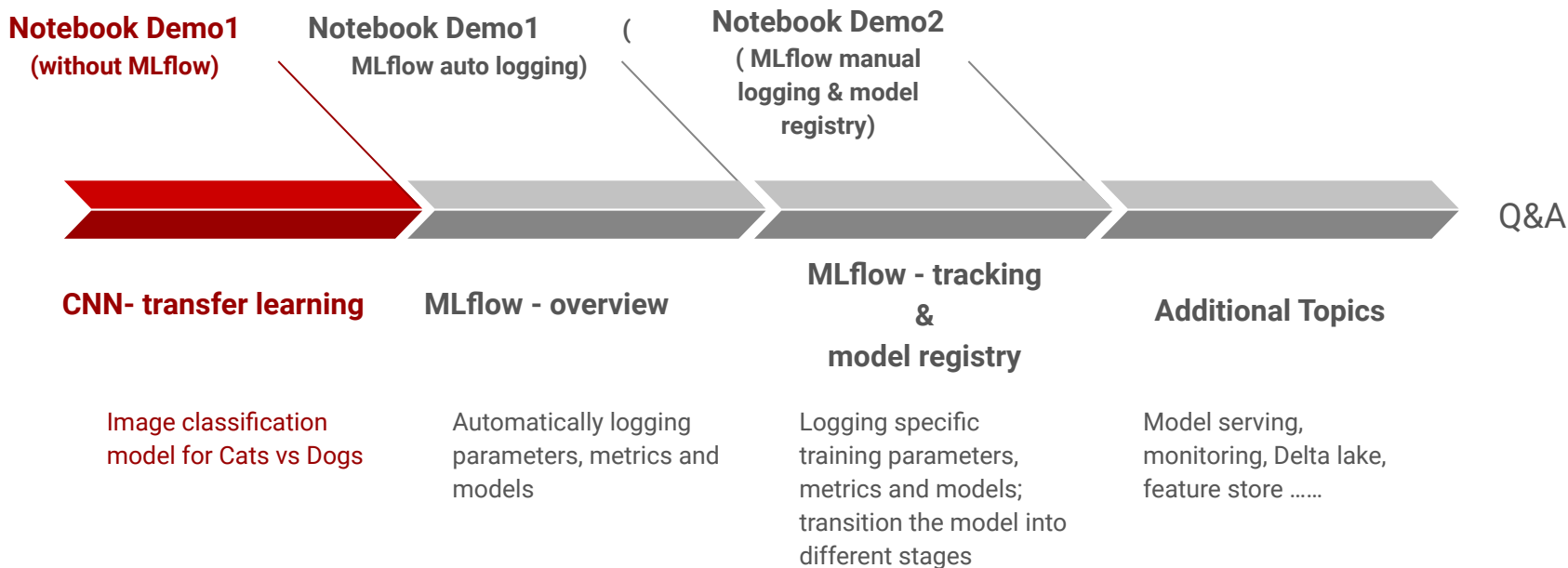
Outline



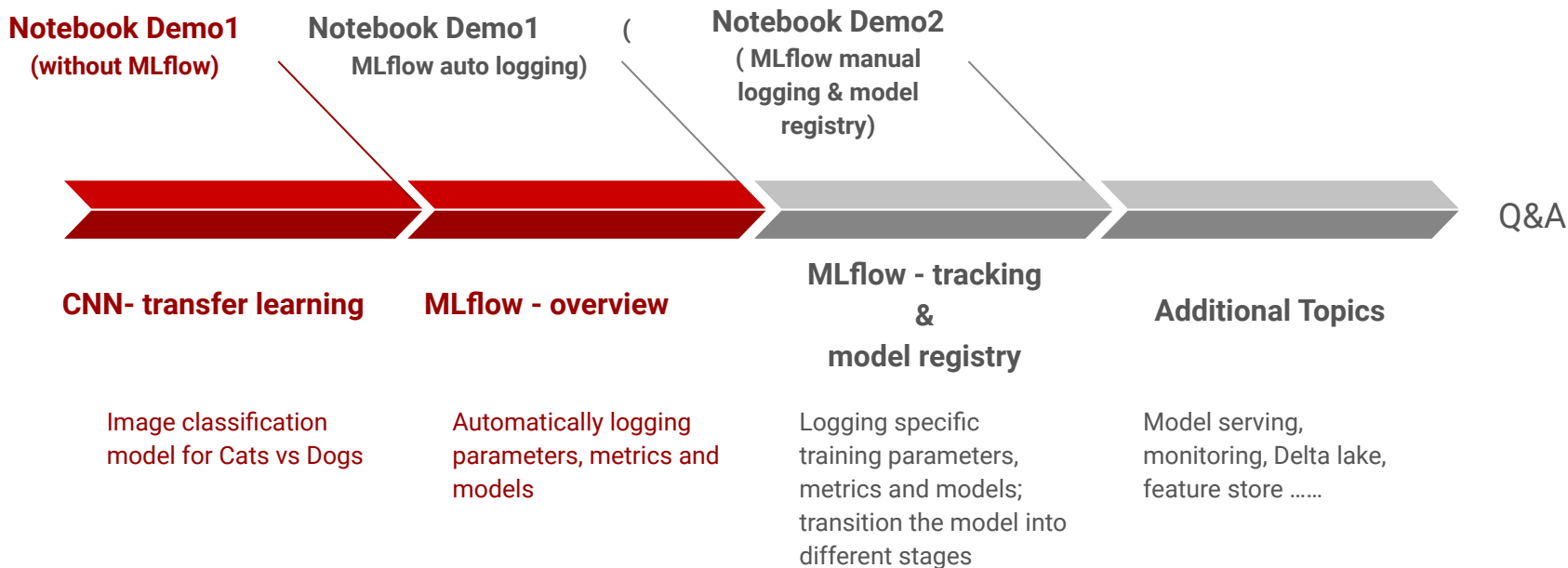
Outline



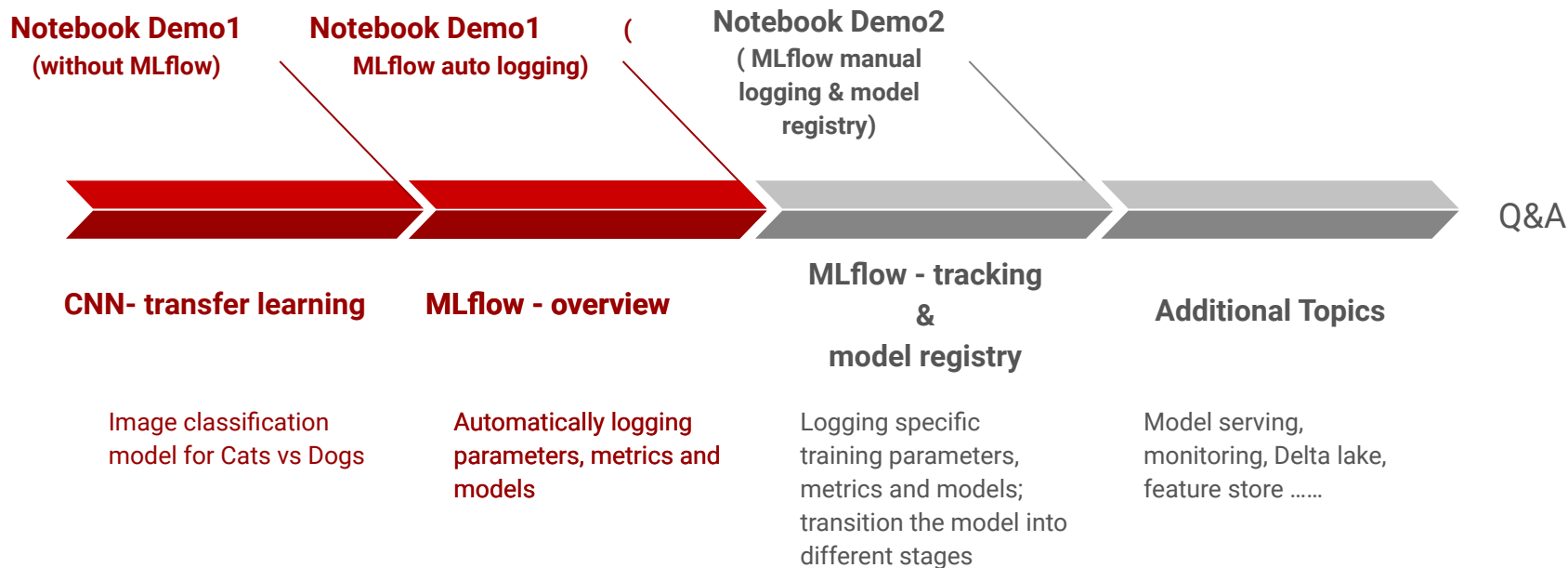
Outline



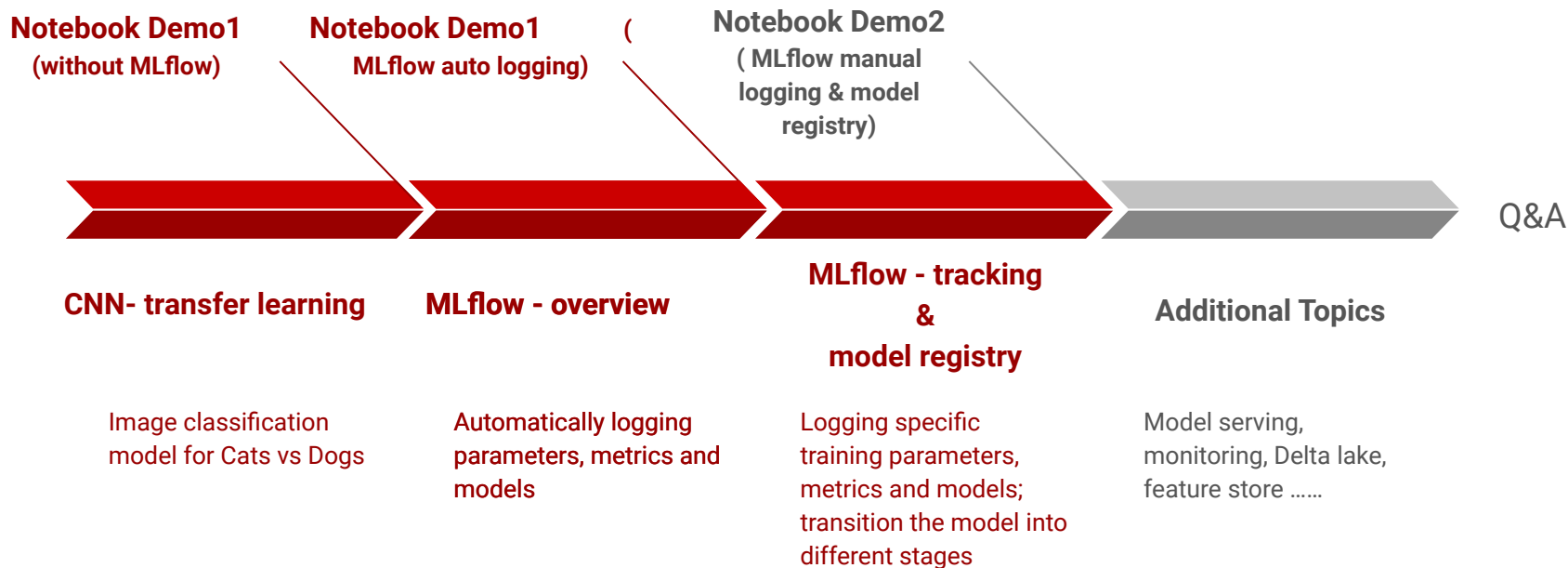
Outline



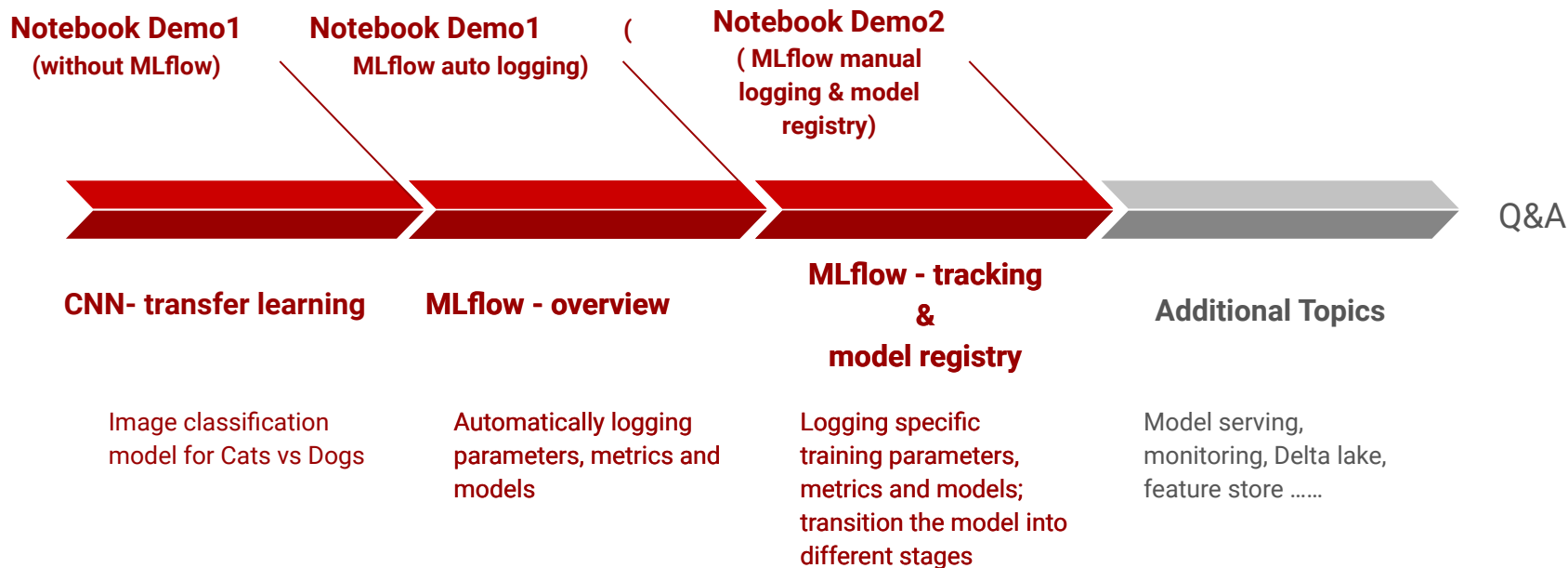
Outline



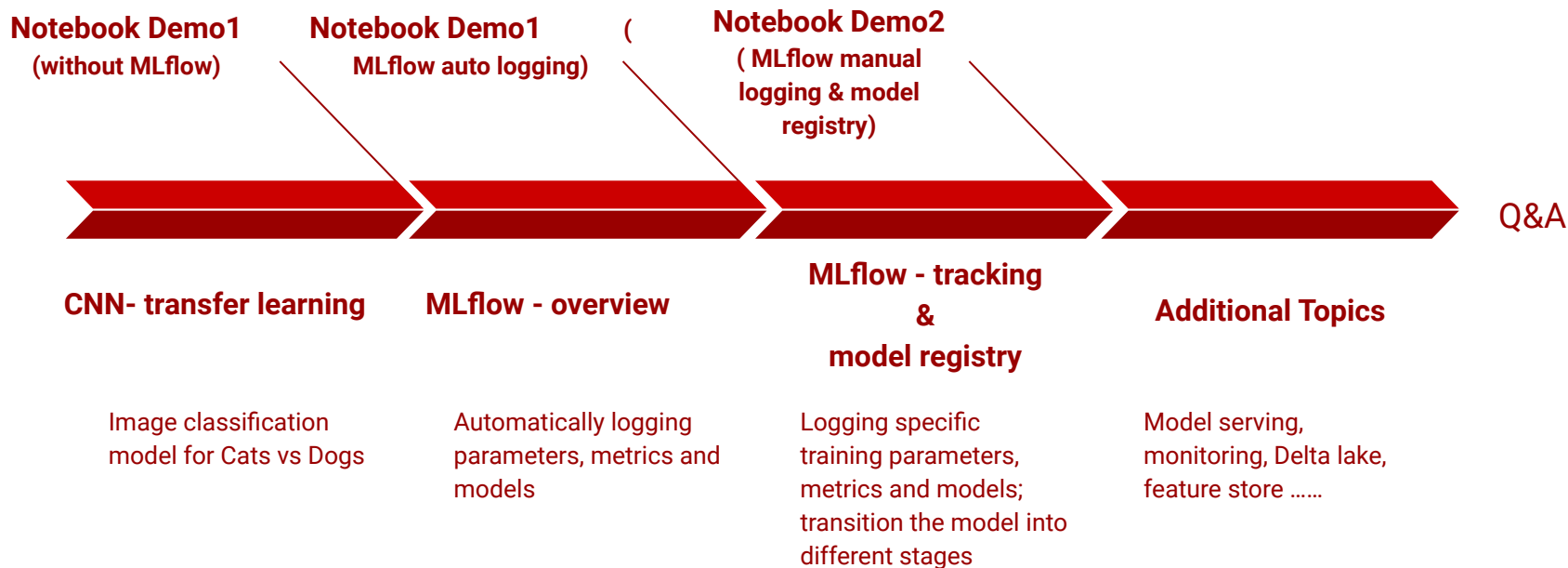
Outline



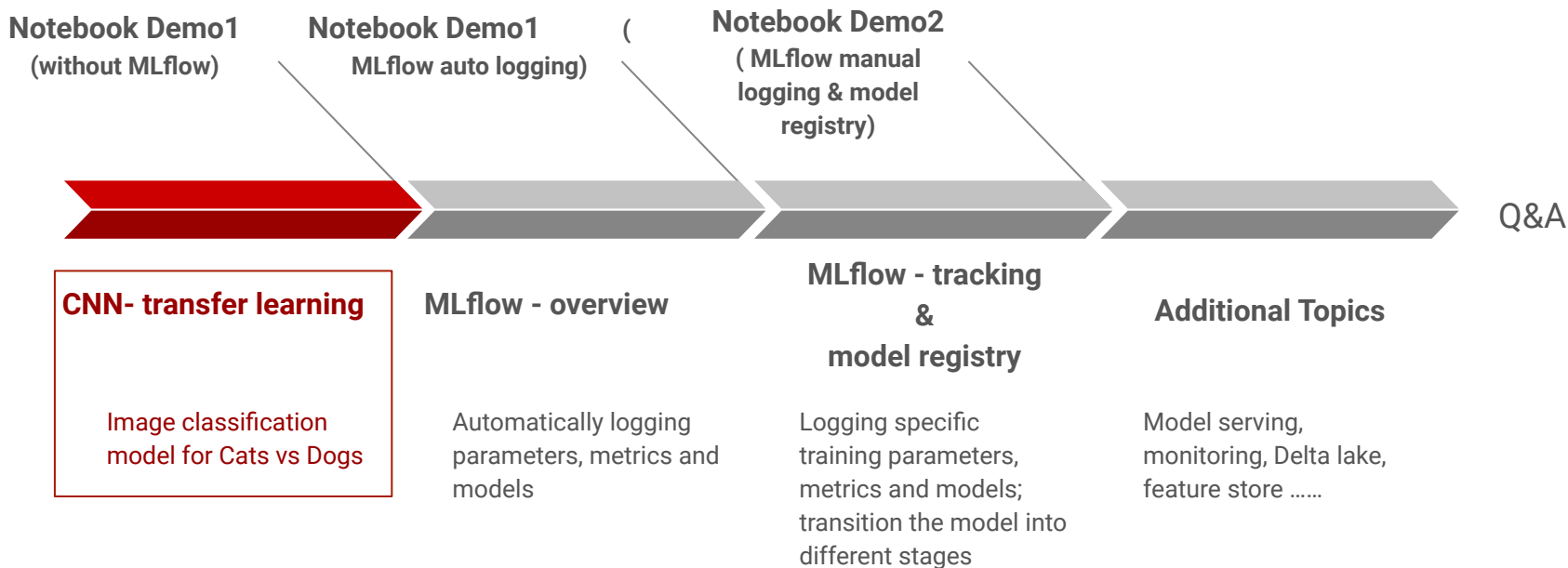
Outline



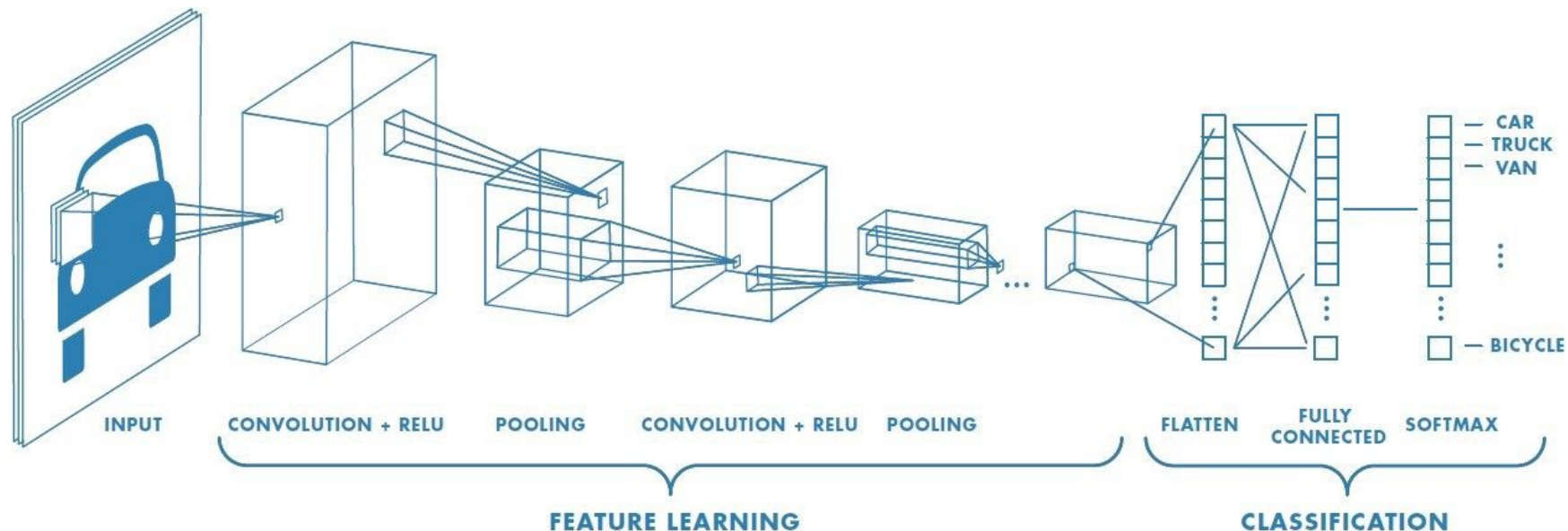
Outline



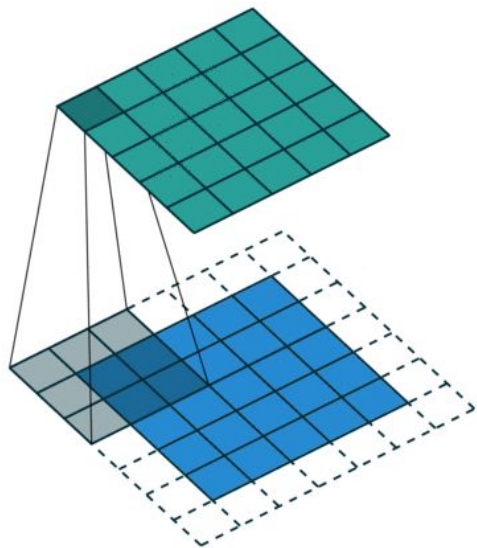
Outline



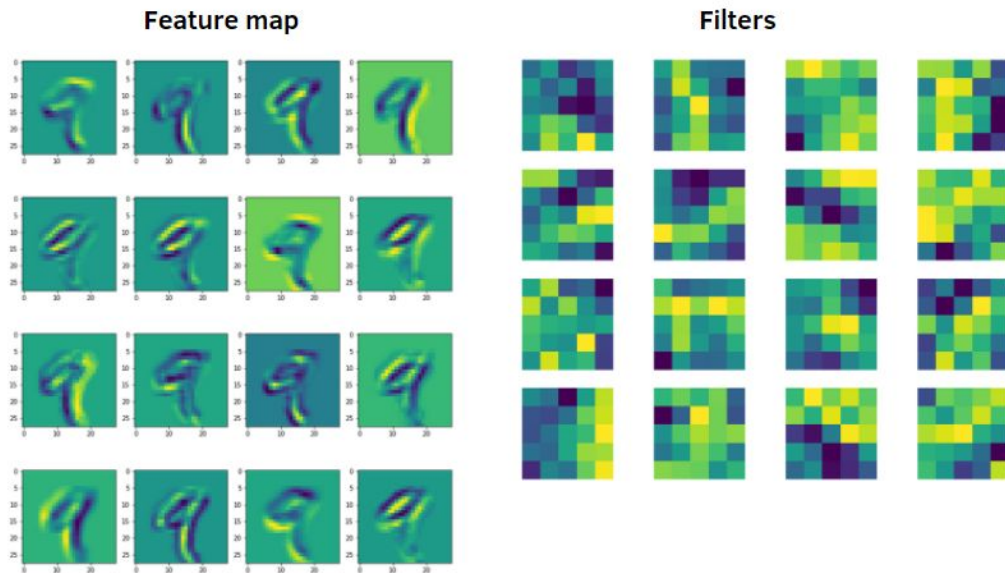
Convolutional Neural Networks



Filters and Feature Maps

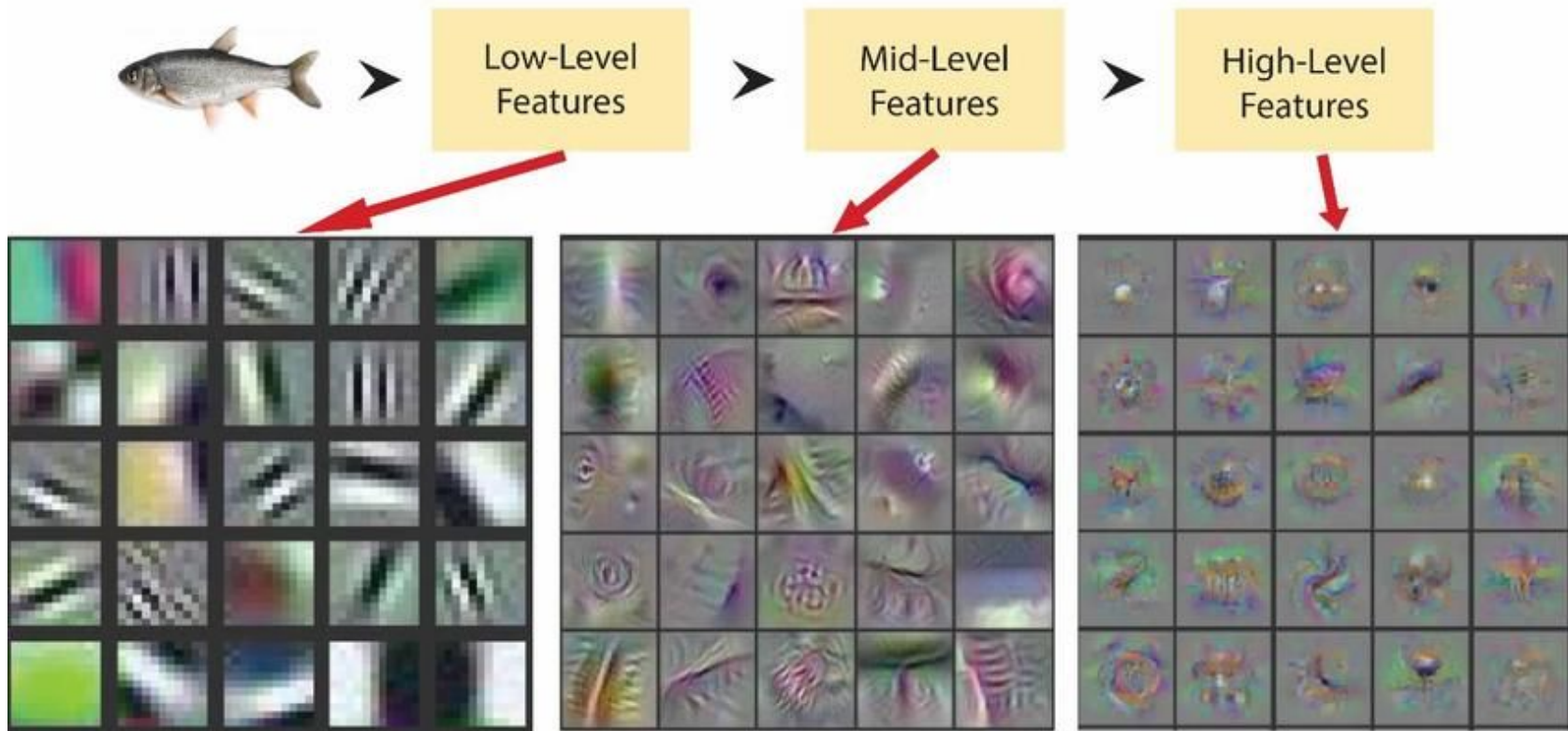


[Image source](#)

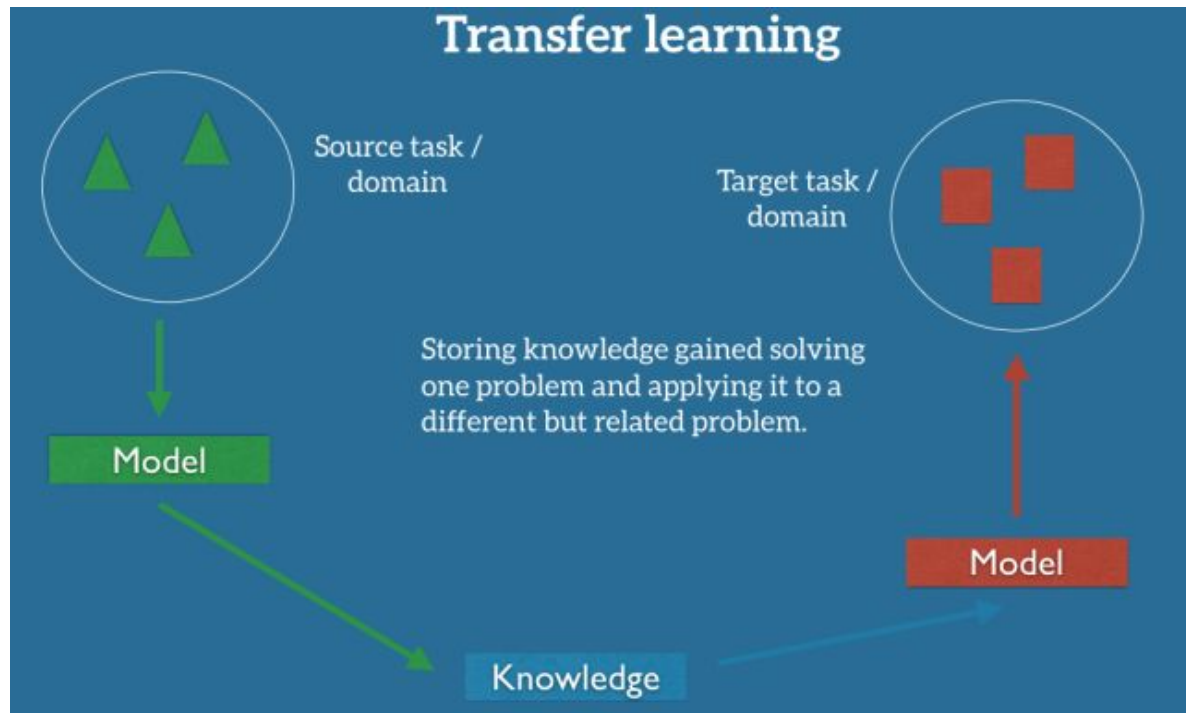
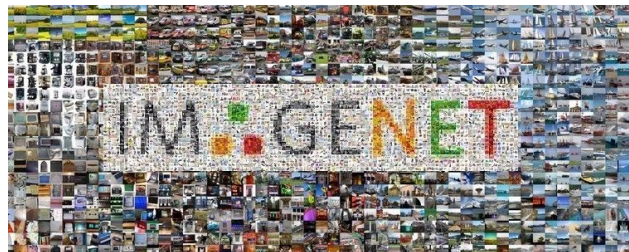


[Image
source](#)

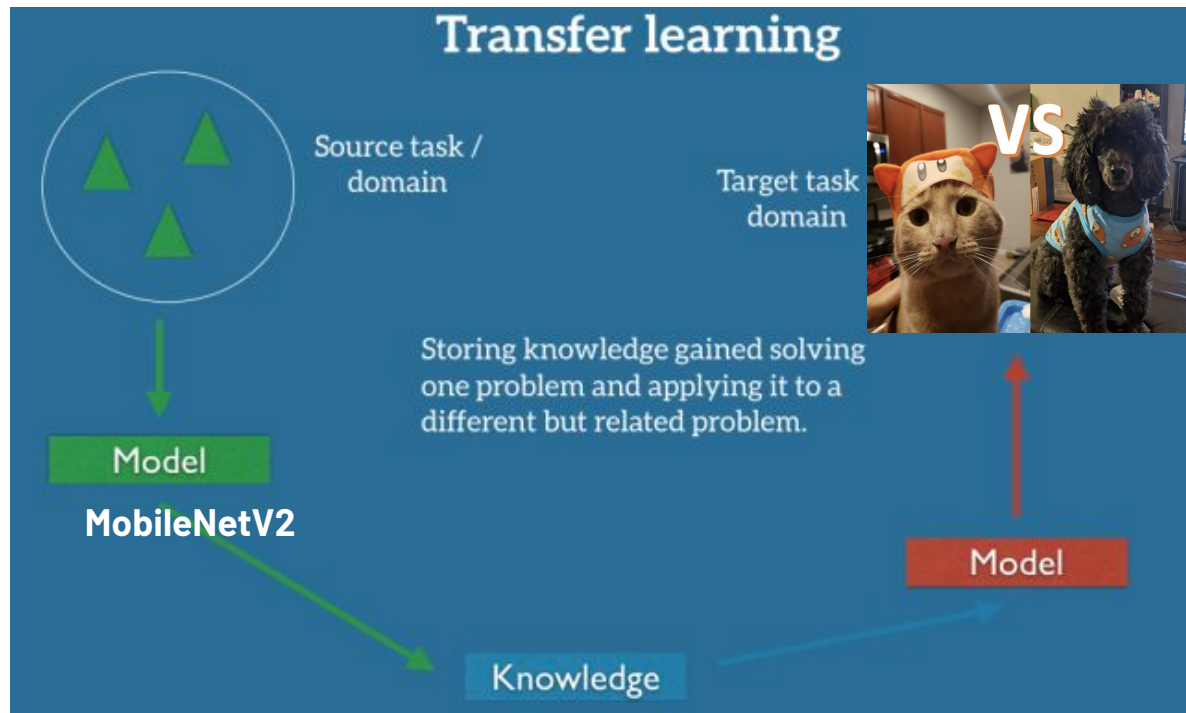
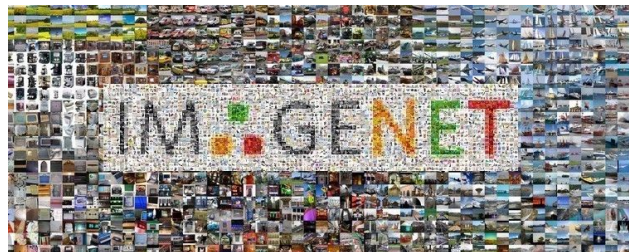
Deeper Layers → More Complex Features



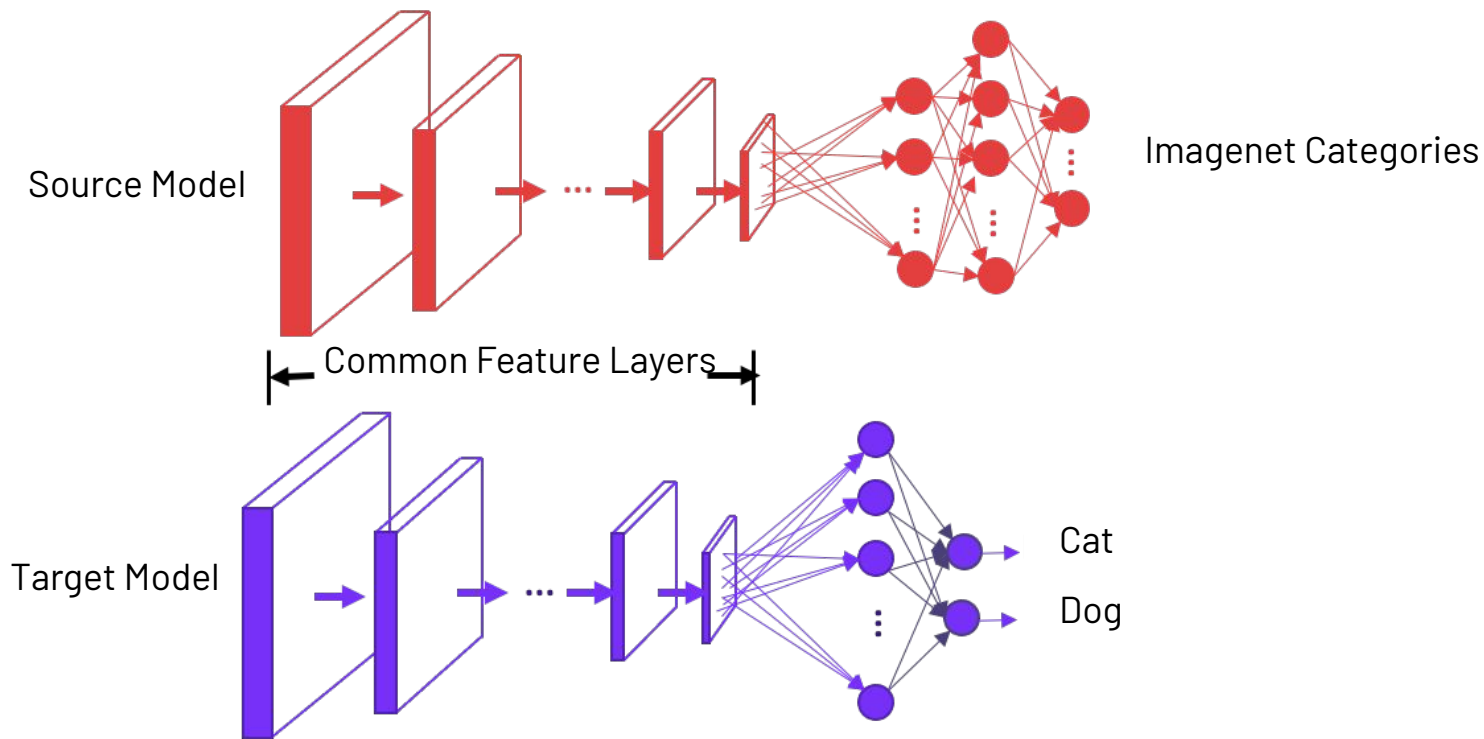
Transfer Learning



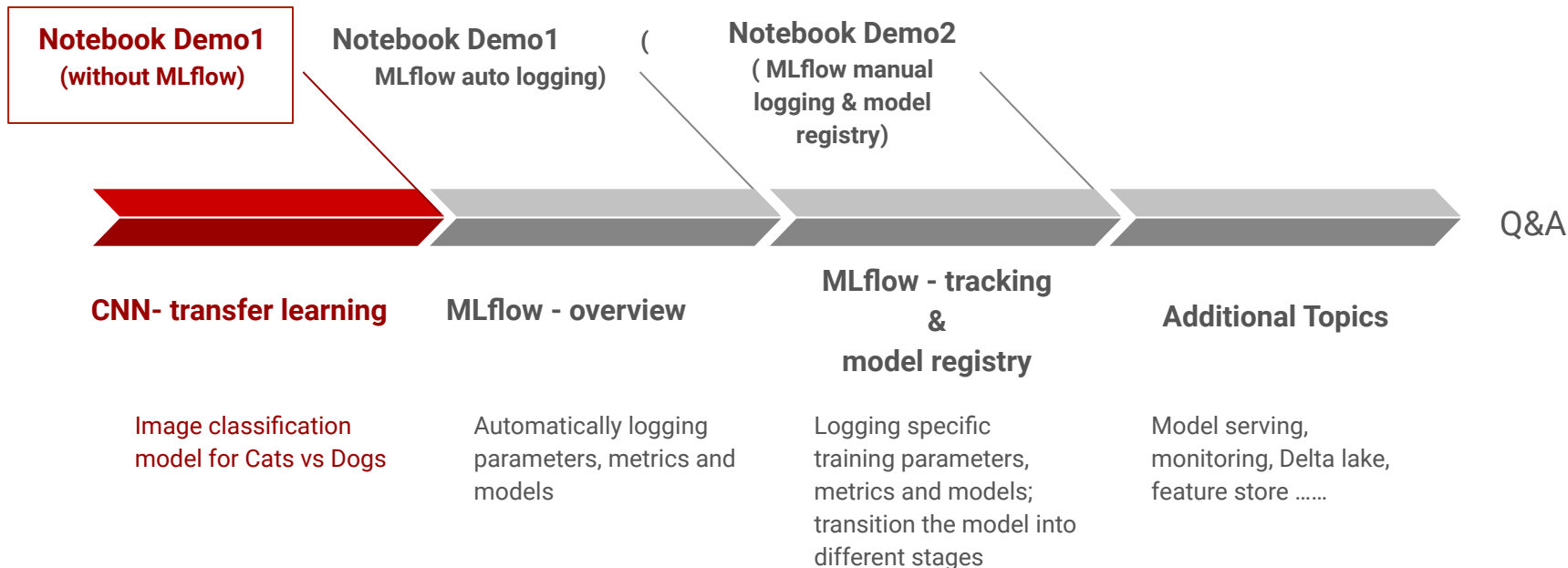
Transfer Learning



Feature Extraction



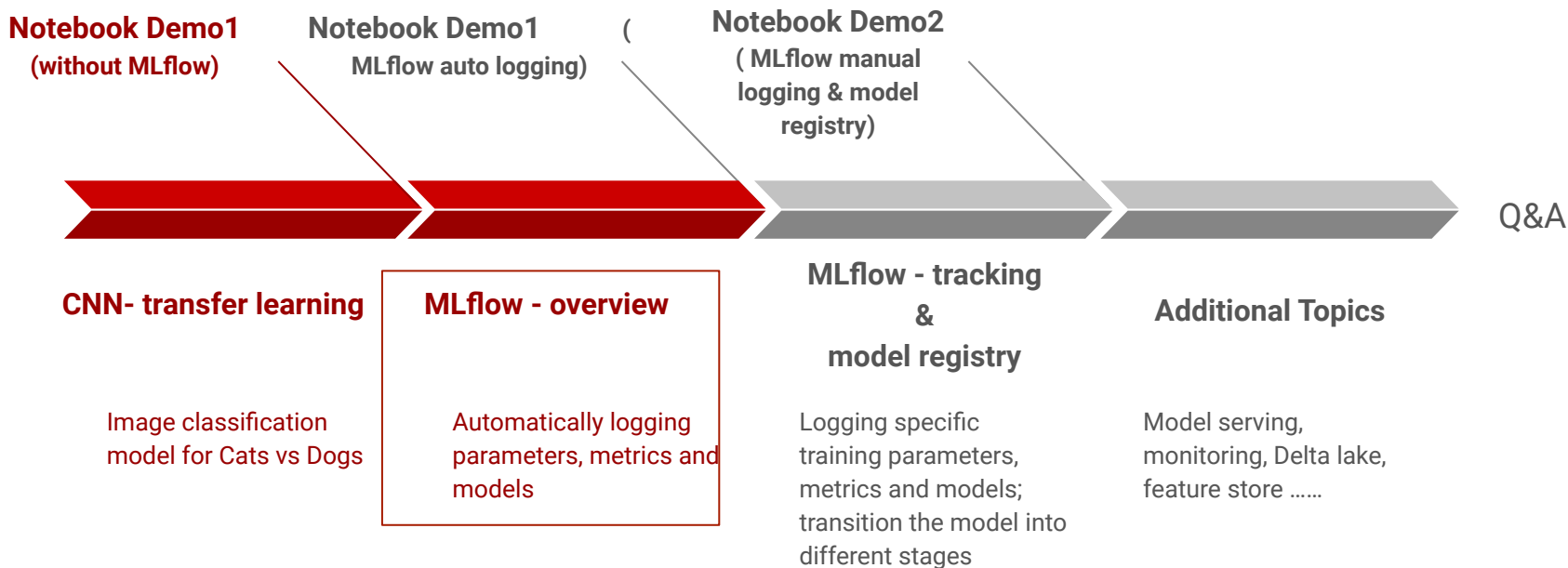
Outline



Notebook Demo 1

(without MLflow)

Outline



MLflow is the most successful MLOps project

Sci-kit Learn: **34M** / month

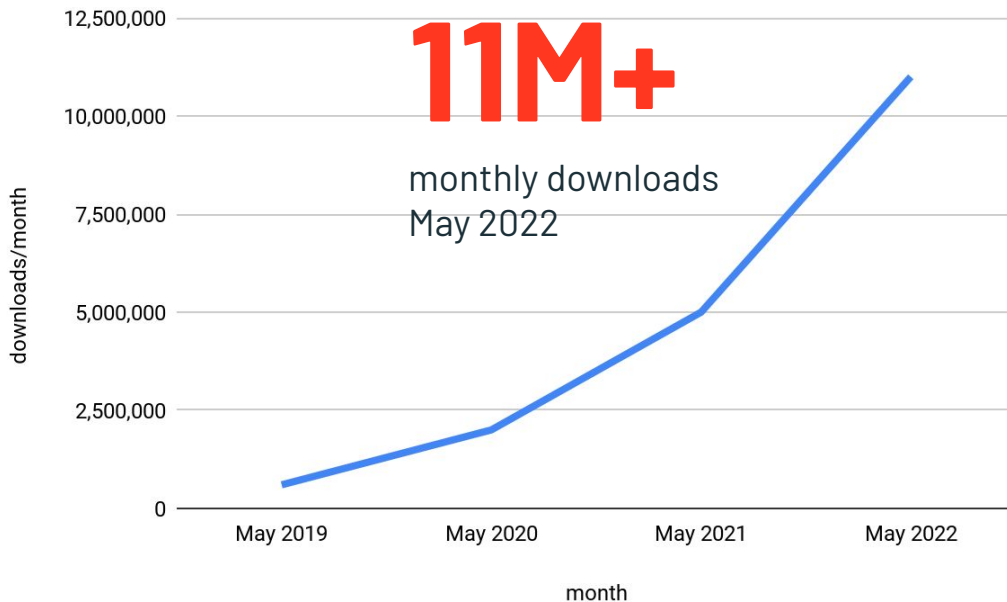
TensorFlow: **19M** / month

MLflow: **11M** / month

PyTorch: **10M** / month

...

TFX: **319K** / month



MLflow

An open source platform for the machine learning lifecycle

pip install mlflow

mlflow
Tracking

Record and query
experiments: code,
data, config,
results

mlflow
Projects

Packaging format
for reproducible
runs on any
platform

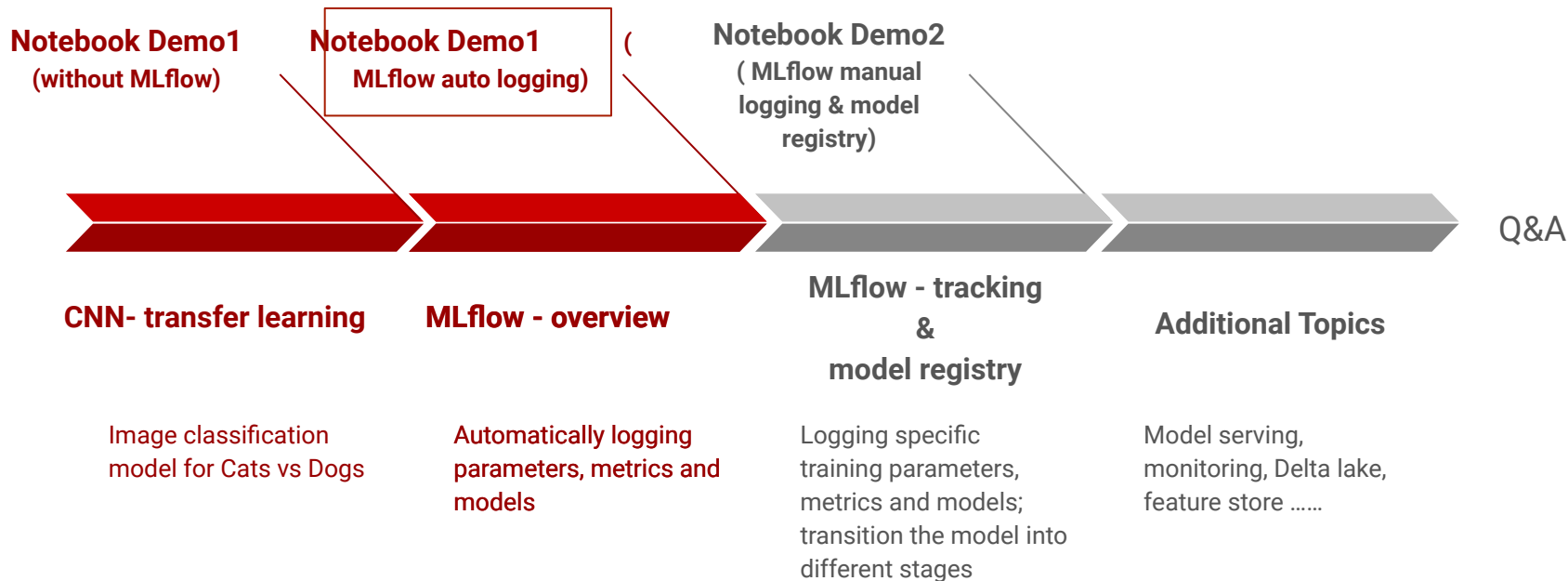
mlflow
Models

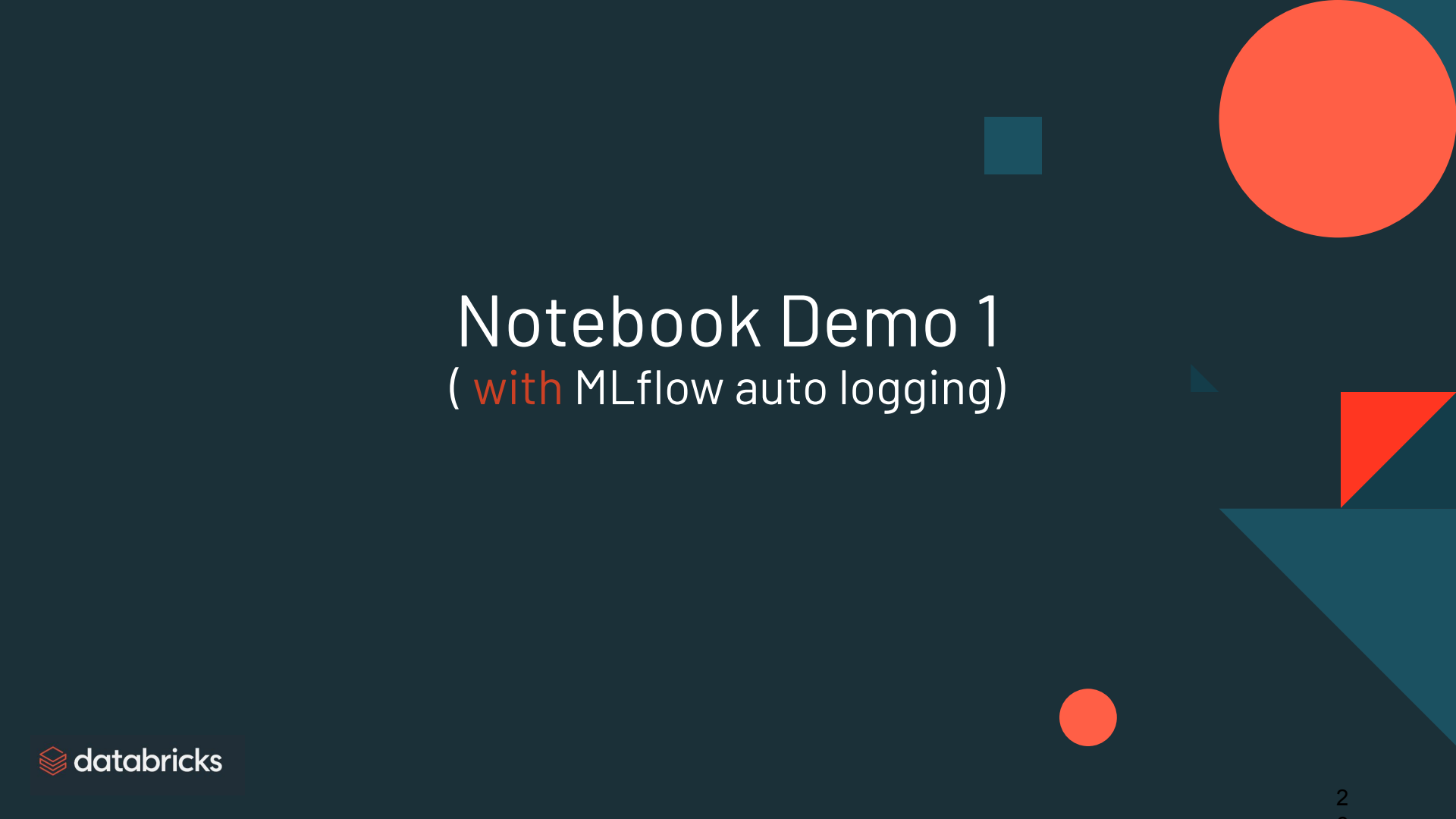
General format
that standardizes
deployment paths

mlflow
Model
Registry

Centralized and
collaborative
model lifecycle
management

Outline

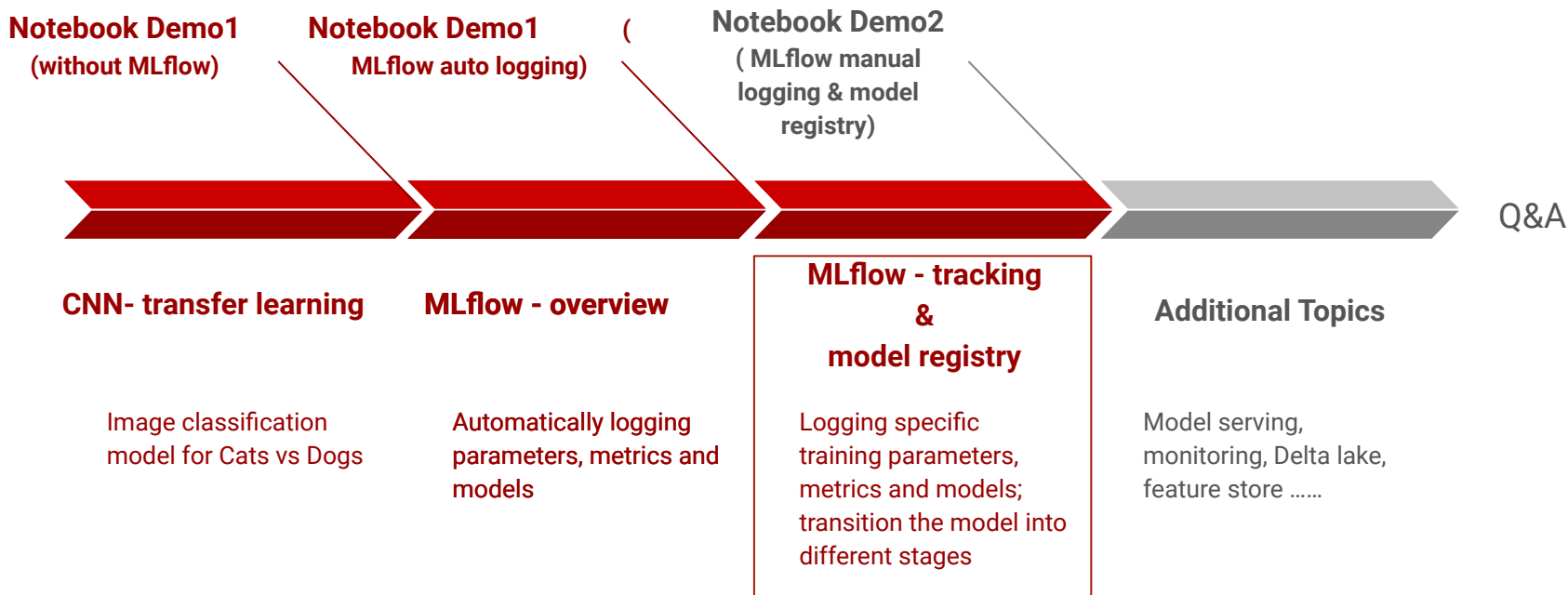




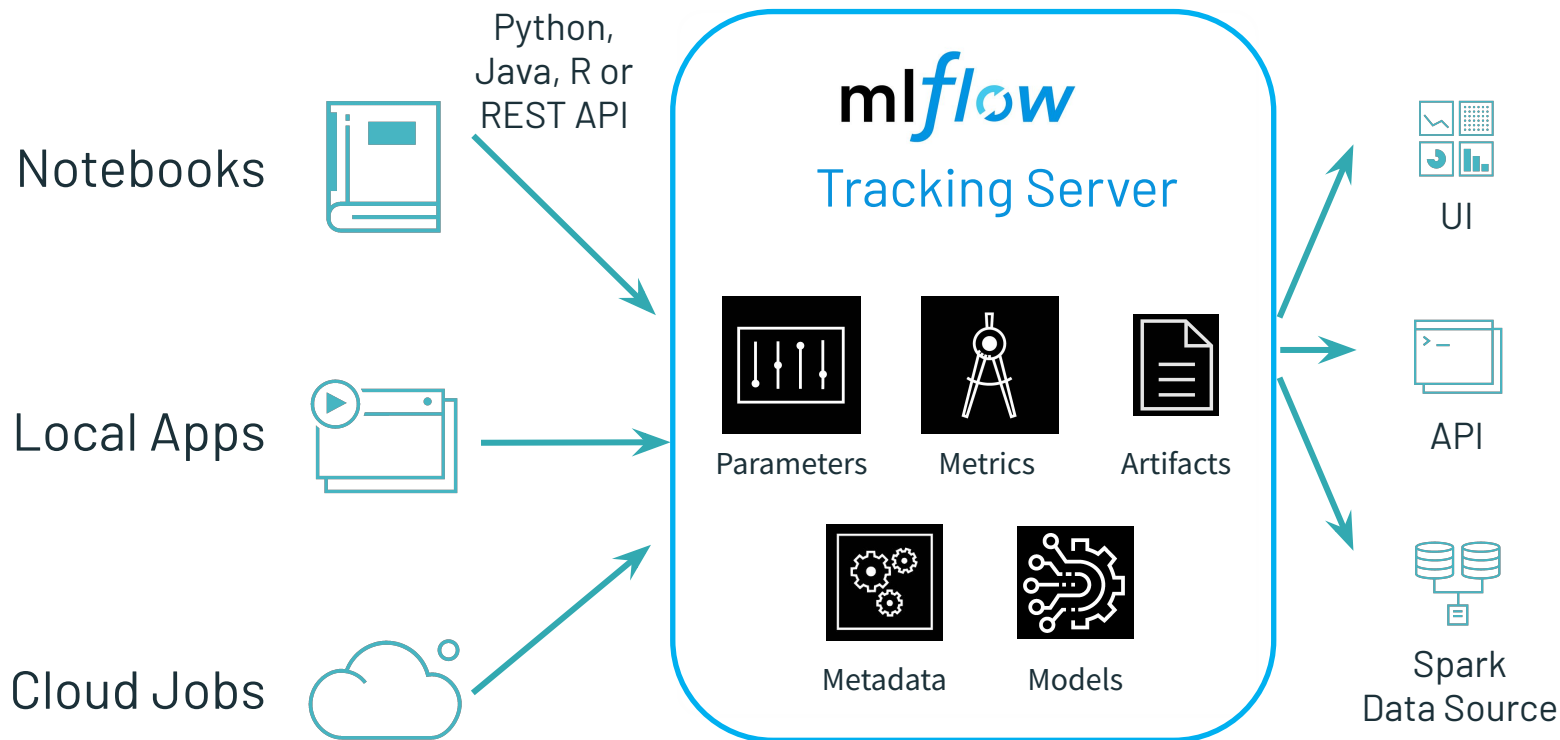
Notebook Demo 1

(**with** MLflow auto logging)

Outline



MLflow Tracking



MLflow Model Registry

- **Central Repository:** Unique named registered models for discovery across data teams
- **Model Registry Workflow:** Provides UI and API for registry operations
- **Model Versioning:** Allow multiple versions of model in different stages
- **Model Stages:** Allow stage transition: none, staging, production, or archived
- **CI/CD Integration:** Easily load a specific version for testing and inspection
- **Model Lineage:** Provides model description, lineage and activities

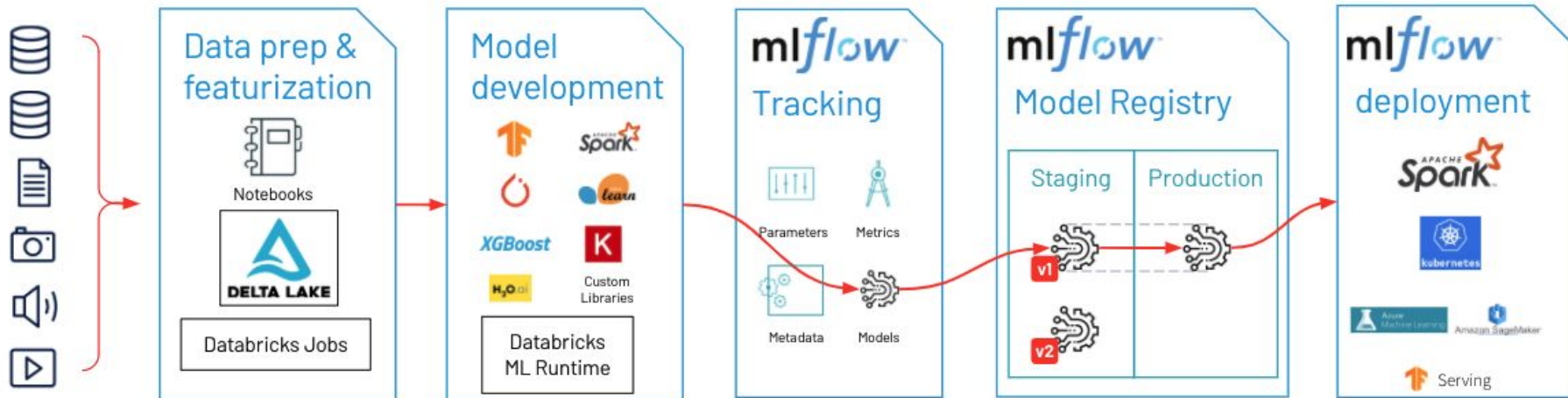
The Full ML Lifecycle

Data Scientists build features.
Data Engineers provide infra for automating featurization.

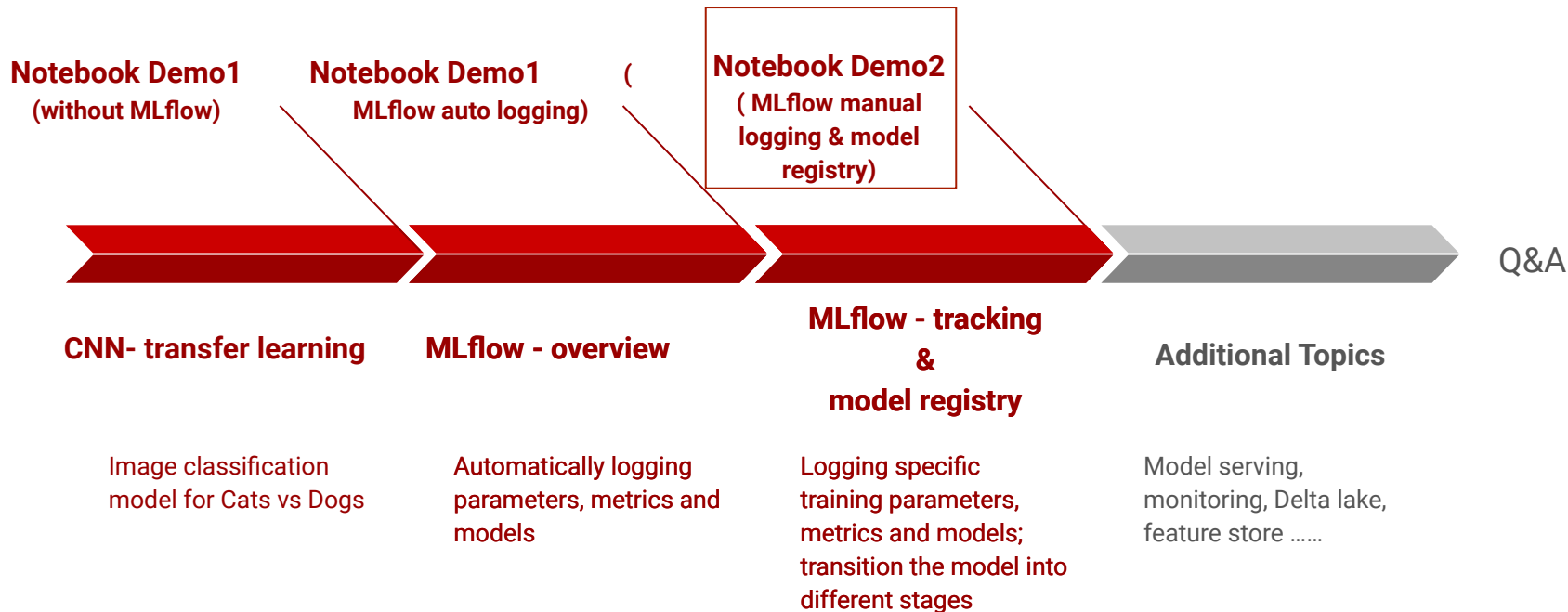
Data Scientists build models and log them to MLflow, which records environment info.

Data Scientists move models to Staging.

Deployment Engineers manage CI/CD tools which promote models to Production.



Outline

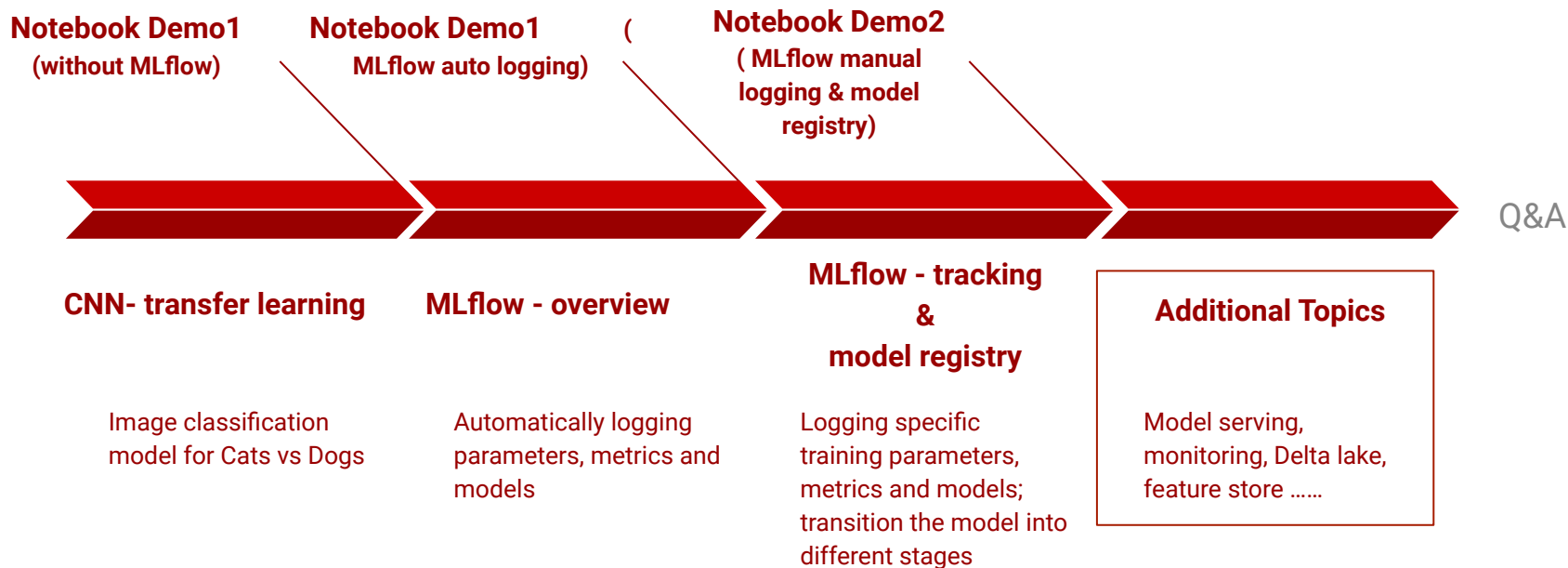




Notebook Demo 2

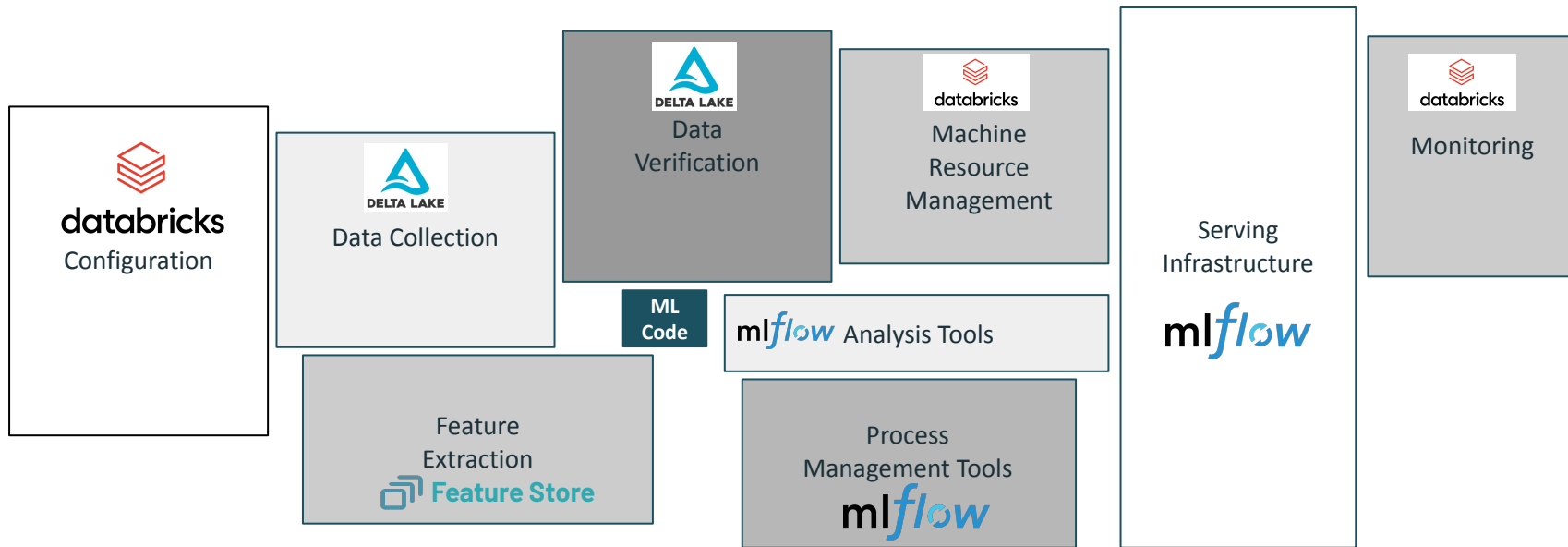
(MLflow manual logging + model registry)

Outline



Hardest Part of ML isn't ML

"Hidden Technical Debt in Machine Learning Systems," Google NIPS 2015



Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.



github.com/feifeiwww

https://github.com/feifeiwww/20220726_Databricks_Demo_Transfer_Learning_with_MLflow

References

- **CNN transfer learning tensorflow example:**
https://www.tensorflow.org/tutorials/images/transfer_learning
- **Sign up for Databricks community edition:**
<https://docs.databricks.com/getting-started/community-edition.html>
- **Databricks community edition login:**
<https://community.cloud.databricks.com/login.html>
- **Get started with Databricks as a data scientist:**
<https://docs.databricks.com/getting-started/quick-start.html>
- **Image classification on ImageNet state of the art:**
<https://paperswithcode.com/sota/image-classification-on-imagenet>
- **My Github repo:**
https://github.com/feifeiwww/20220726_Databricks_Demo_Transfer_Learning_with_MLflow



Thank you! Questions?

feifei.wang@databricks.com