Declarative Transfer Learning from Deep CNNs at Scale

Supun Nakandala and Arun Kumar

{snakanda, arunkk}@eng.ucsd.edu

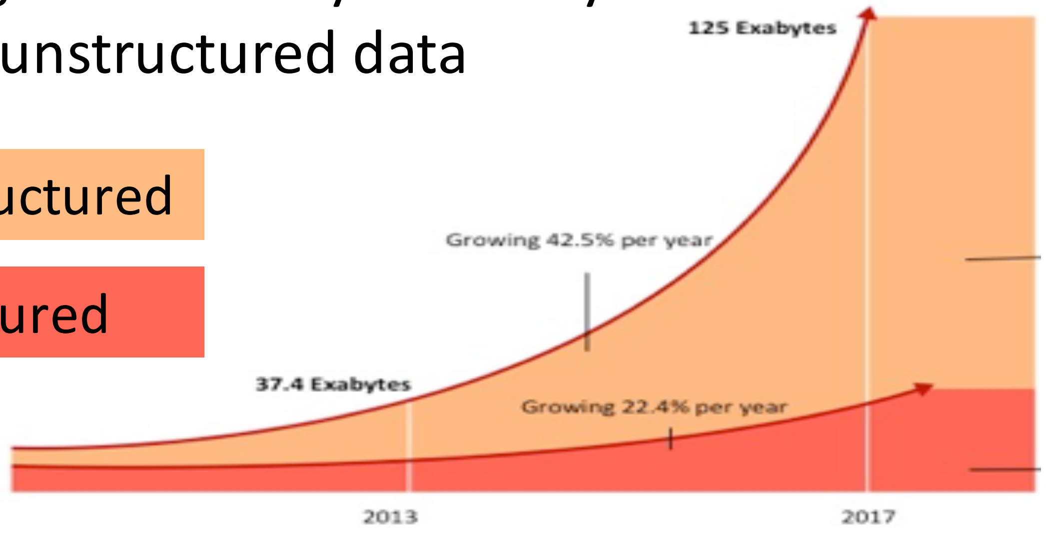
UCSDCSE
Computer Science and Engineering

Growth of unstructured data

Data growth mainly driven by unstructured data

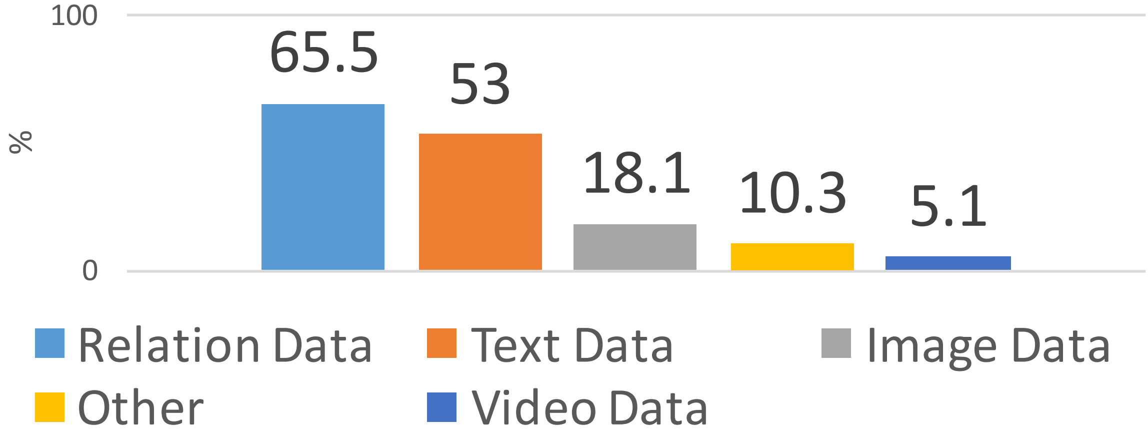
unstructured

structured



Source: ETC@USC

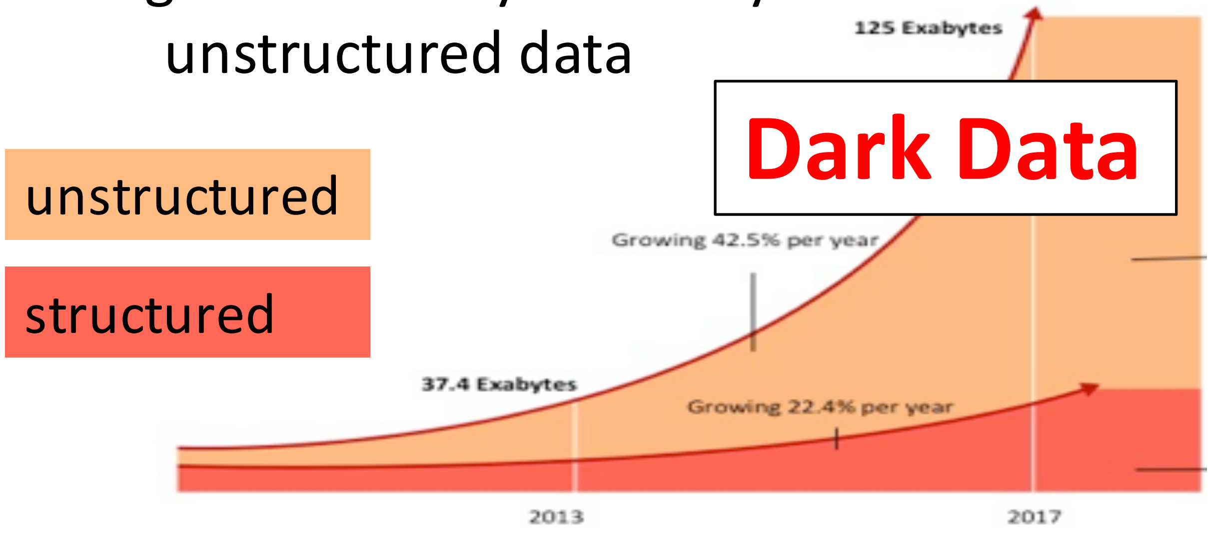
What type of data is used by Data Scientists?



Source: 2017 Kaggle Survey

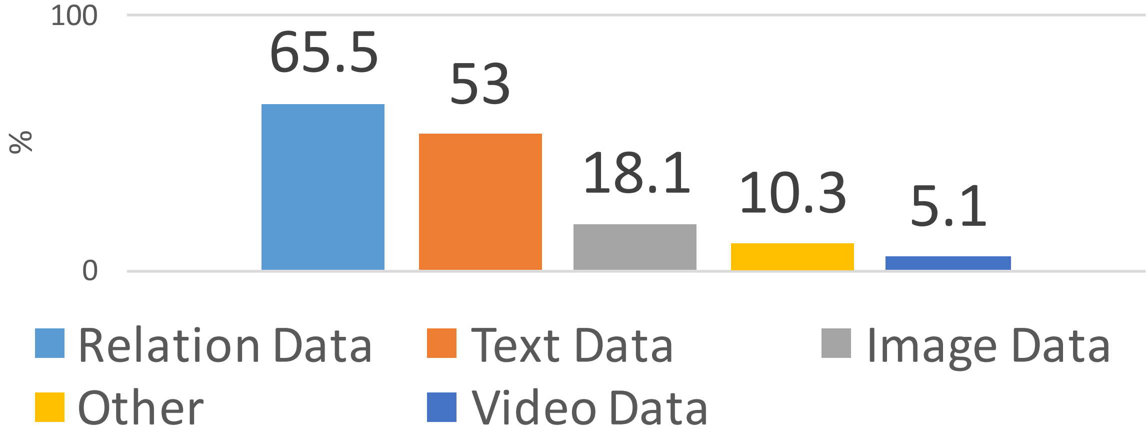
Growth of unstructured data

Data growth mainly driven by unstructured data

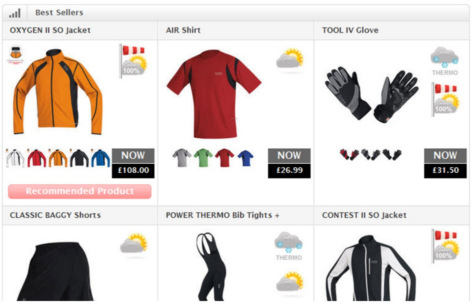


Source: ETC@USC

What type of data is used by Data Scientists?

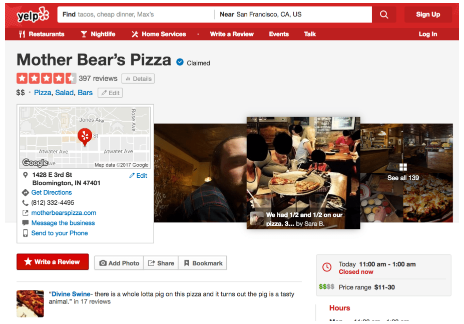


Source: 2017 Kaggle Survey



e-Commerce

Healthcare



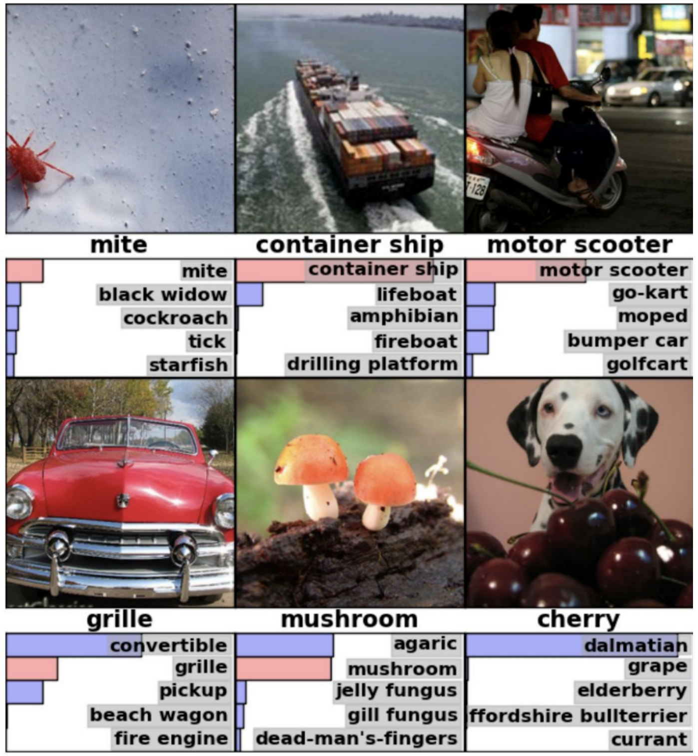
Social Media

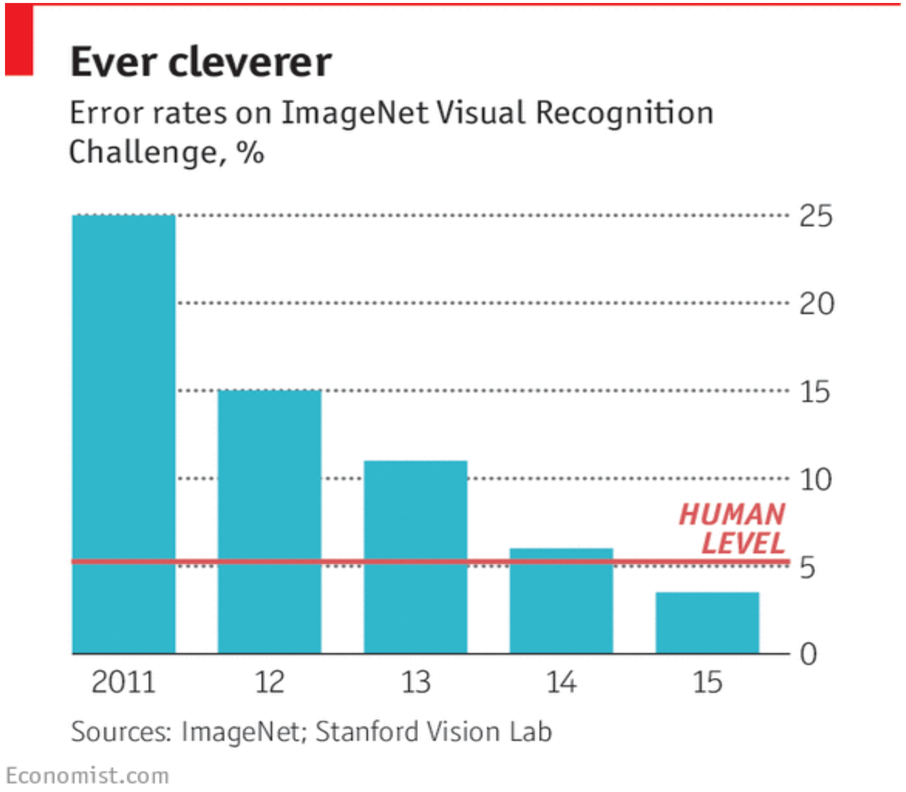
Opportunity: CNN

Deep Convolution Neural Networks (CNN) provide opportunities to holistically integrate image data with analytics pipelines.

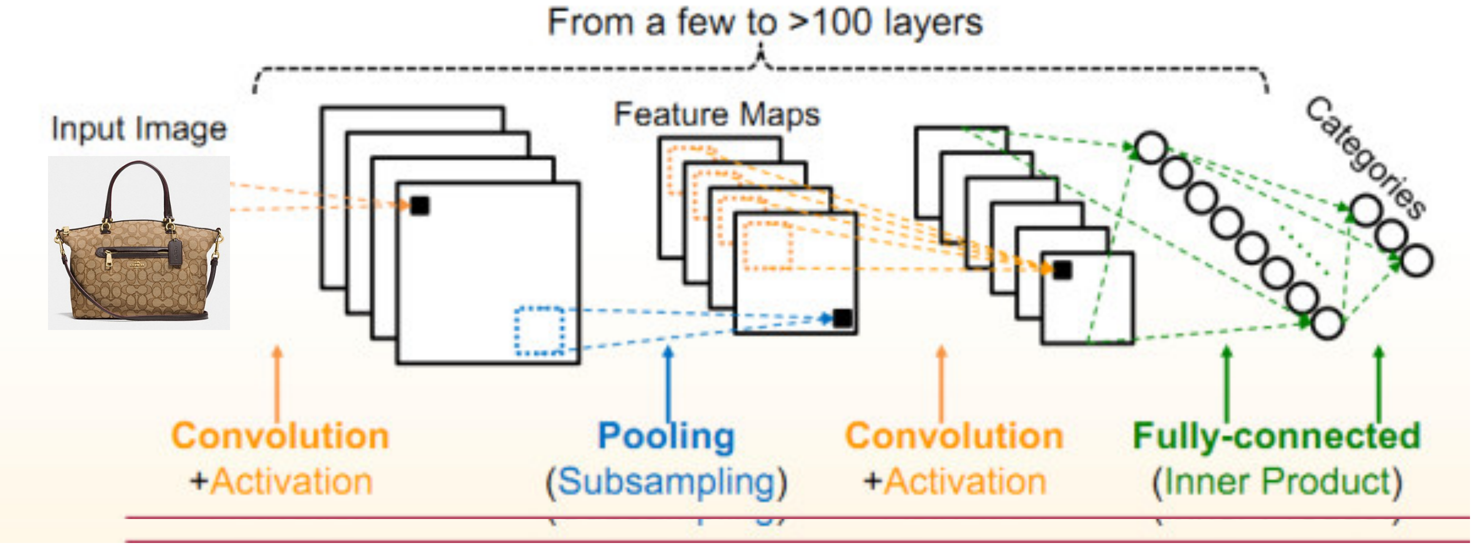
IMAGENET

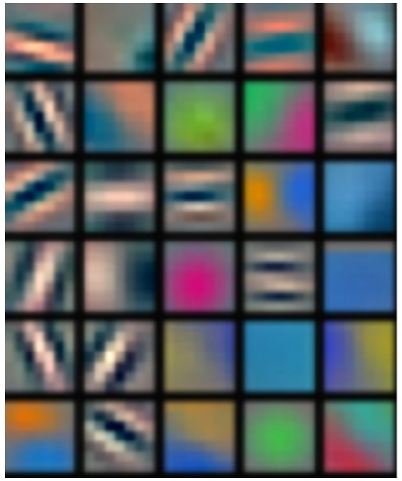
- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



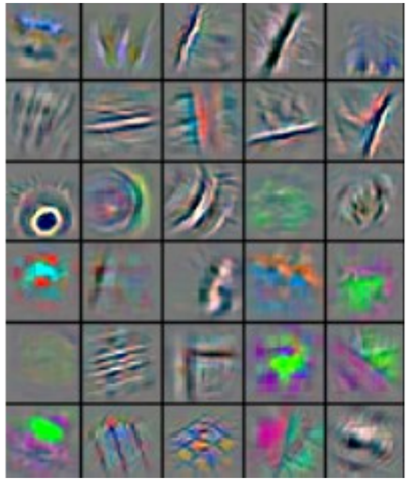


CNN: Hierarchical Feature Extractors

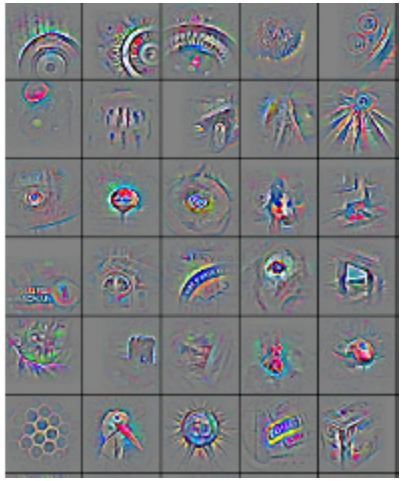




Low level features



Mid level features



High level features

CNN: Training Limitations

Lot of labelled training data

Lot of compute power

Time consuming

"Dark art" of hyperparameter tuning

CNN: Training Limitations

Lo

“Transfer Learning” mitigates these limitations

“Dark art” of hyperparameter tuning

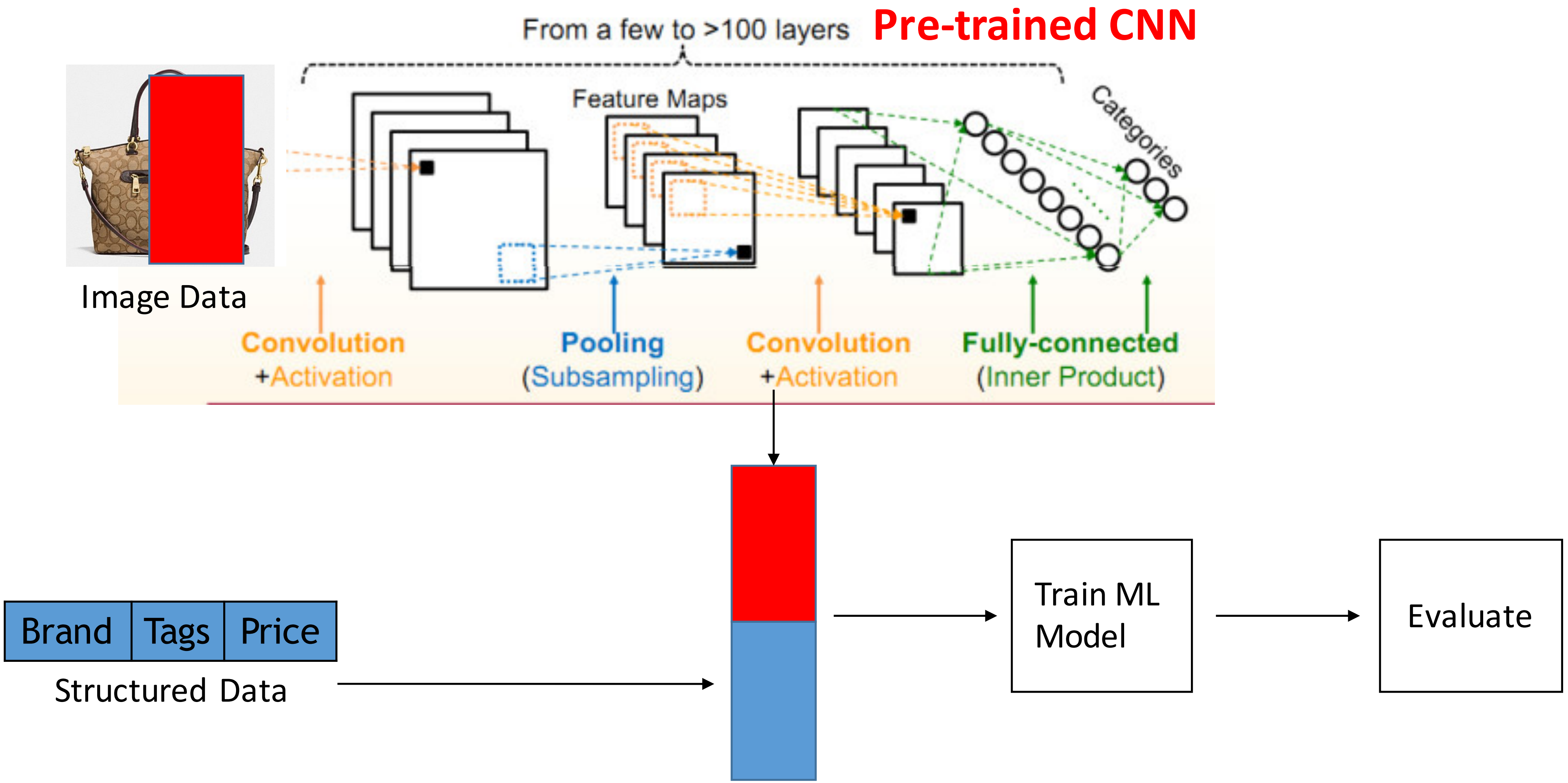
Outline

Example and Motivations

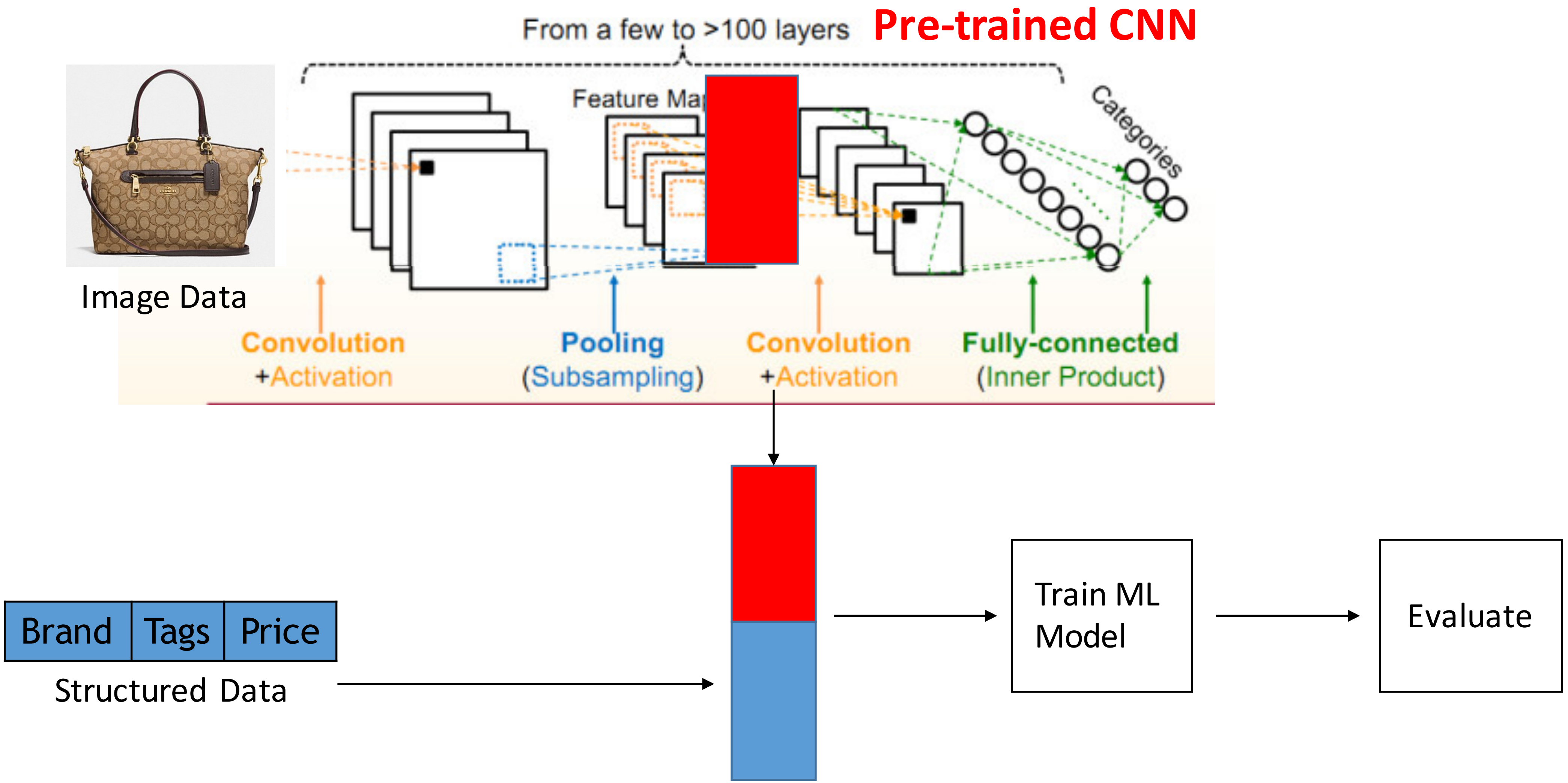
Our System Vista

Experimental Evaluation

Transfer Learning: CNNs for the other 90%

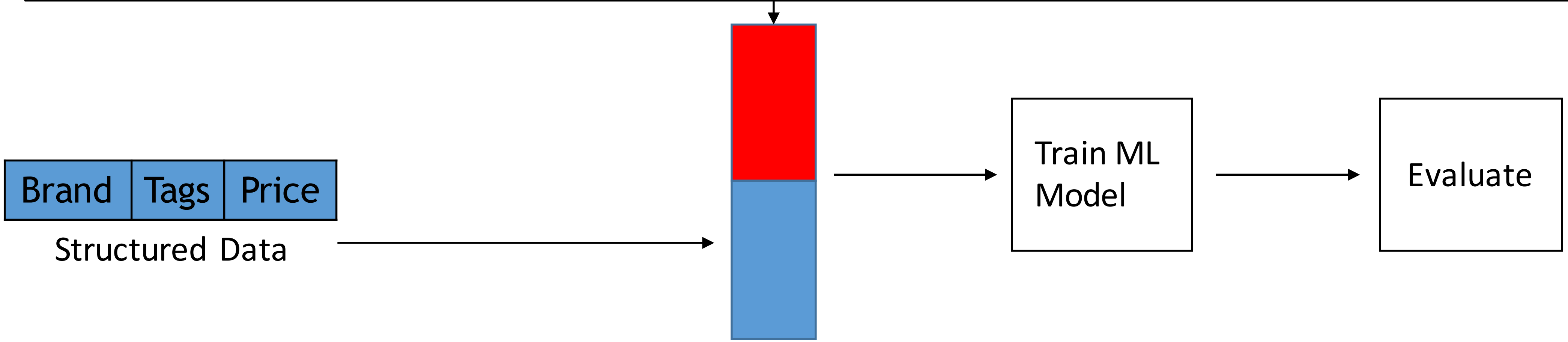


Transfer Learning: CNNs for the other 90%

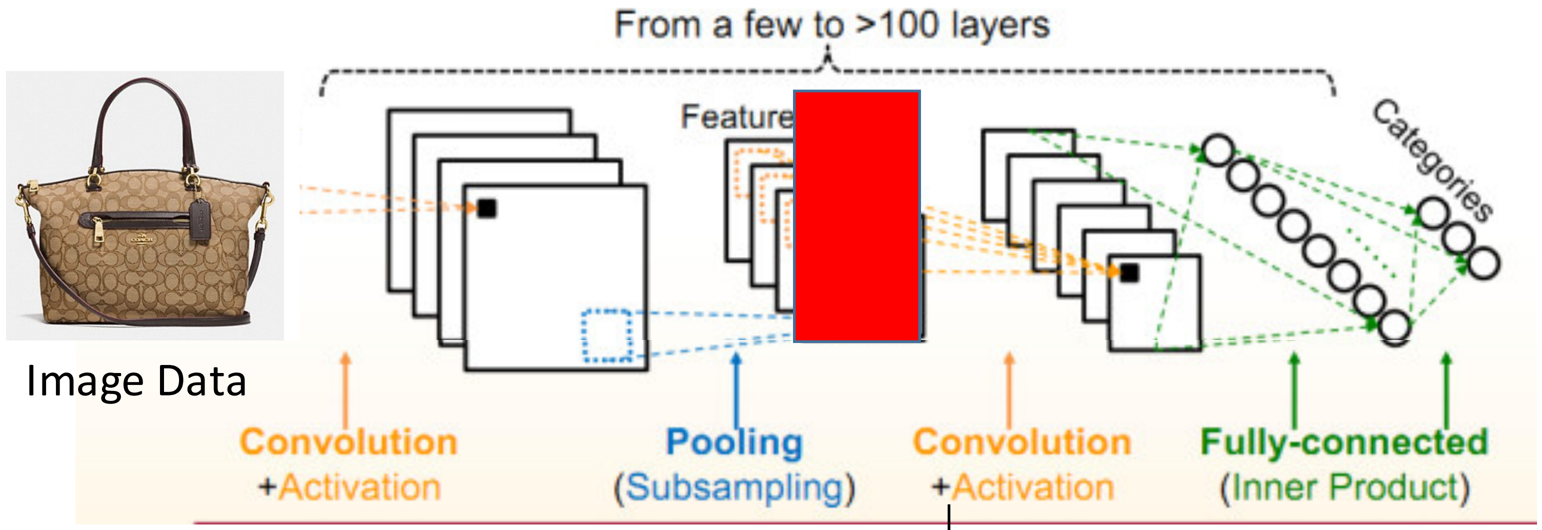


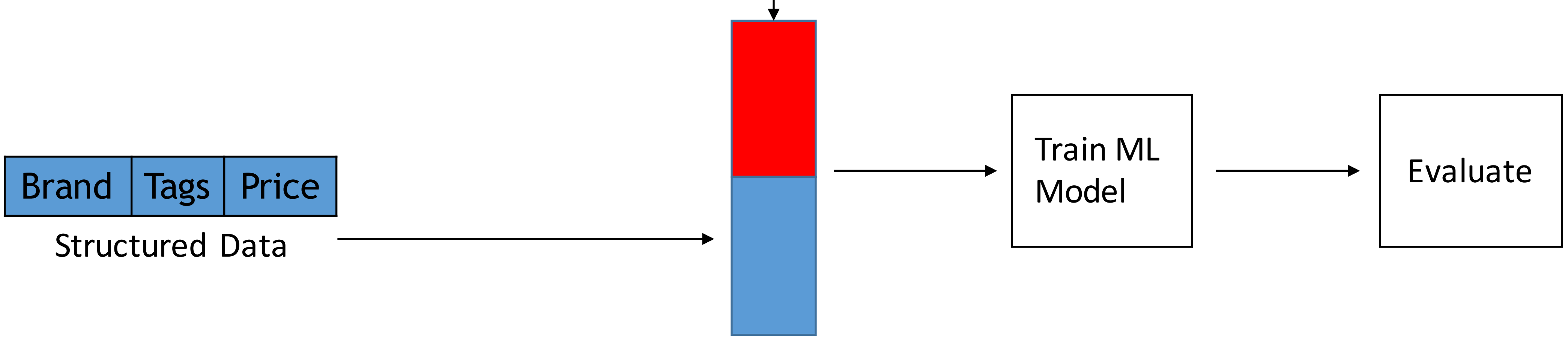
Transfer Learning: CNNs for the other 90%

Which layer will result in the best accuracy?

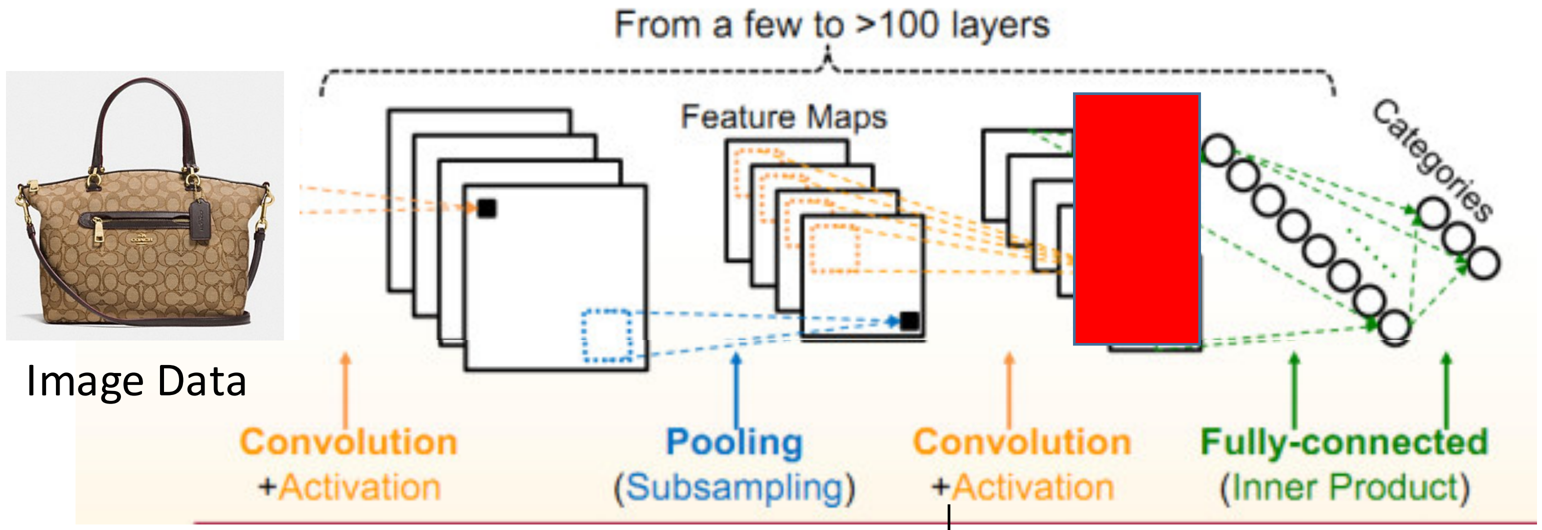


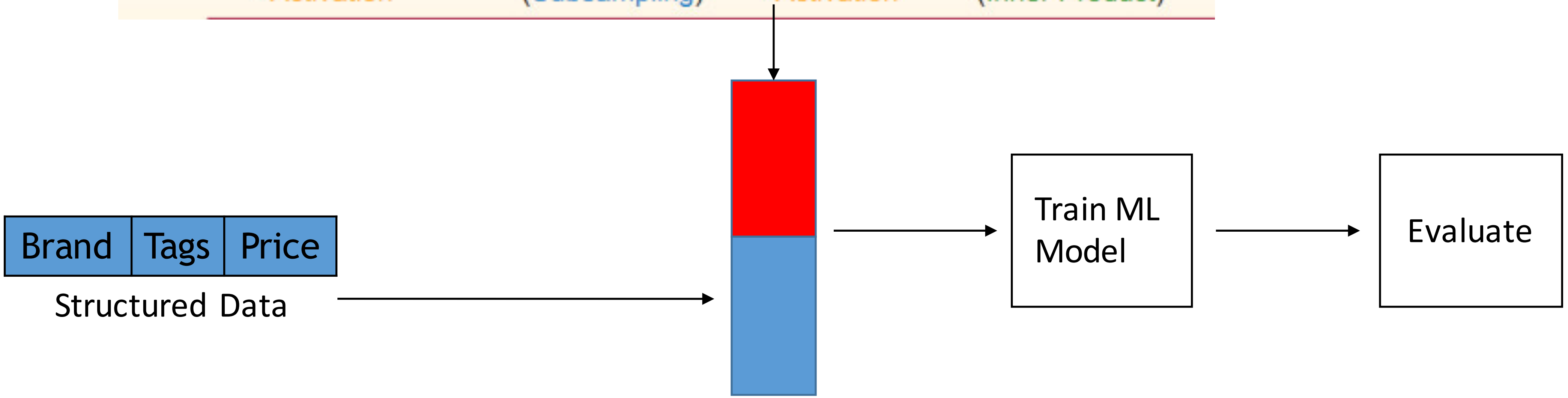
Transfer Learning: Bottleneck



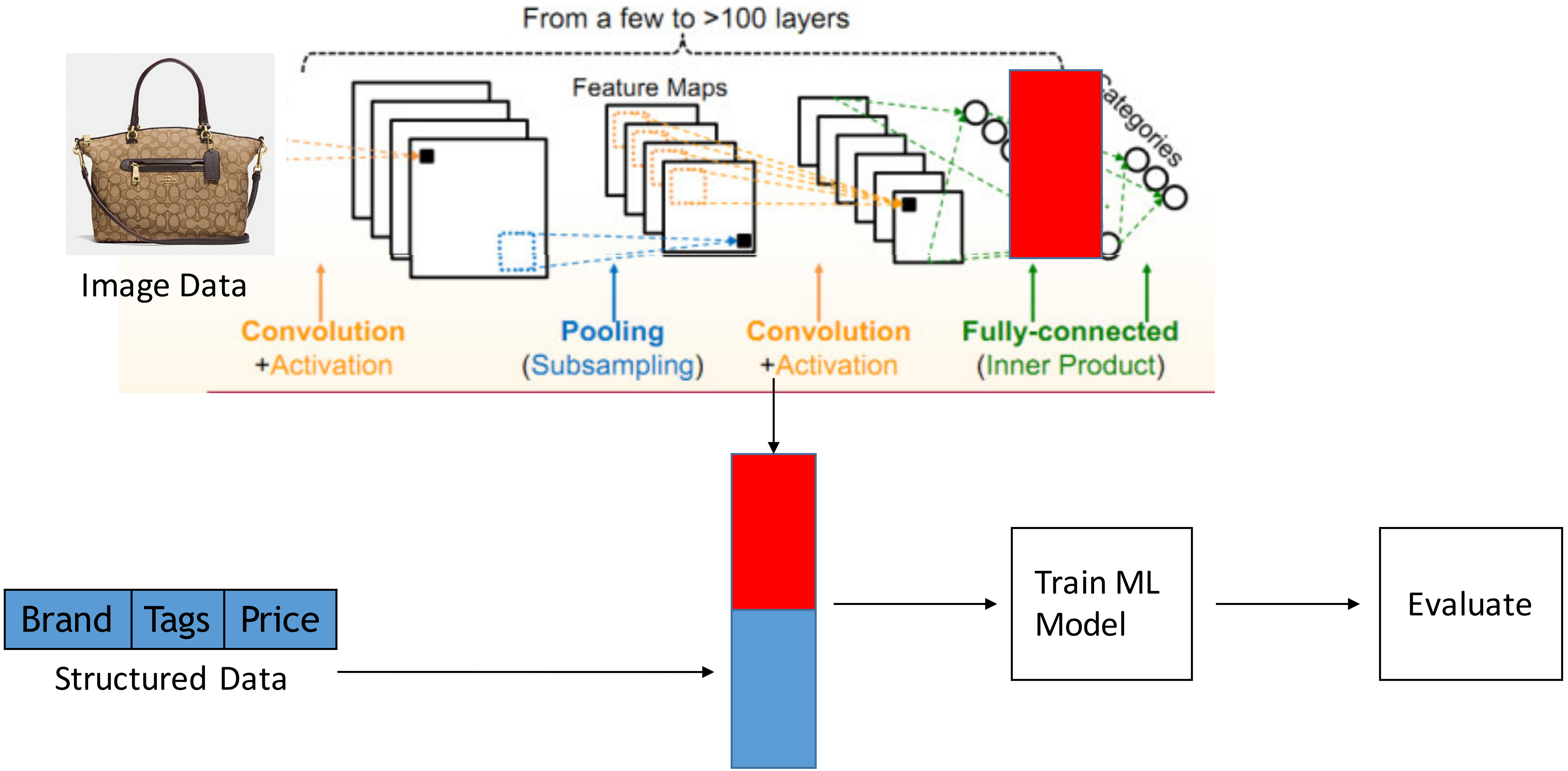


Transfer Learning: Bottleneck

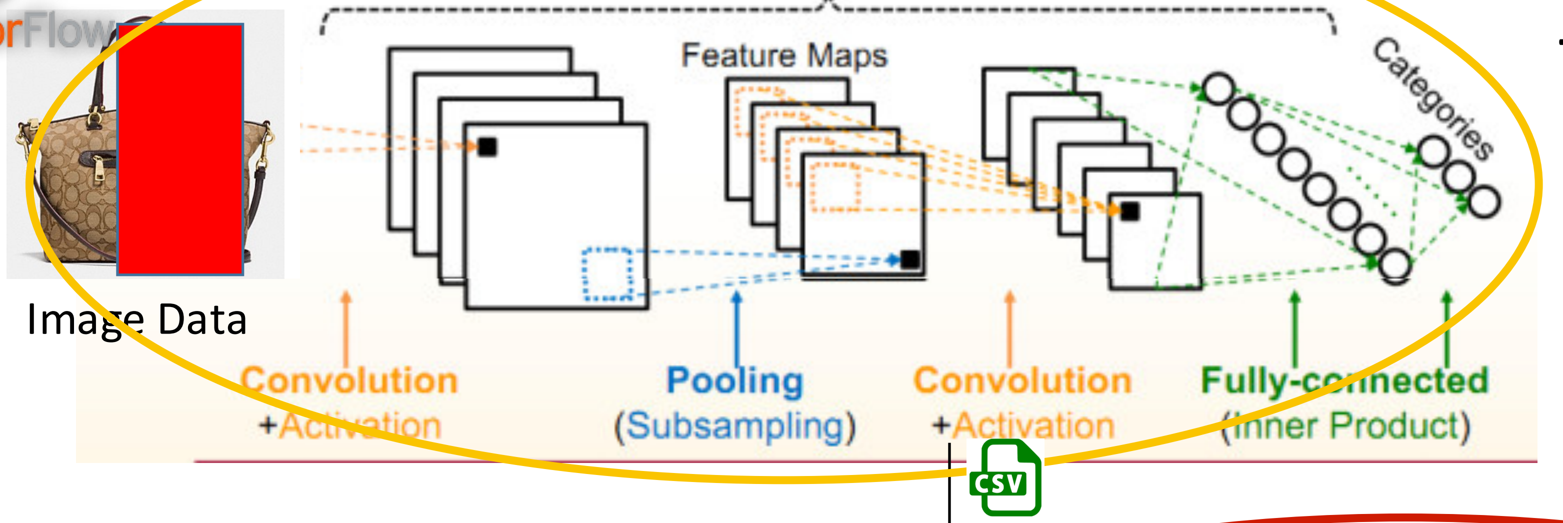




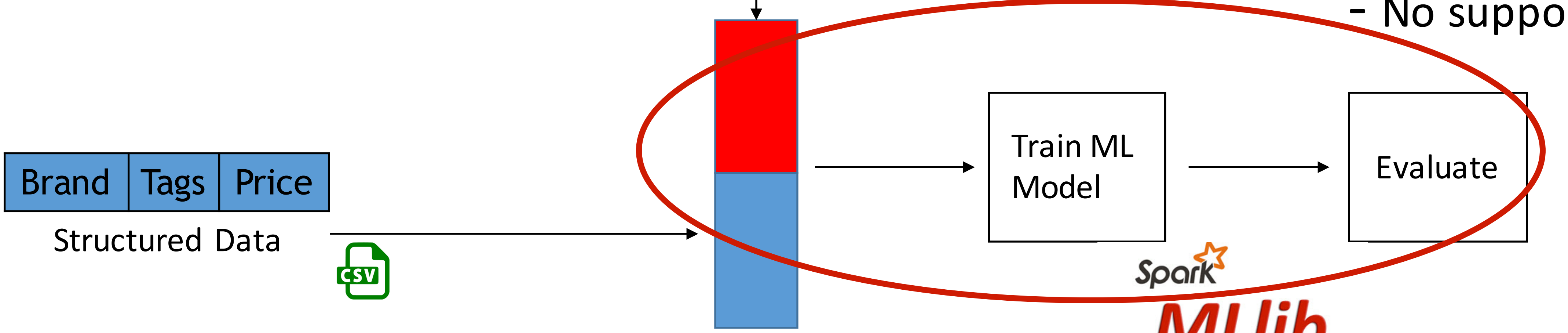
Transfer Learning: Bottleneck



Transfer Learning: Current Practice



+ Efficient CNN inference
- Doesn't support scalable processing

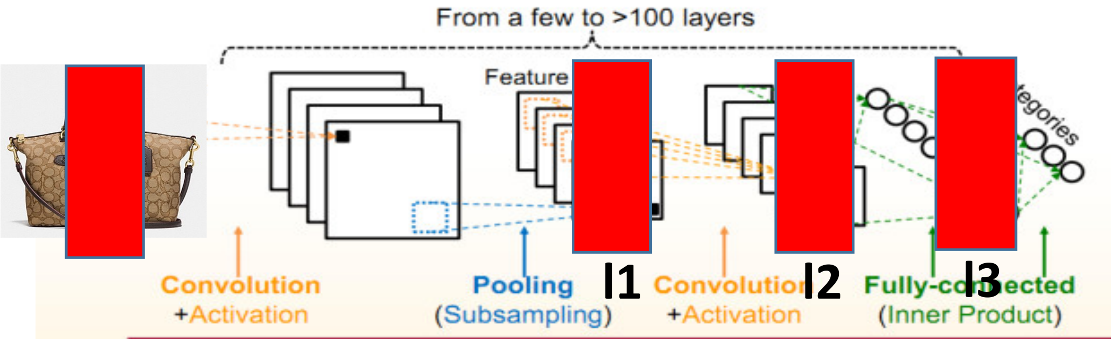


+ Scalable processing
+ Fault tolerant
- No support for CNNs

Problems with Current Practice

Usability: Manual management of CNN features.

Efficiency: From image inference for all feature layers has computational redundancies.



Reliability: CNN layers are big, requires careful memory configuration.

Disk spills

System crashes!

Outline

Example and Motivations

Our System Vista

- Overview
- System Architecture
- System Optimizations

Experimental Evaluation

Vista: Overview

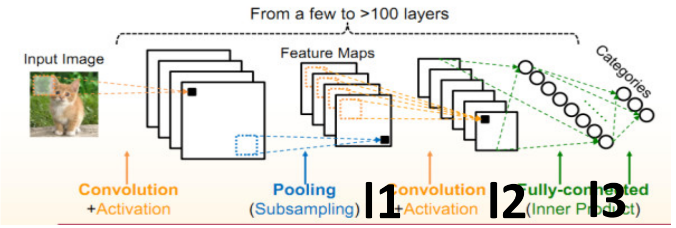
Vista is a declarative system for scalable feature transfer from deep CNNs for multimodal analytics.

Vista takes in:

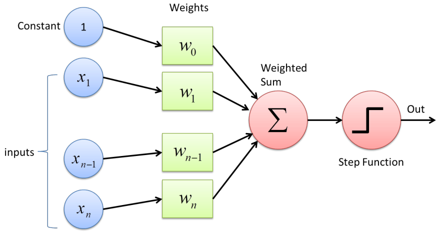
Brand	Tags	Price

Structured Data

Image Data



Pre-trained CNN and layers of interest



ML model

Vista optimizes the CNN feature transfer workload and reliably runs it.

Outline

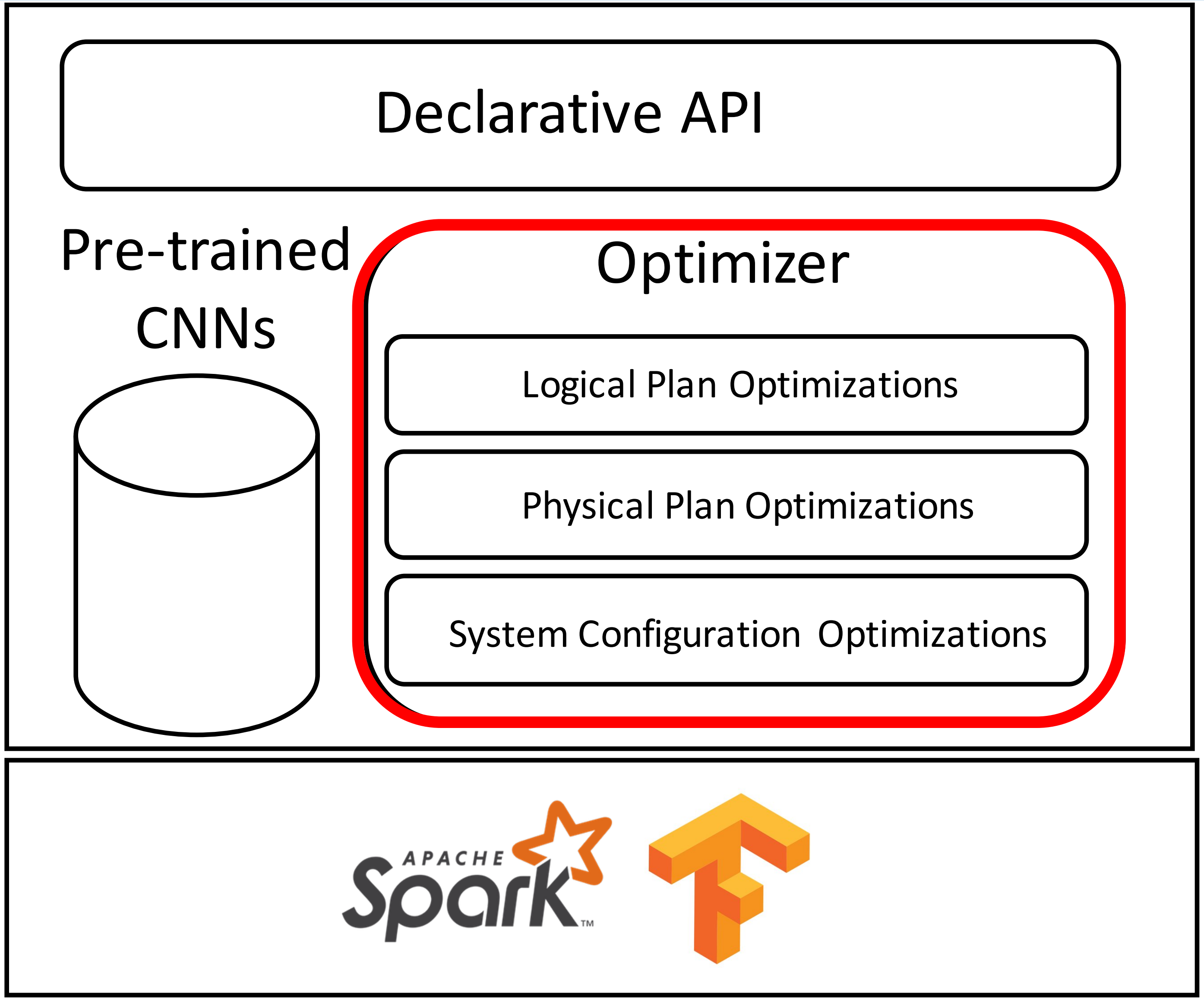
Example and Motivations

Our System Vista

- Overview
- System Architecture
- System Optimizations

Experimental Evaluation

Vista: Architecture



Benefit: Usability

Benefit: Efficiency and Reliability

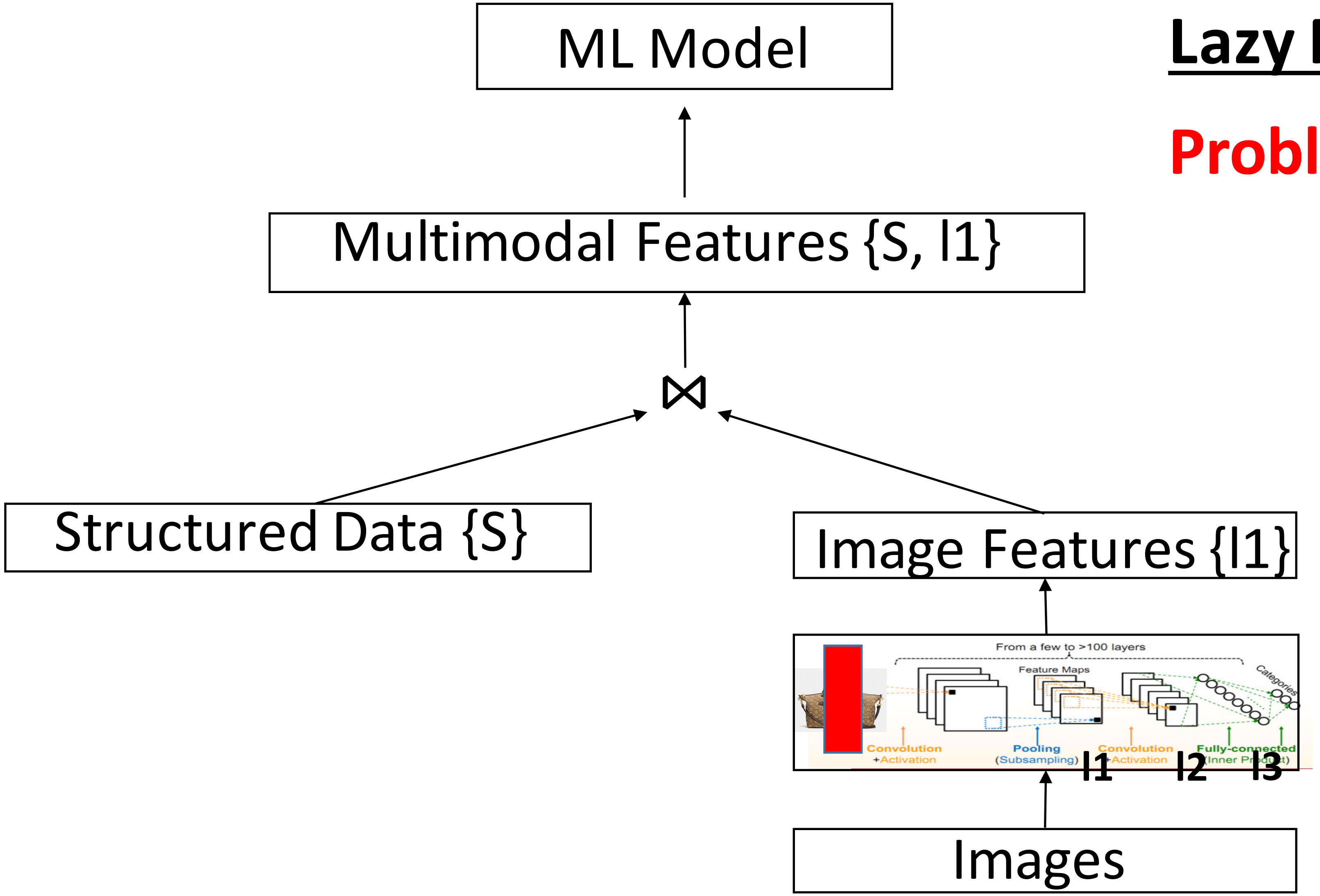
Benefit: Efficiency, Scalability, and Fault Tolerance

Outline

Our System Vista

- System Optimizations
  - Logical Plan Optimizations
  - Physical Plan Optimizations
  - System Configuration Optimizations

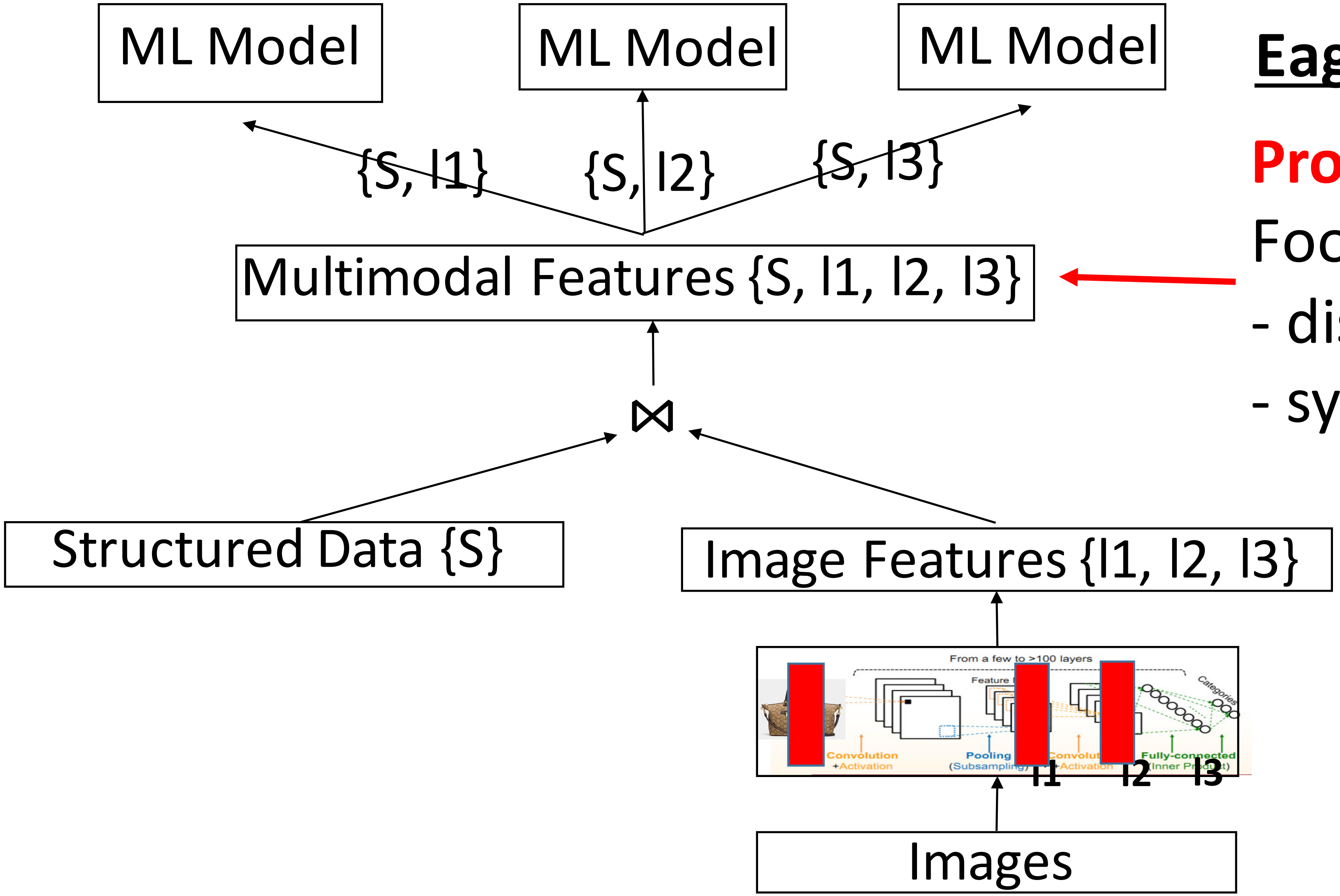
Current Practice: Repeated Inference



Lazy Materialization

Problem: Repeated inferences

Extract all layers in one go

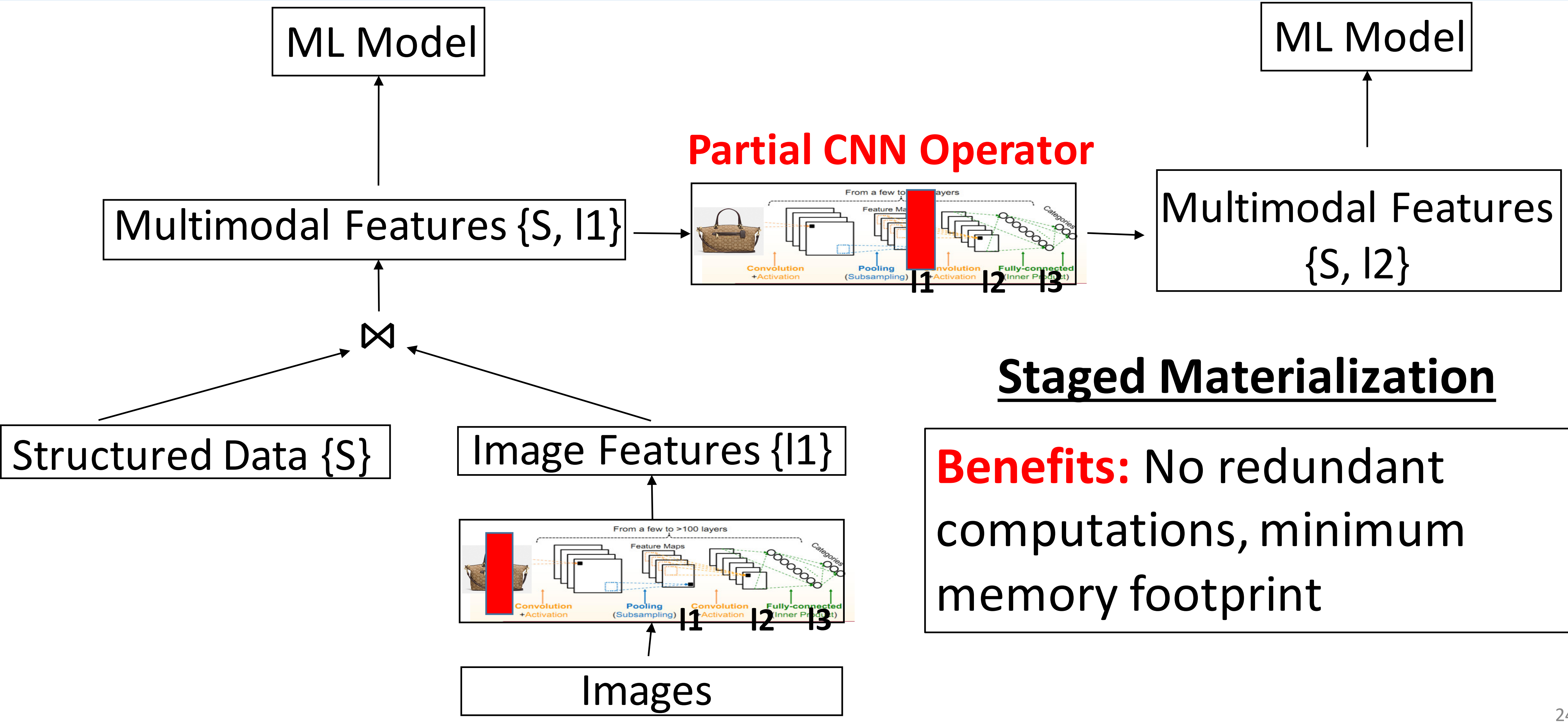


Eager Materialization

Problem: High Memory Footprint

- disk spills/cache misses
- system crashes!

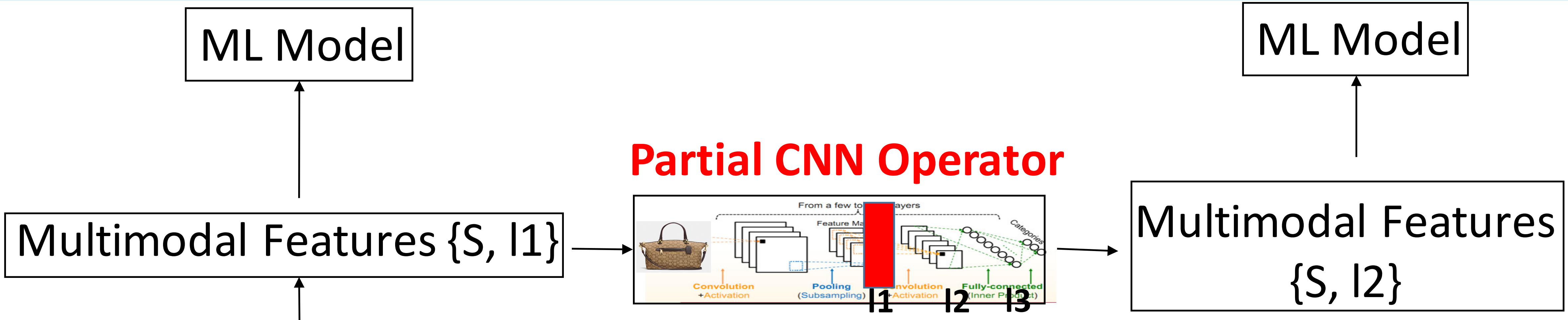
Our Novel Plan: Staged CNN Inference

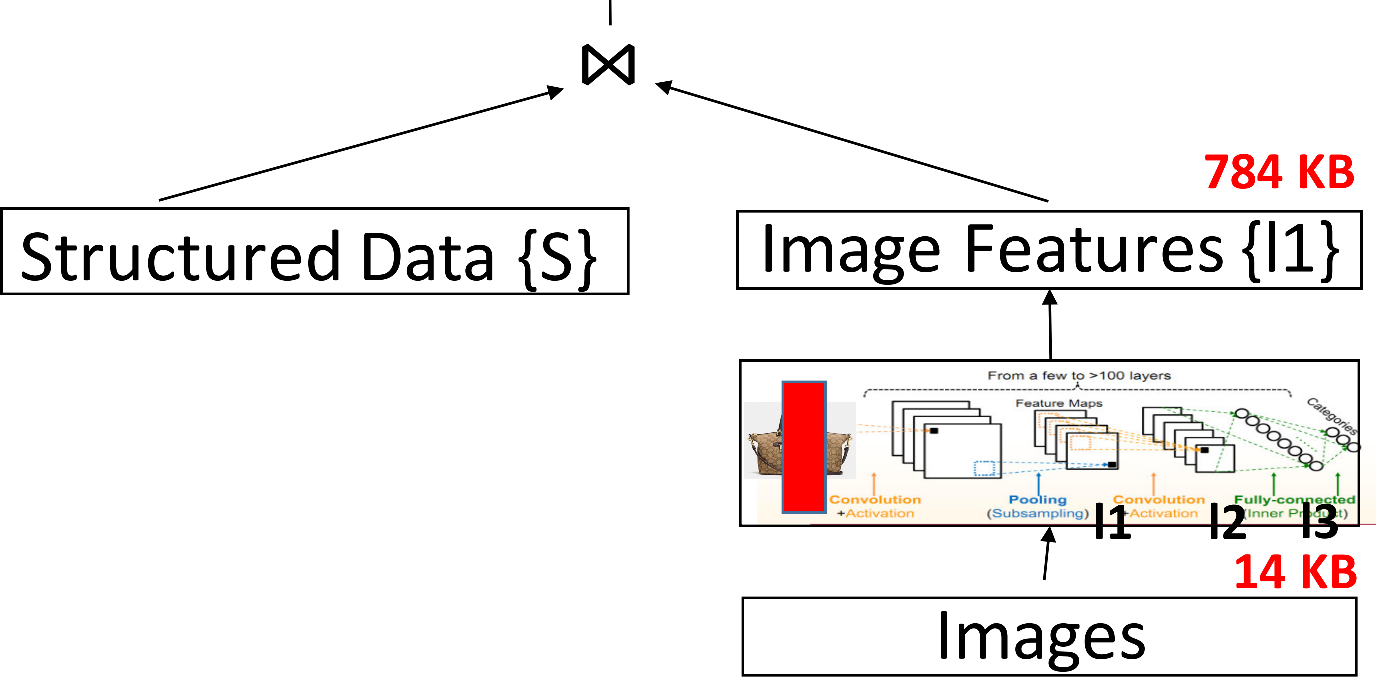


Staged Materialization

Benefits: No redundant computations, minimum memory footprint

Our Novel Plan: Staged CNN Inference





Staged Materialization

Benefits: No redundant computations, minimum memory footprint

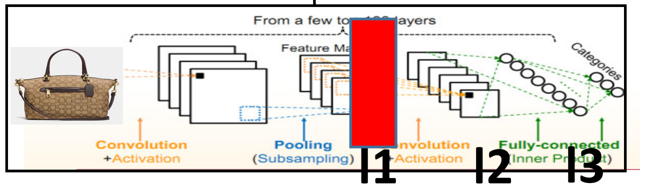
Problem: High join overhead

Our Novel Plan: Staged CNN Inference - Reordered

ML Model

Multimodal Features {S, l1}

Image Features {l1}

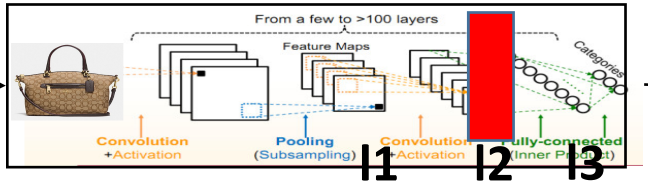


⋈

Structured Data {S}

Images

Partial CNN Operator



Multimodal Features {S, l2}

ML Model

Staged Materialization

Benefits: No redundant computations, minimum memory footprint

~~**Problem:** High join overhead~~

Outline

Our System Vista

- System Optimizations
  - Logical Plan Optimizations
  - Physical Plan Optimizations
  - System Configuration Optimizations

Vista: Physical Plan Optimizations

Join Operator

Options: Broadcast vs Shuffle join

Trade-Offs: Memory footprint vs Network cost

Storage Format

Options: Compressed vs Uncompressed

Trade-Offs: Memory footprint vs Compute cost

Benefit: Vista automatically picks the physical plan choices.

Outline

Our System Vista

- System Optimizations
  - Logical Plan Optimizations
  - Physical Plan Optimizations
  - System Configuration Optimizations

Vista: System Configuration Optimizations

Memory allocation

Query parallelism

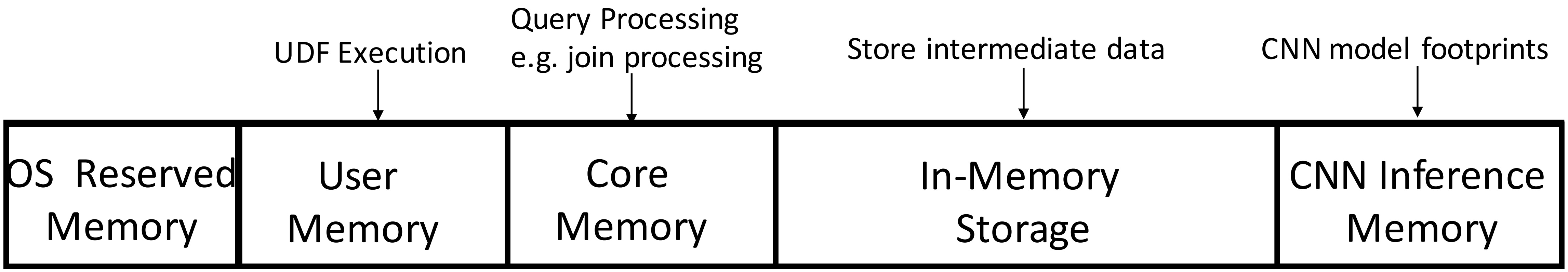
Data partition size

Memory Allocation

Challenge: Default configurations won't work

CNN features are big

Non trivial CNN model inference memory



Benefit: Vista frees the data scientist from manual memory and system configuration tuning.

Query Parallelism and Data Partition Size

Query Parallelism

Increase Query Parallelism → Increase CNN Inference Memory → Less Storage Memory

> **Benefit:** Vista sets Query Parallelism to improve utilization and reduce disk spills.

Data Partition Size

Too big → System Crash, Too small → High overhead

> **Benefit:** Vista sets optimal data partition size to reduce overheads and avoid crashes.

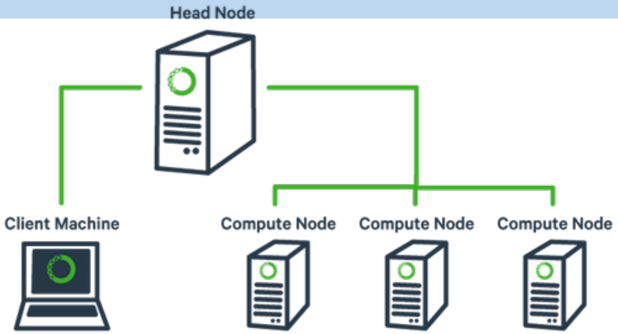
Outline

Example and Motivations

Our System Vista

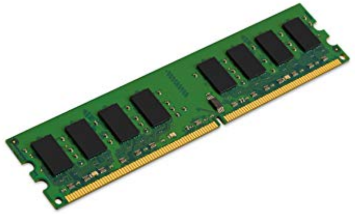
Experimental Evaluation

Experimental Setup



8 worker nodes
and 1 master node

Intel Xeon @ 2.00GHz CPU with 8 cores



32 GB RAM

300 GB HDD

Ubuntu 16.04 LTS

APACHE Spark™

Version 2.2.0
Runs in standalone mode

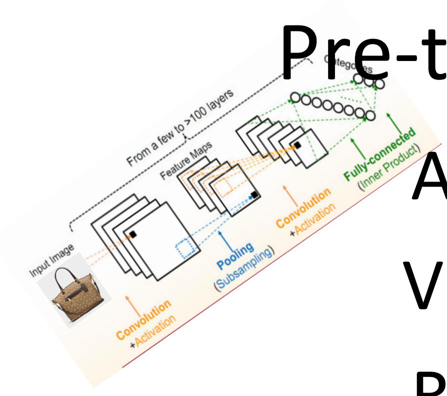
TensorFlow

Version 1.3.0

Dataset & Workloads

amazon product reviews dataset

Number of Records	200,000
Number of Structured Features	200 – price, category embedding, review embedding
Image	Image of each product item
Target	Predict each product is popular or not



Pre-trained CNNs:

AlexNet – Last 4 layers

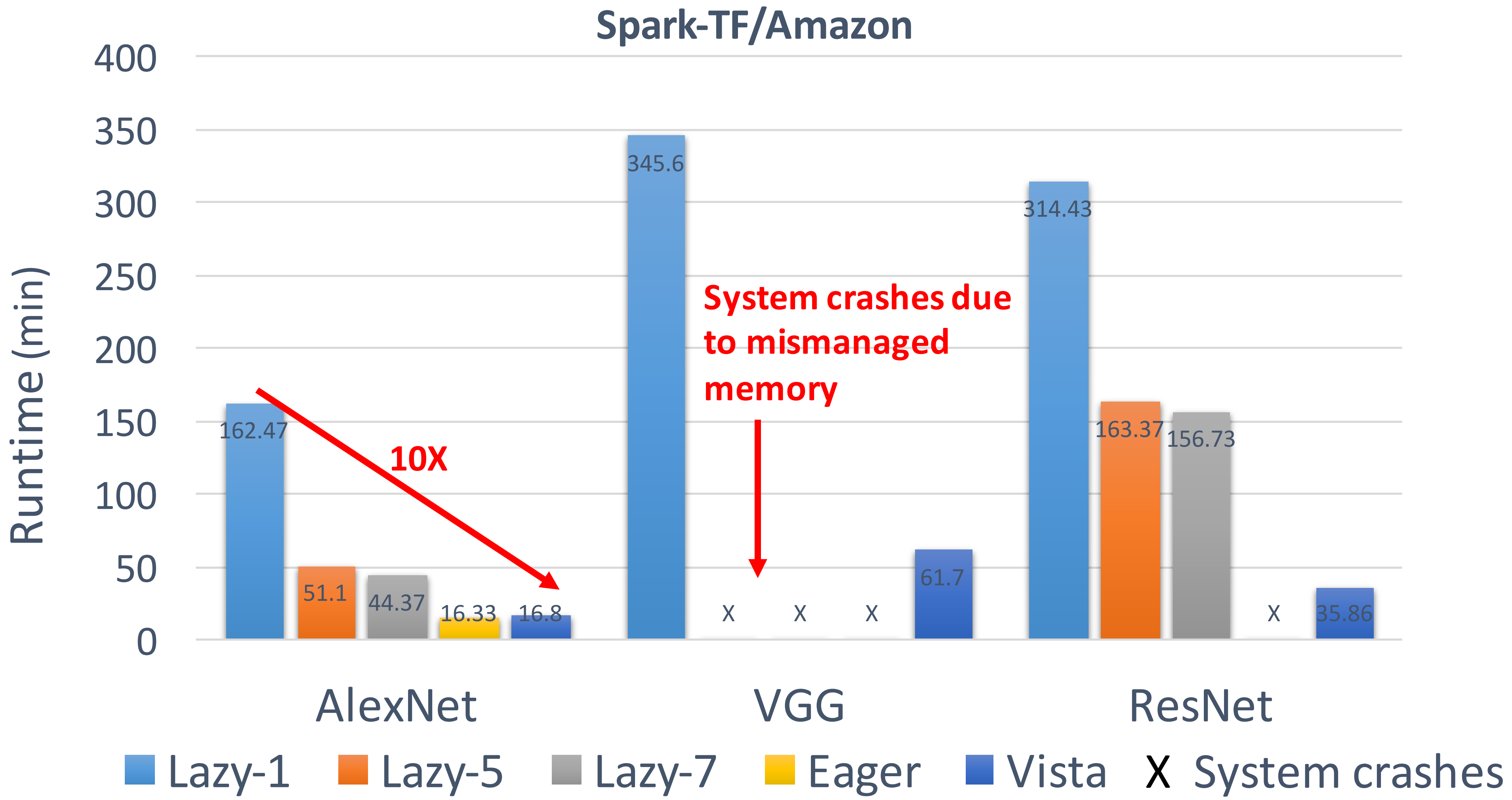
VGG16 – Last 3 layers

ResNet50 – Last 5 layers

ML model: Spark MLlib The Machine Learning Library

Logistic Regression for 10 iterations

End-to-end reliability and efficiency



Experimental results for other data systems, datasets, and drill-down experiments can be found in our paper.

Summary of Vista

Declarative system for scalable feature transfer from CNNs.

Performs DBMS inspired logical plan, physical plan, and system configuration optimizations.

Improves efficiency by up to 90% and avoids unexpected system crashes.

Thank You!

Project Webpage: https://adalabucsd.github.io/vista.html