



. arm



UNIVERSITY of  
ROCHESTER

RWTH AACHEN  
UNIVERSITY

<http://synergy.ece.gatech.edu>

# SCALE-Sim: Systolic CNN accelerator simulator

L

Tutorial @ ASPLOS 2021  
April 16 2021

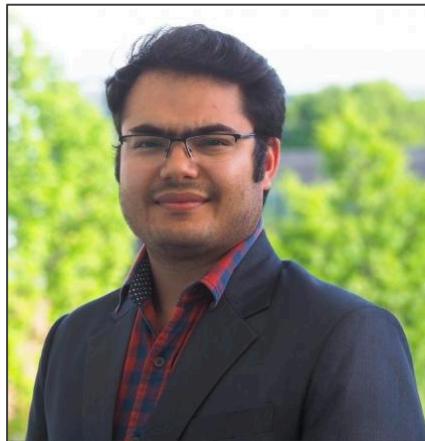
<https://scalesim-project.github.io/>  
<https://scalesim-project.github.io/tutorials-2021-asplos.html>

# Presenters



**Tushar Krishna**

*Associate Professor  
Georgia Tech*



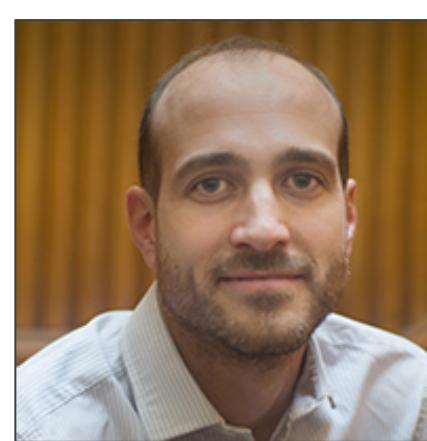
**Ananda Samakdar**

*PhD Student  
Georgia Tech*



**Jan Moritz Joseph**

*Post Doc Researcher  
RWTH Aachen University*



**Paul Whatmough**

*Principal Research Engineer  
ARM Research, Boston*



**Yuhao Zhu**

*Assistant Professor  
Univ. of Rochester*

# Schedule (EST)

Time slot	Topic	
10:05 – 10:10	Welcome and Overview	Tushar
10:10 to 10:45	Introduction to DNNs and Accelerator Design	Paul & Tushar
10:45 to 11:15	Overview of SCALE-Sim	Anand + Moritz + Paul
11:15 to 11:50	Tutorial 1: Design Space Exploration using SCALE-Sim	Anand
12:00 to 12:40	Tutorial 2: Modifying SCALE-Sim to add custom features	Moritz
12:45 to 1:30	Tutorial 3: Using SCALE-Sim to build larger simulators	Anand
1:30 to 2:00	Discussion on future roadmap, planned features, and ideas from the community	Yuhao

Brief Q/A at the end of each talk.

Attention: Tutorial is being recorded!

Slides + Videos will be available on the SCALE-sim tutorial website

<https://scalesim-project.github.io/tutorials-2021-asplos.html>

# Outline

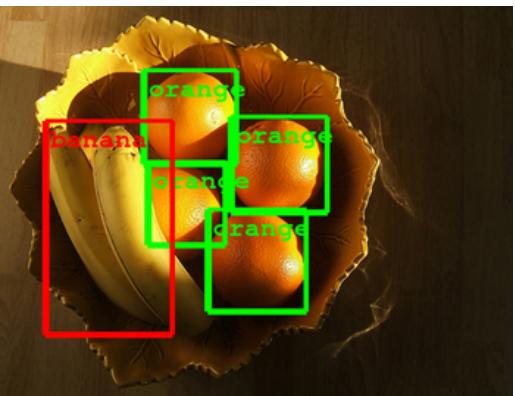
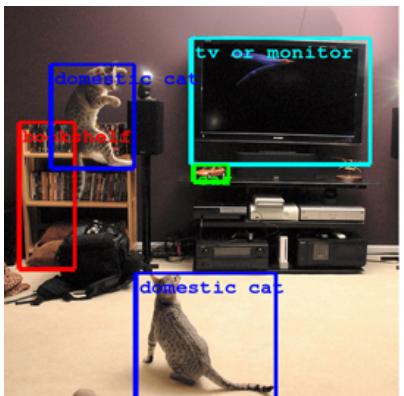
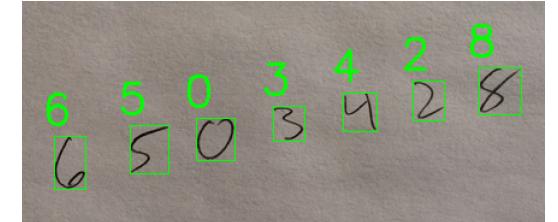
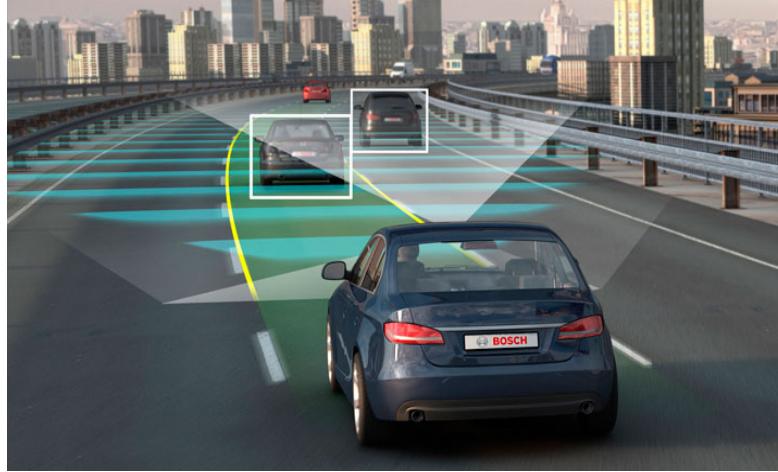
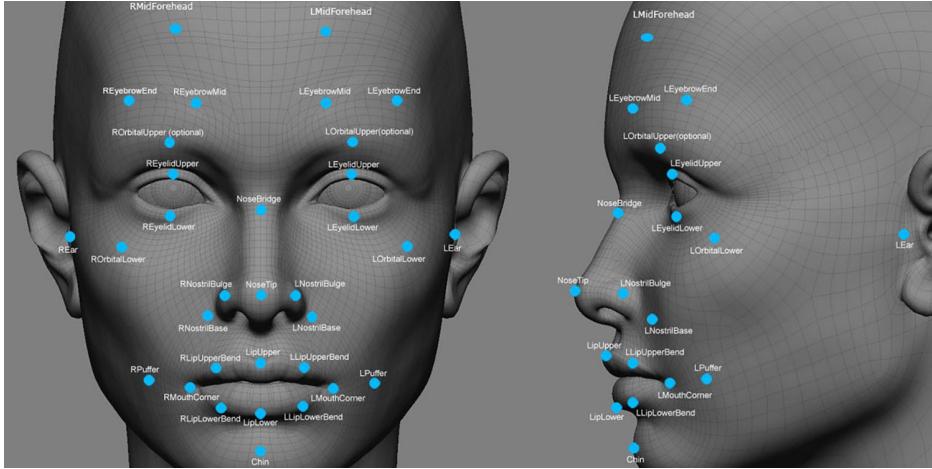
- **What this tutorial is about**
  - Role within the Deep Learning landscape
- **Resources**
  - SCALE-sim
  - Synthesis Lectures
- **Relevant Background**
  - Deep Learning Landscape
  - DNN Accelerators



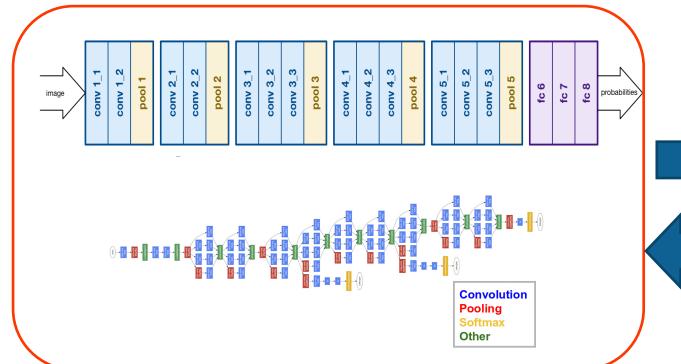
L What this tutorial is about

---

# Deep Learning Applications



# Deep Learning Landscape



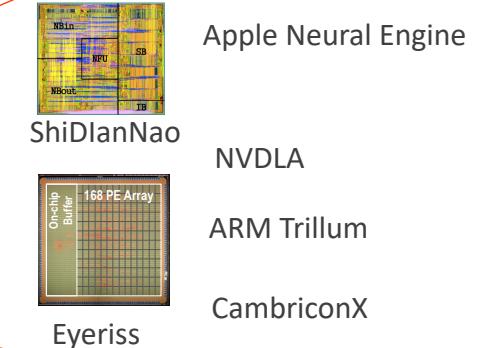
Model Creation



Training

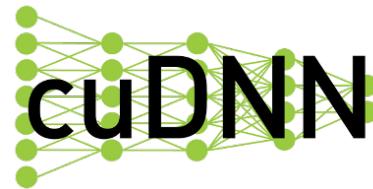


Inference

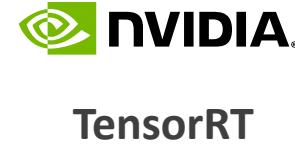


  
TensorFlow

 Caffe



Design Tools



TensorRT

SCALE-sim

This Tutorial

Resources

L

# Resources

7

# SCALE-sim Webpage, Github and Tutorial

<https://scalesim-project.github.io/>

Systolic CNN AcceLerator Simula

SCALE-Sim is a CNN accelerator simulator, that provides cy bandwidth and trace results for a specified accelerator con

Get Started Sources

**SOURCES**  
SCALE-Sim sources are available from Github  
[Go to Github](#)

**GETTING STARTED**  
Learn, how to use SCALE-Sim in 30 seconds!  
[SCALE-Sim in 30 sec.](#)

<https://scalesim-project.github.io/tutorials-2021-asplos.html>

<https://github.com/scalesim-project/scale-sim-v2>

**SCALE-SIM TUTORIAL - ASPLOS 2021**

**README.md**

**Systolic CNN AcceLerator Simulator (SCALE Sim) v2**

SCALE Sim is a simulator for systolic array based accelerators for Convolution, Feed Forward, and any layer that uses GEMMs. This is a refreshed version of the simulator with feature enhancements, restructured code to aid feature additions, and ease of distribution.

The previous version of the simulator can be found [here](#).

**Getting started in 30 seconds**

**Installing the package**

Getting started is simple! SCALE-Sim is completely written in python and is available both as a package and could be run from source.

You can install SCALE-Sim in your environment using the following command

```
$ pip3 install scalesim
```

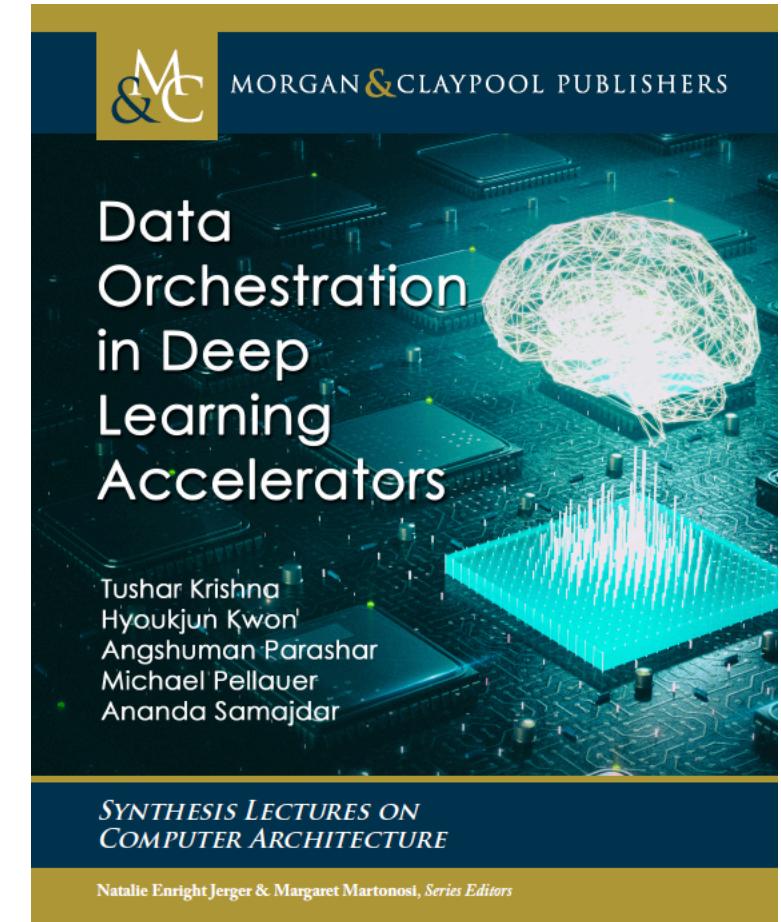
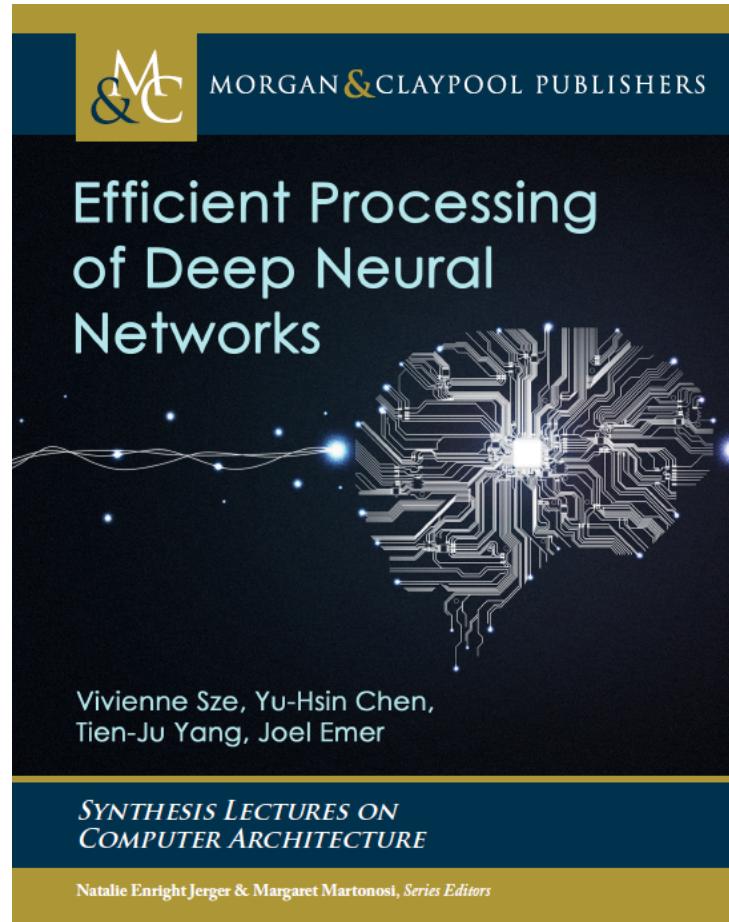
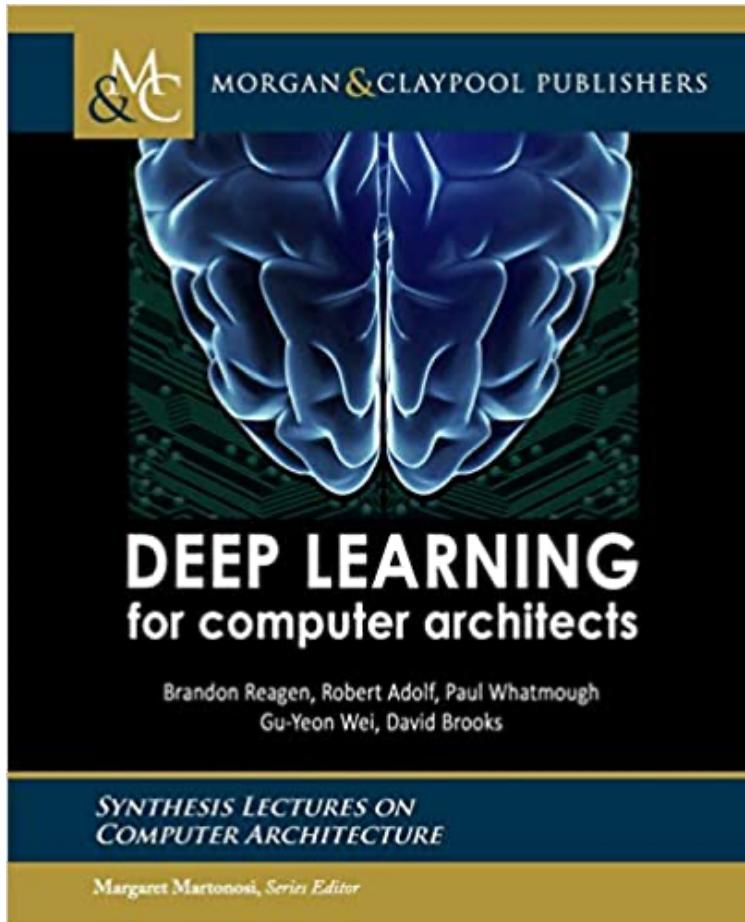
Alternatively you can install the package from the source as well

```
$ python3 setup.py install
```

**Launching a run**

<https://github.com/scalesim-project/scale-sim-v2>

# Synthesis Lectures





# Welcome to the Tutorial!

---

# Schedule (EST)

Time slot	Topic	
10:05 – 10:10	Welcome and Overview	Tushar
10:10 to 10:45	Introduction to DNNs and Accelerator Design	Paul & Tushar
10:45 to 11:15	Overview of SCALE-Sim	Anand + Moritz + Paul
11:15 to 11:50	Tutorial 1: Design Space Exploration using SCALE-Sim	Anand
12:00 to 12:40	Tutorial 2: Modifying SCALE-Sim to add custom features	Moritz
12:45 to 1:30	Tutorial 3: Using SCALE-Sim to build larger simulators	Anand
1:30 to 2:00	Discussion on future roadmap, planned features, and ideas from the community	Yuhao

Brief Q/A at the end of each talk.

Attention: Tutorial is being recorded!

Slides + Videos will be available on the SCALE-sim tutorial website

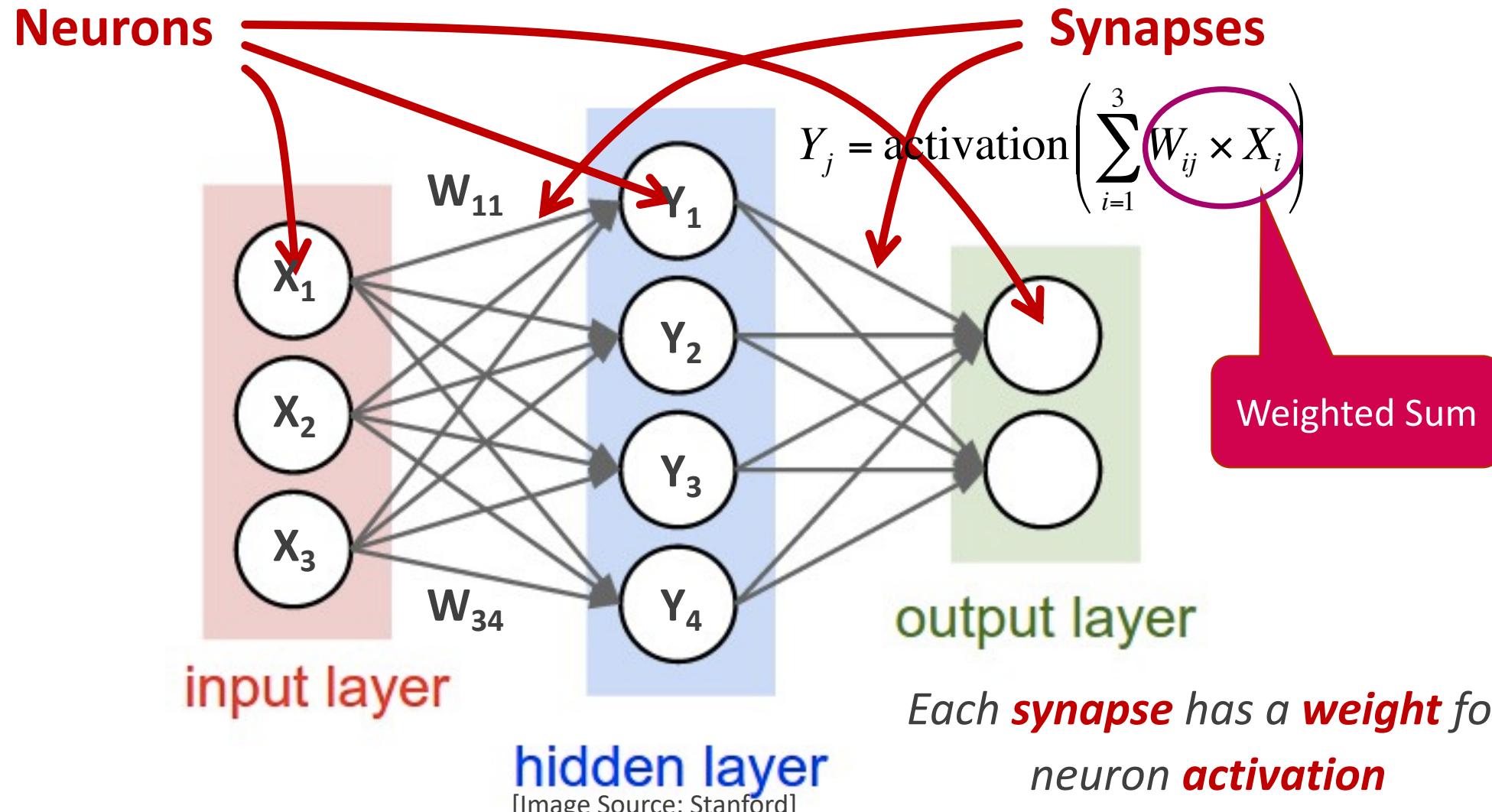
<https://scalesim-project.github.io/tutorials-2021-asplos.html>

L

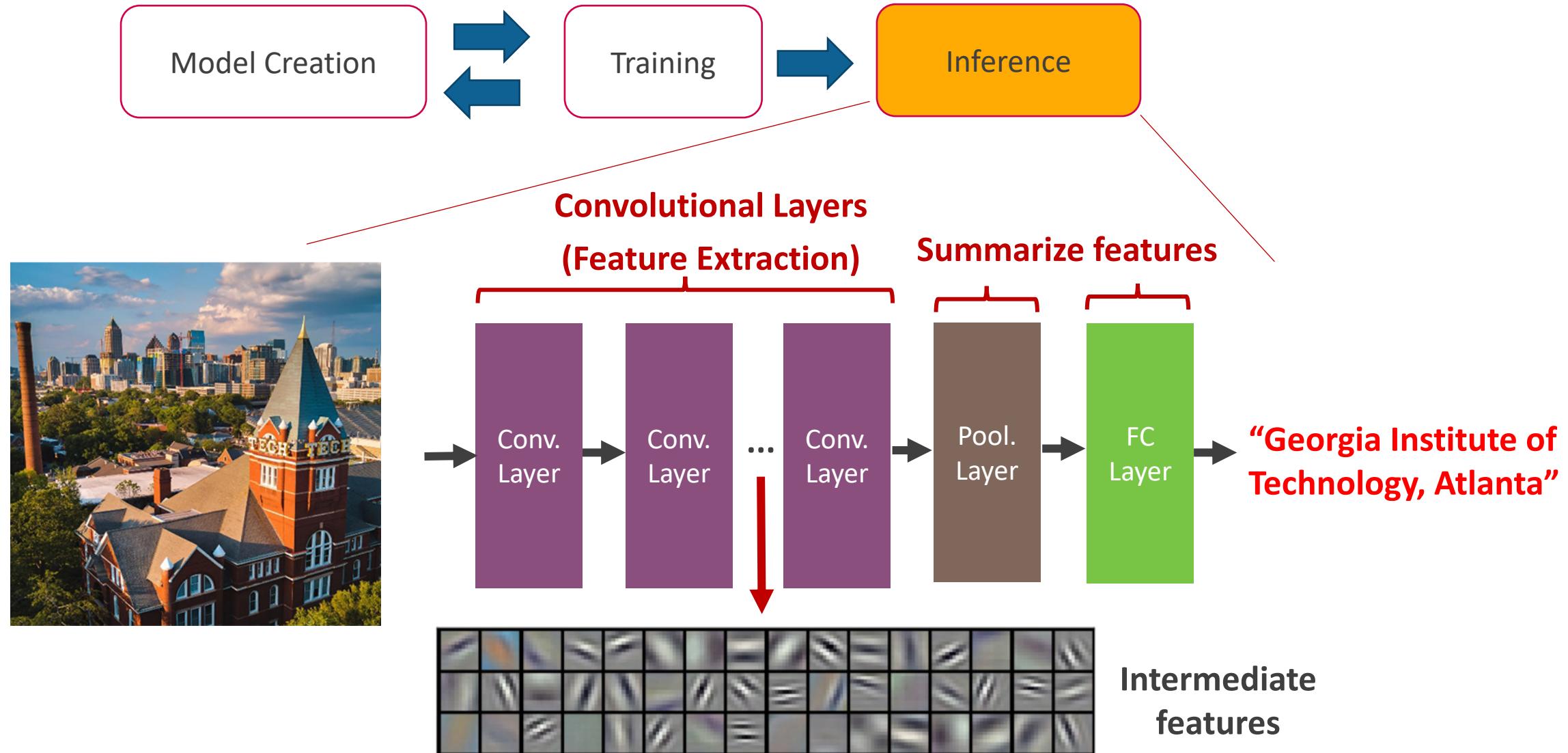
# Relevant Background

7

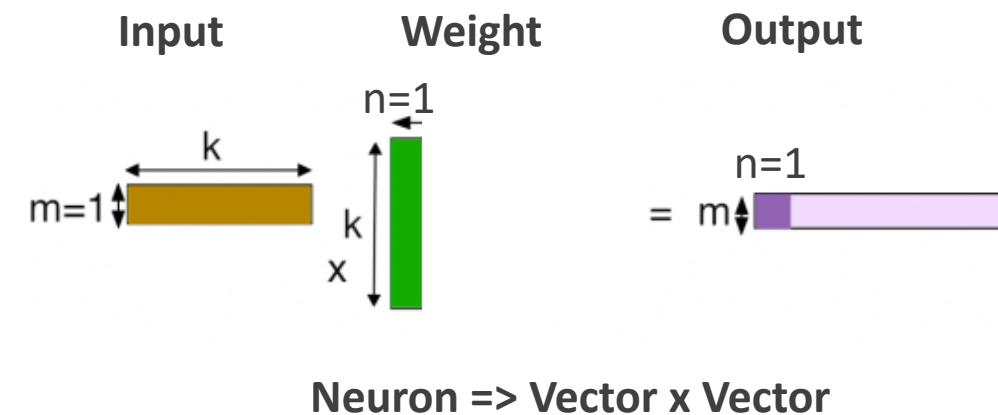
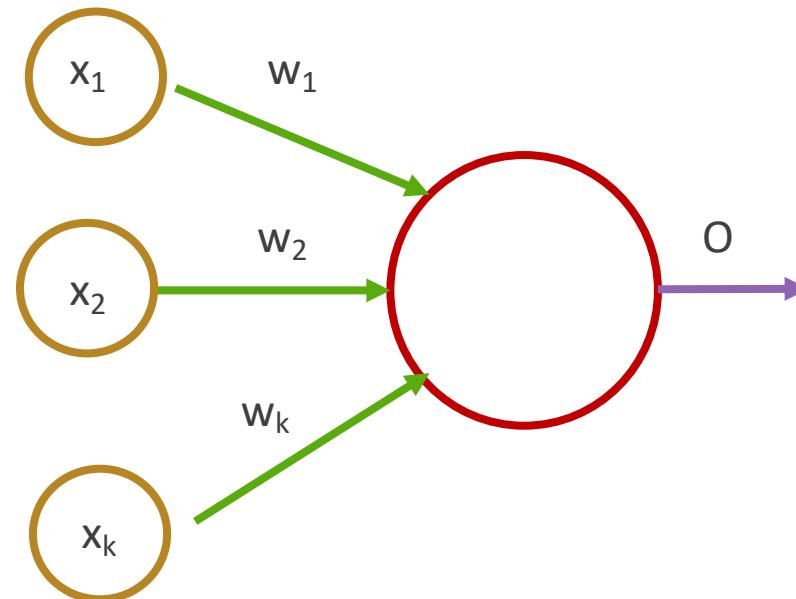
# What is a Deep Neural Network?



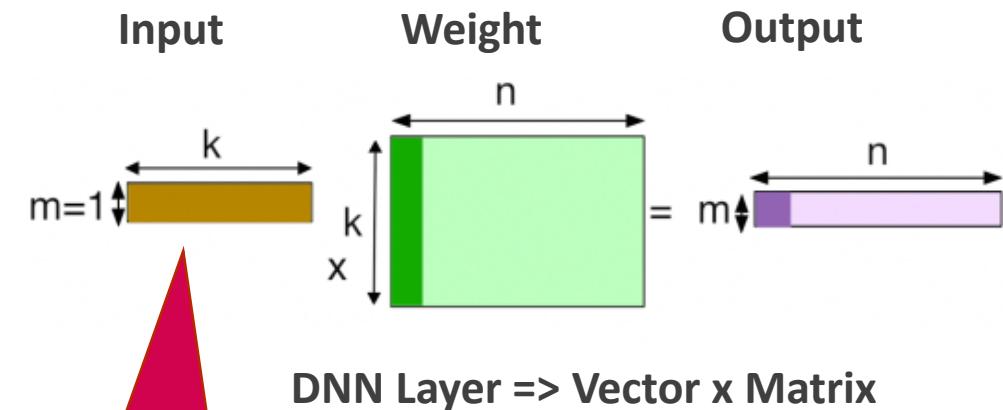
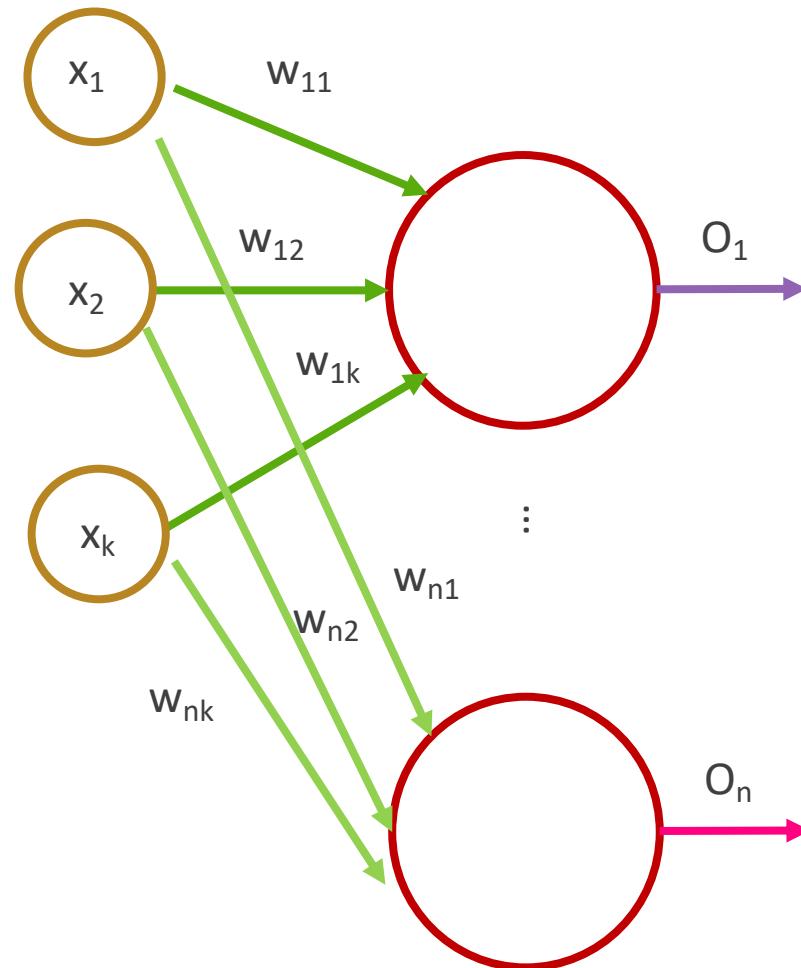
# Modern Deep Learning Landscape



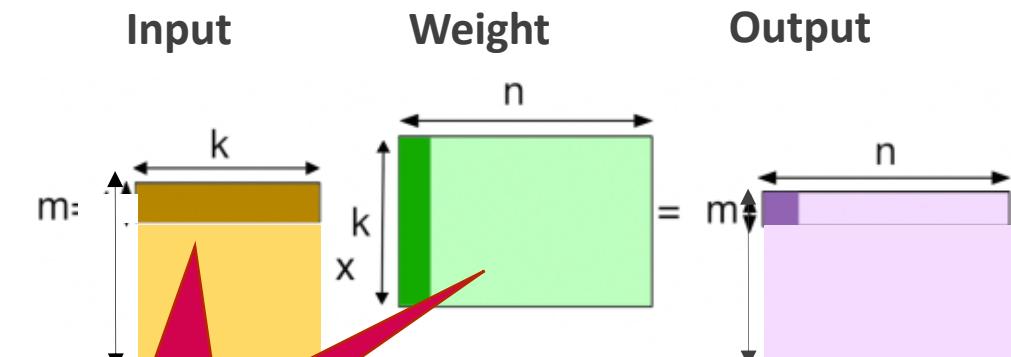
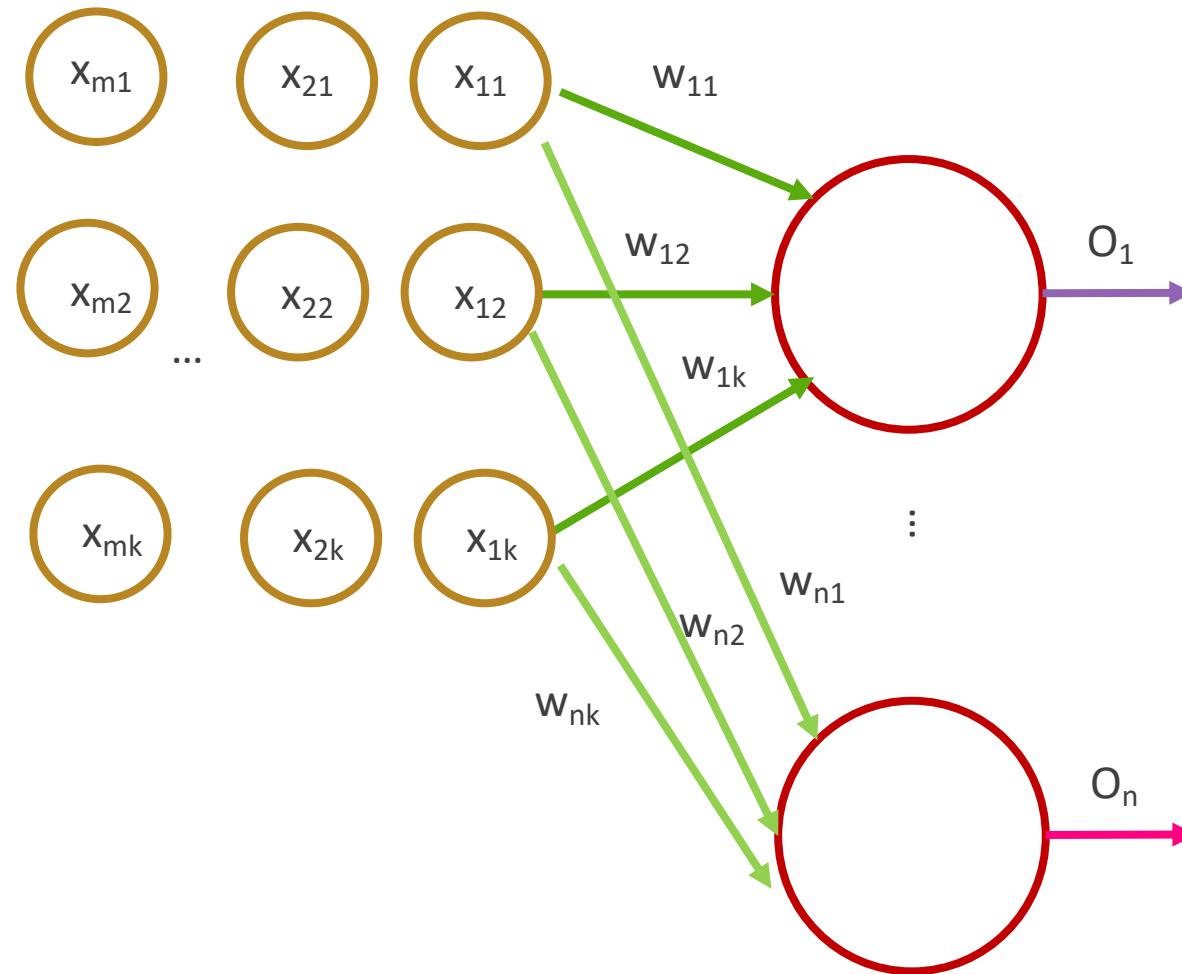
# Computations in a DNN → Linear Algebra



# Computations in a DNN → Linear Algebra



# Computations in a DNN → Linear Algebra

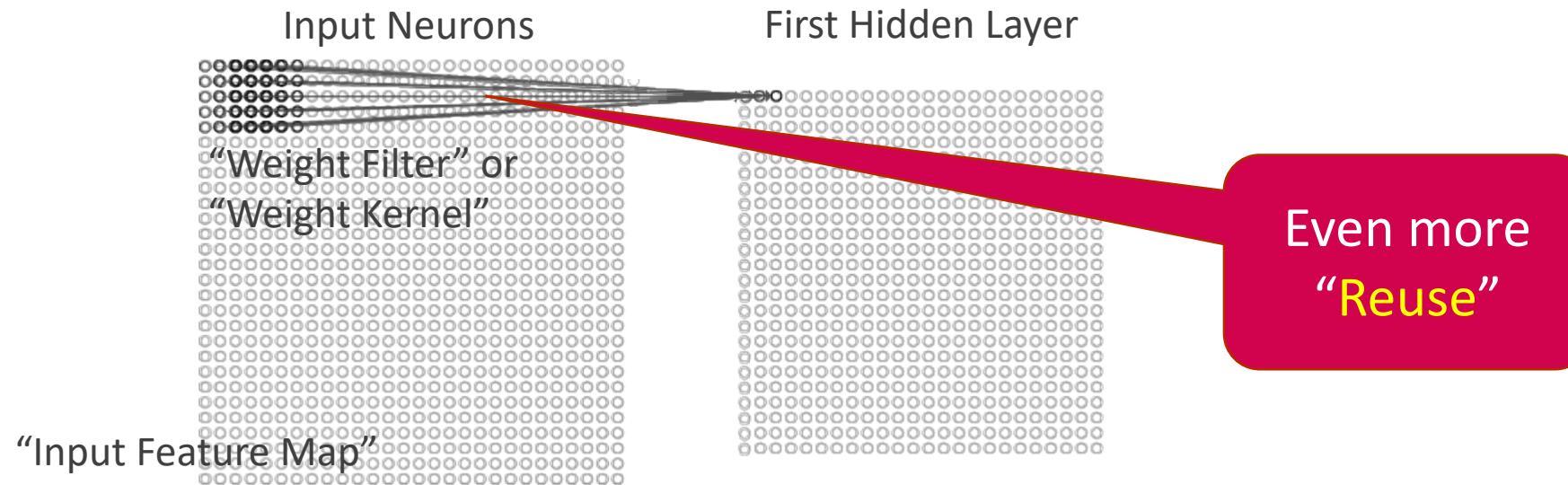


Data “Reuse”

Batching => Matrix x Matrix

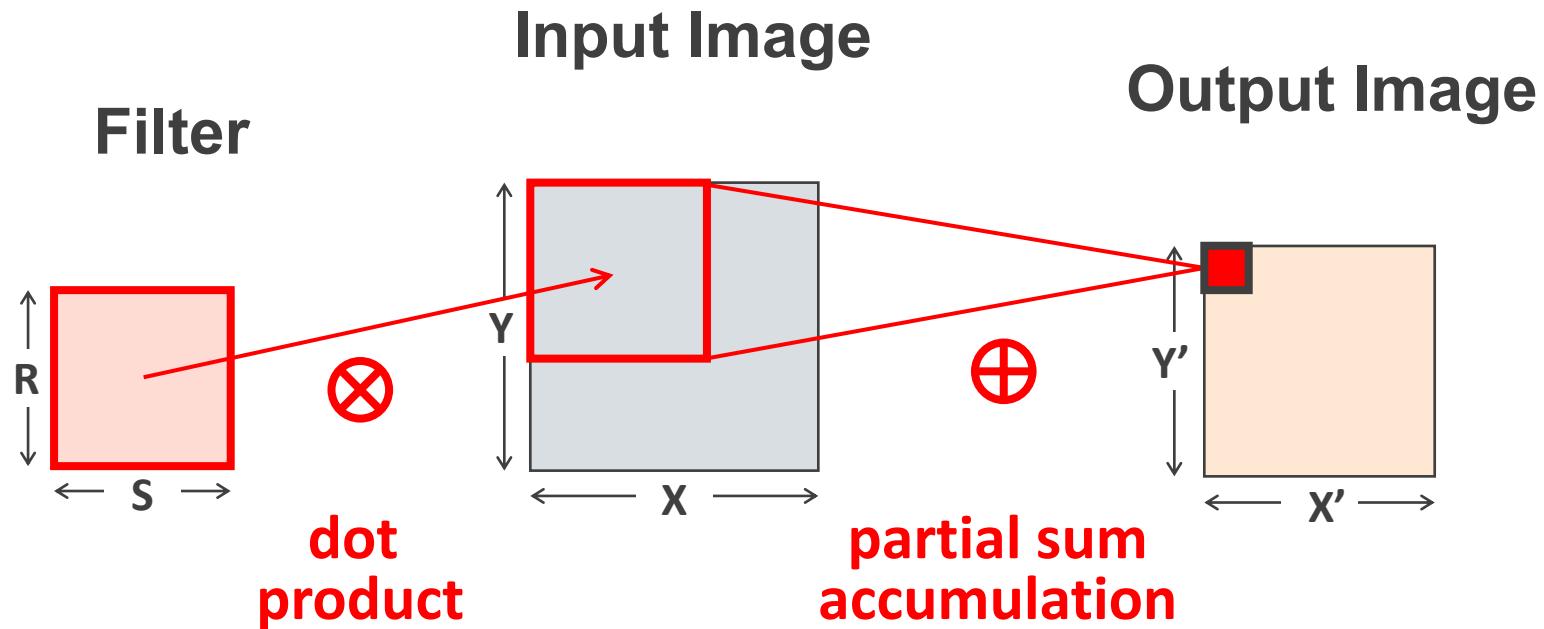
GEMM

# Convolutional Neural Networks

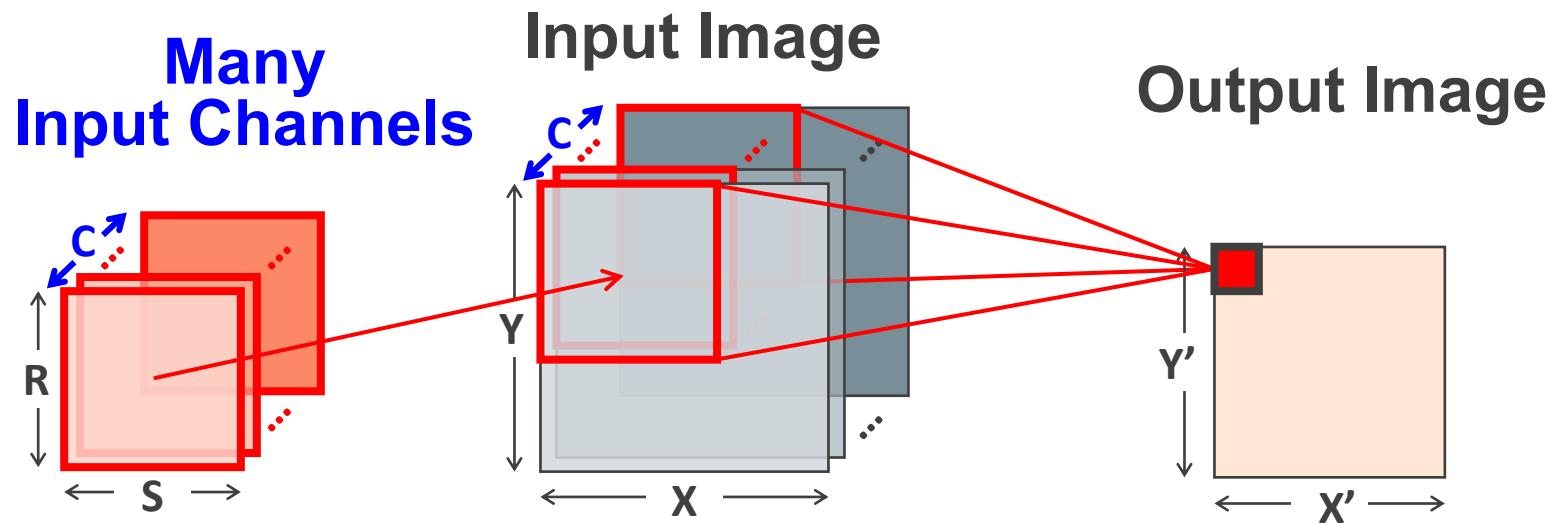


**Shared Weights:**  
All neurons use the *same* filter weights

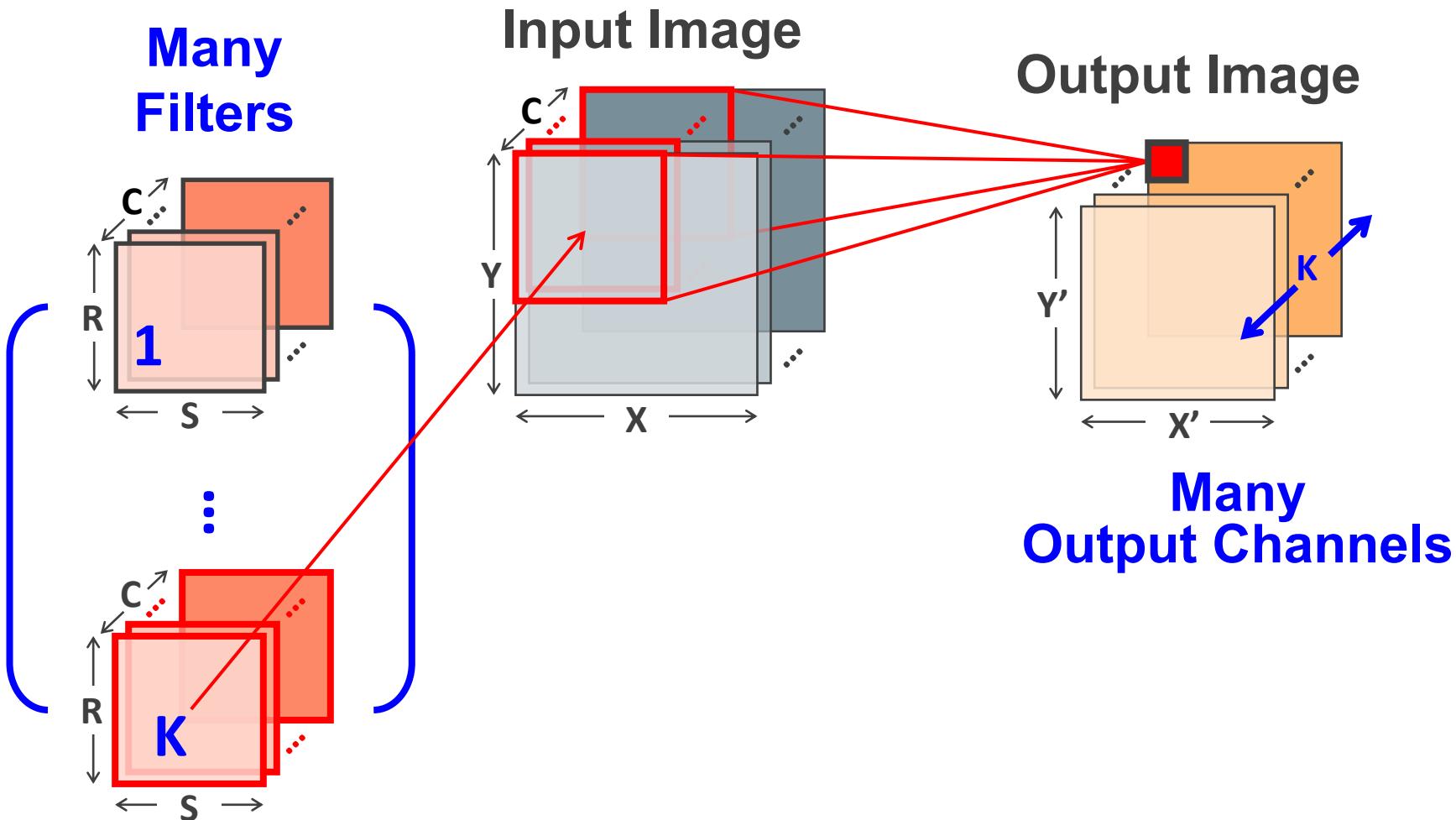
# Convolution in CNN



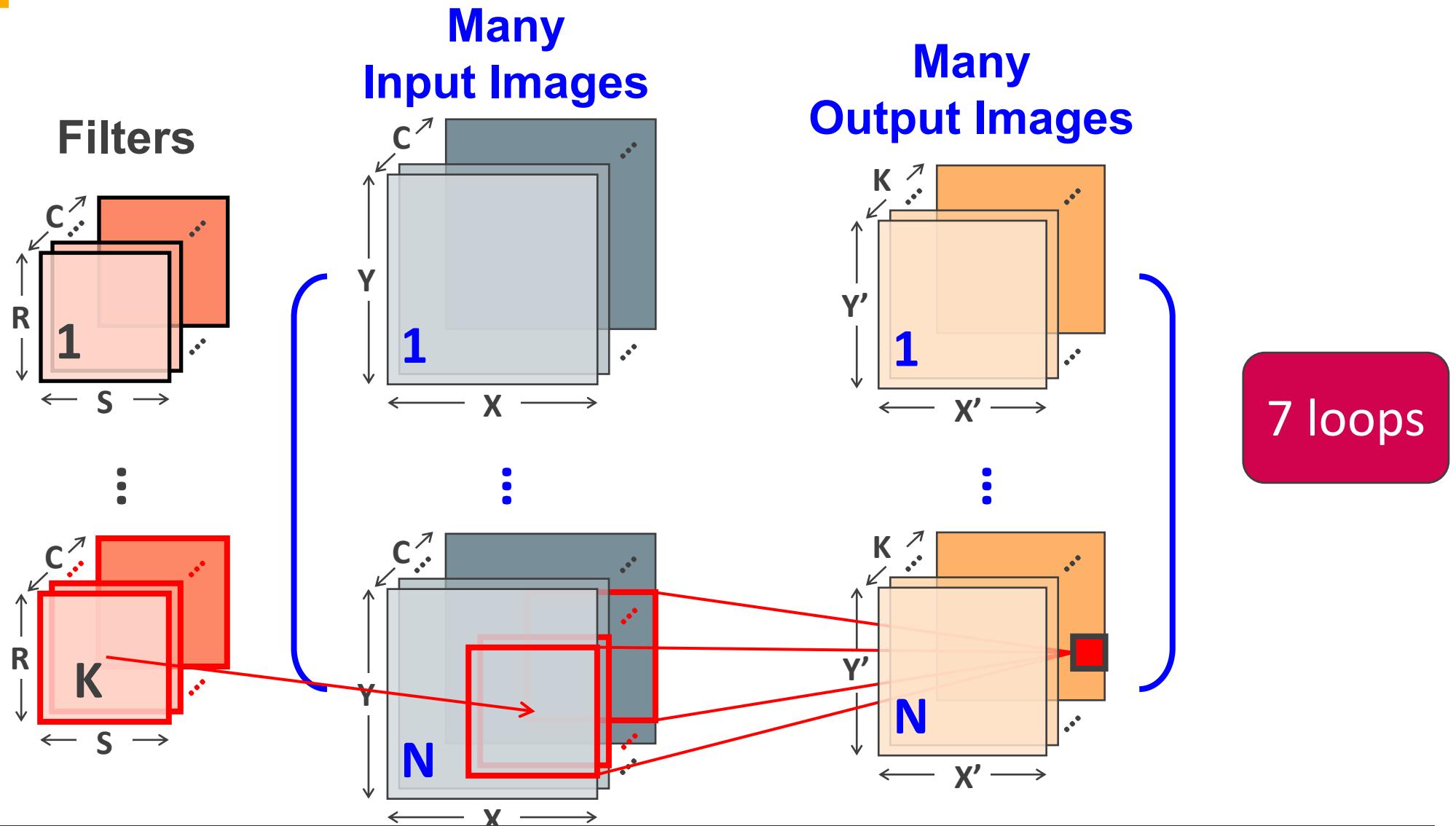
# Convolution in CNN



# Convolution in CNN



# Convolution in CNN



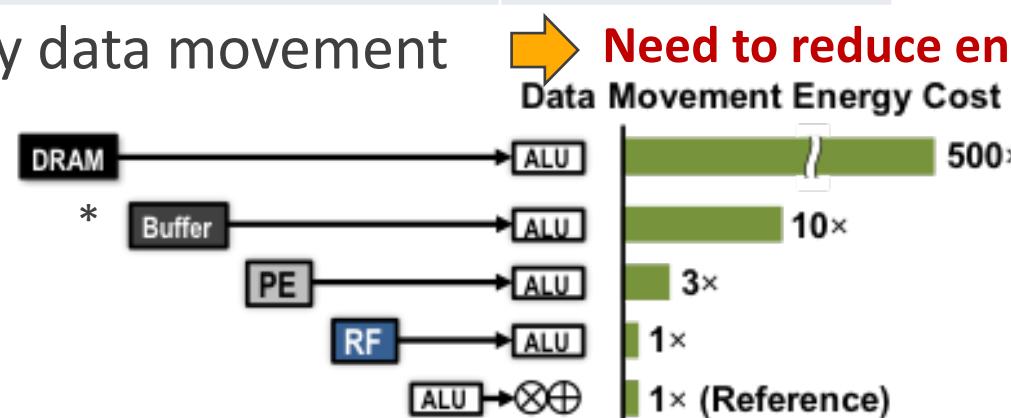
# Challenges with DNN Computations

- Millions of Parameters (i.e., weights)
  - Billions of computations → **Need lots of parallel computations**

DNN Topology	Number of Weights
AlexNet (2012)	3.98M
VGGnet-16 (2014)	28.25M
GoogleNet (2015)	6.77M
Resnet-50 (2016)	23M
DLRM (2019)	540M
Megatron (2019)	8.3B

This makes CPUs inefficient

- Heavy data movement → **Need to reduce energy**



This makes GPUs inefficient

# Spatial (or Dataflow) Accelerators

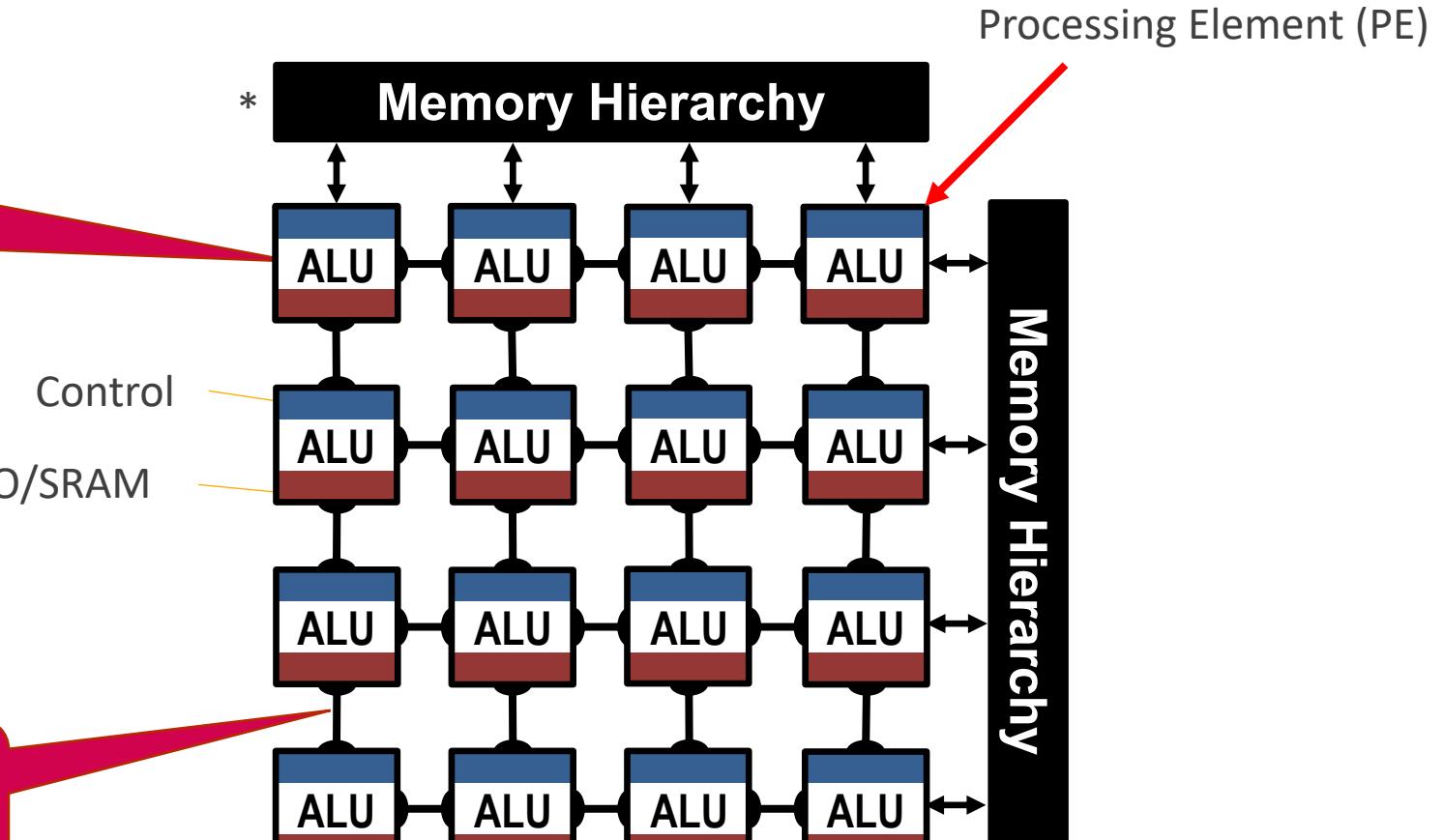
- Millions of Parameters (i.e., weights)

- Billions of computations

Spread computations across hundreds of ALUs

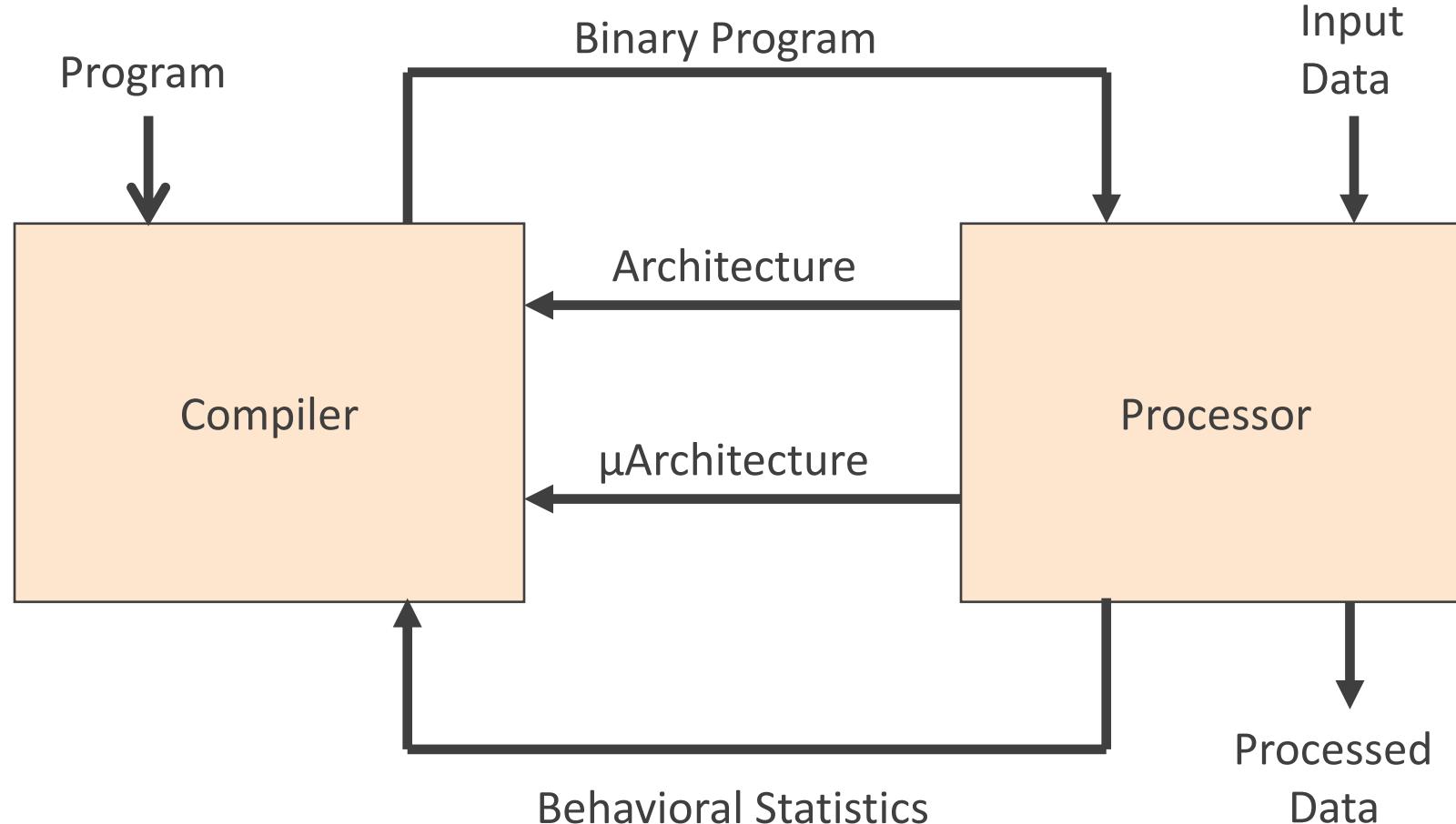
- Heavy data movement

Reuse data within the array via local memories and direct communication

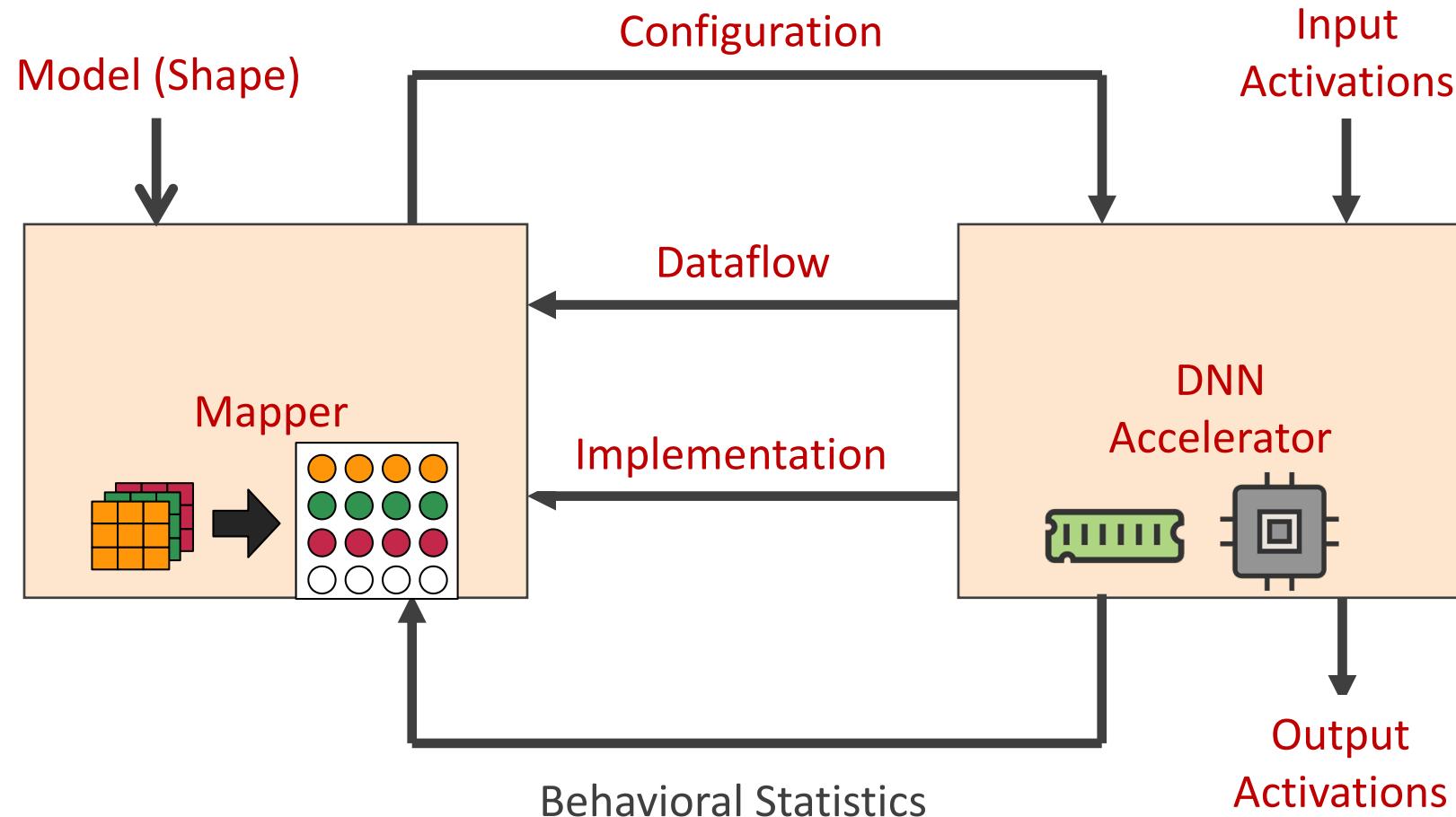


*Examples: MIT Eyeriss, Google TPU, ...*

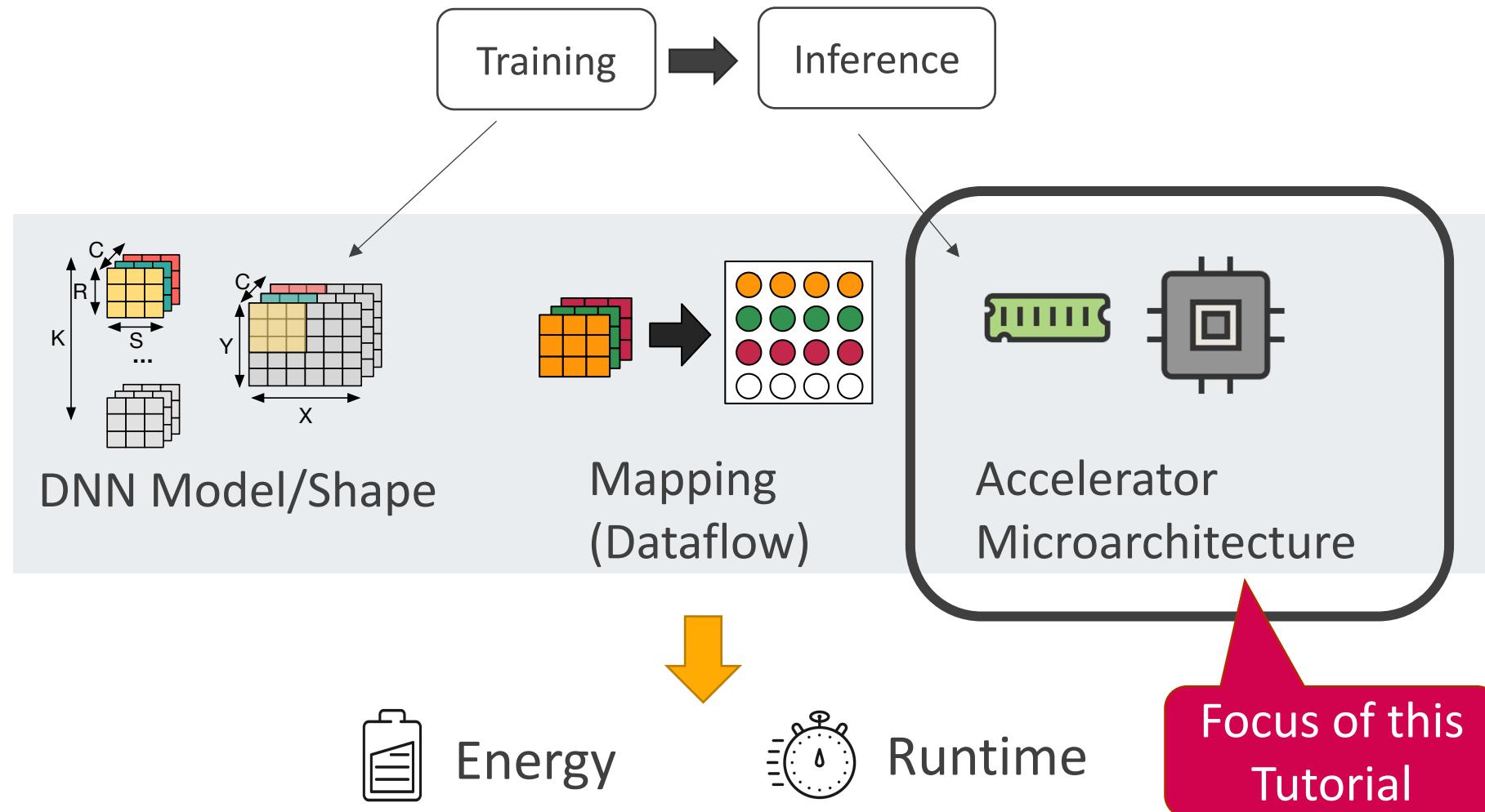
# CPU Compute Model



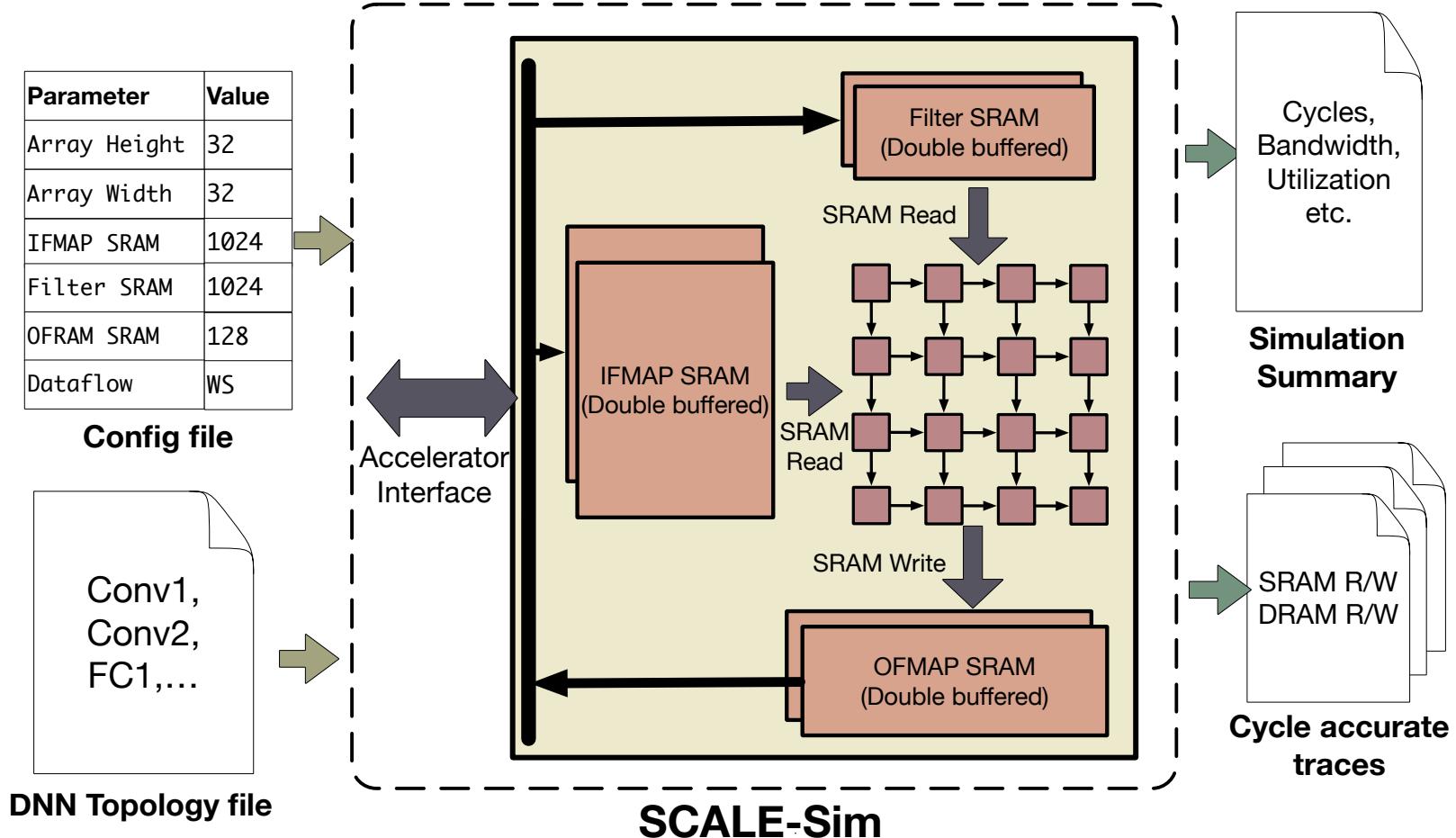
# DNN Compute Model



# Challenges in Design and Deployment



# SCALE-sim



*A Systematic Methodology for Characterizing Scalability of DNN Accelerators using SCALE-sim*  
 Ananda Samajdar, Jan Moritz Joseph, Yuhao Zhu, Paul Whatmough, Matthew Mattina, and Tushar Krishna, ISPASS 2020

# Schedule (EST)

Time slot	Topic	
10:05 – 10:10	Welcome and Overview	Tushar
10:10 to 10:45	Introduction to DNNs and Accelerator Design	Paul & Tushar
10:45 to 11:15	Overview of SCALE-Sim	Anand + Moritz + Paul
11:15 to 11:50	Tutorial 1: Design Space Exploration using SCALE-Sim	Anand
12:00 to 12:40	Tutorial 2: Modifying SCALE-Sim to add custom features	Moritz
12:45 to 1:30	Tutorial 3: Using SCALE-Sim to build larger simulators	Anand
1:30 to 2:00	Discussion on future roadmap, planned features, and ideas from the community	Yuhao

Brief Q/A at the end of each talk.

Attention: Tutorial is being recorded!

Slides + Videos will be available on the SCALE-sim tutorial website

<https://scalesim-project.github.io/tutorials-2021-asplos.html>