AccML 2020: HiPEAC Accelerated Machine Learning

Machine Learning At Scale: *Heterogeneity* & *Scalability* Challenges for ML Systems

Carole-Jean Wu

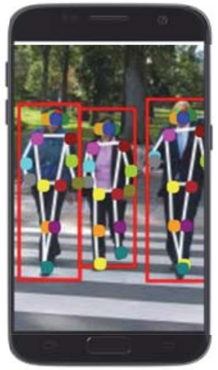
Facebook AI Research – SysML

Machine Learning at Facebook's Scale

- Machine learning is used extensively
  - Ranking posts in Newsfeed
  - Content understanding
  - Object detection, segmentation, and tracking
  - Speech recognition/translation

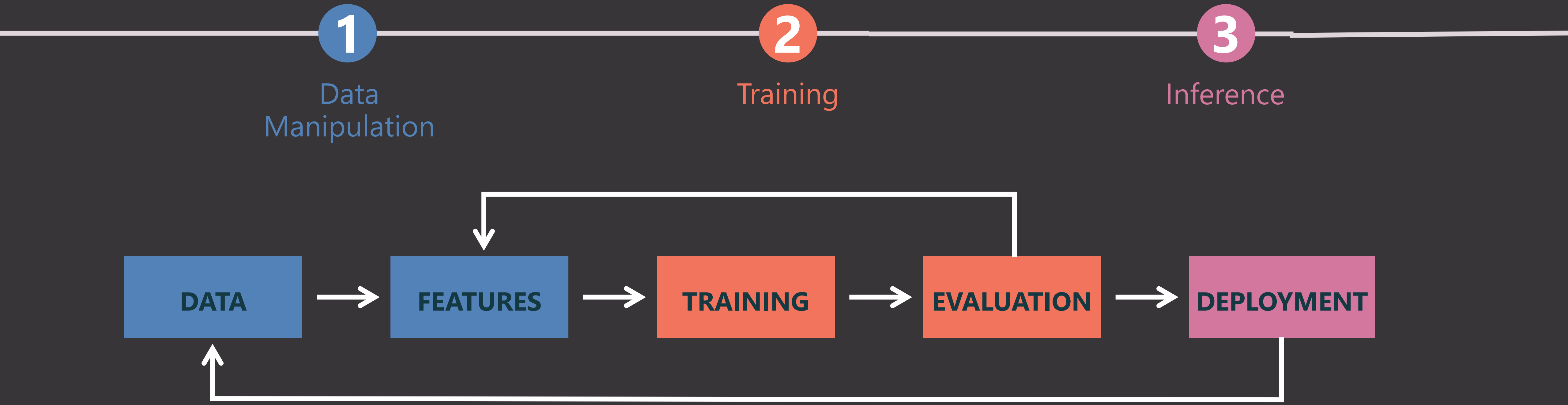
- From data centers to the edge

Keypoints Segmentation

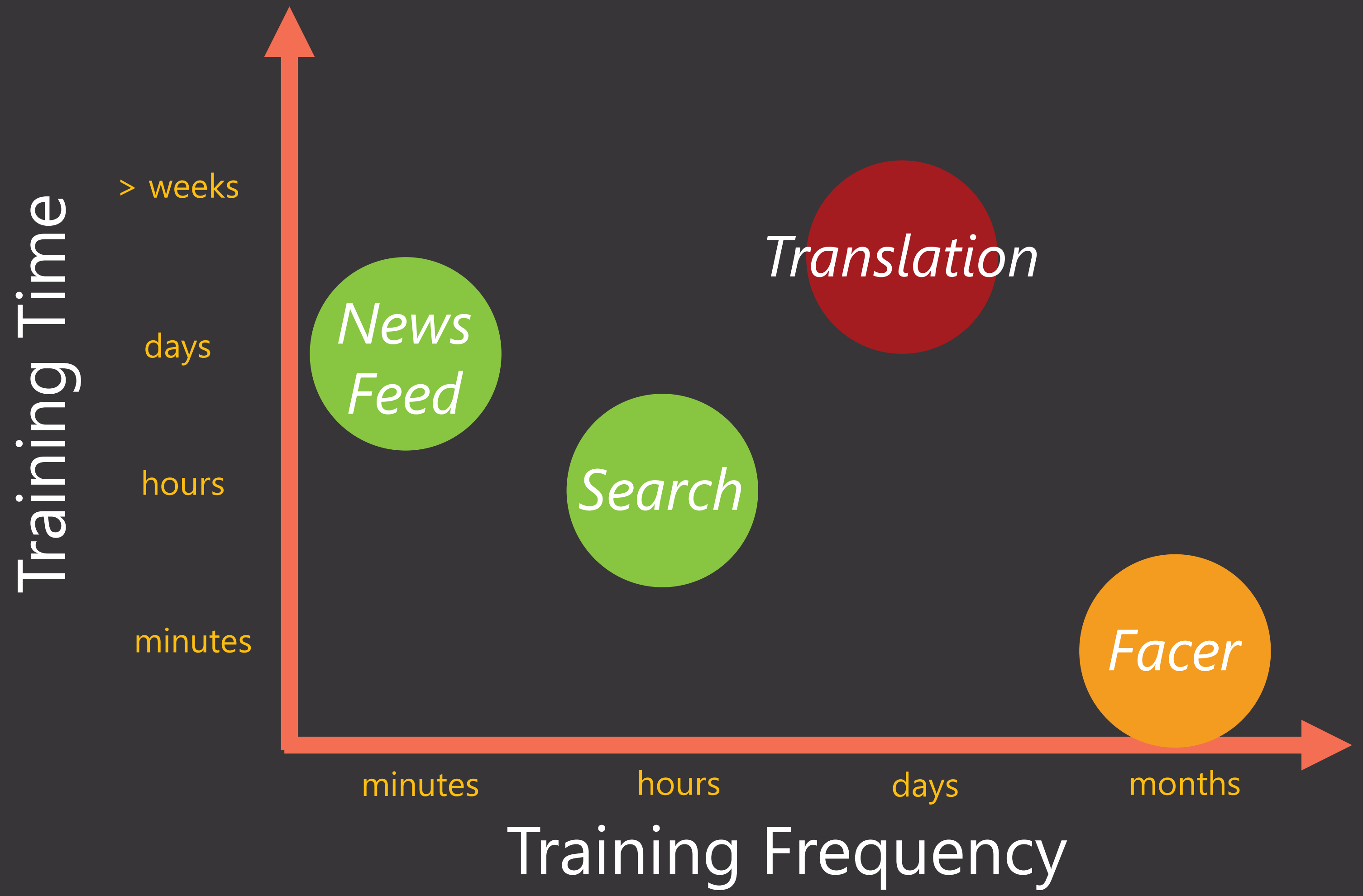


Augmented Reality with Smart Camera

ML Execution Flow



ML Model Training at Facebook



What about Inference?

200+ Trillion

Total Predictions Per Day

6+ Billion

Language Translations Per Day

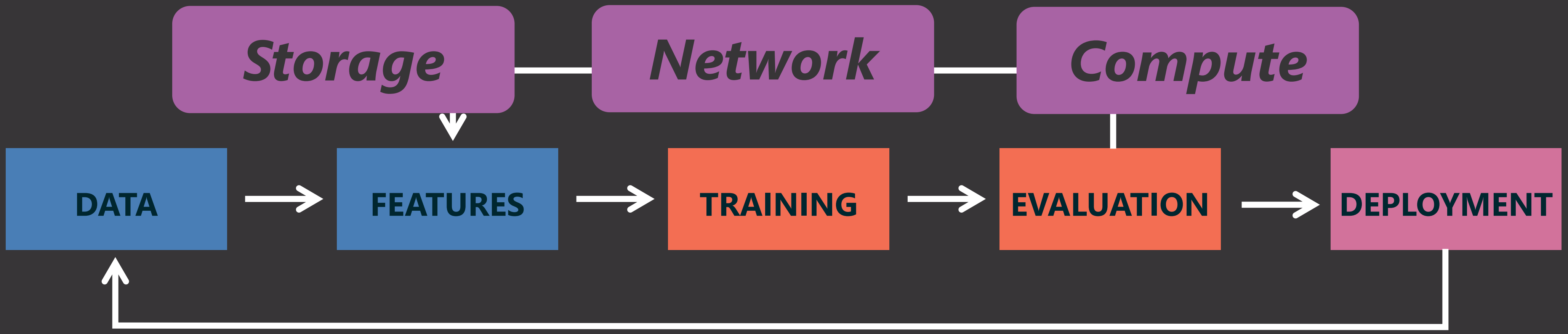
Millions

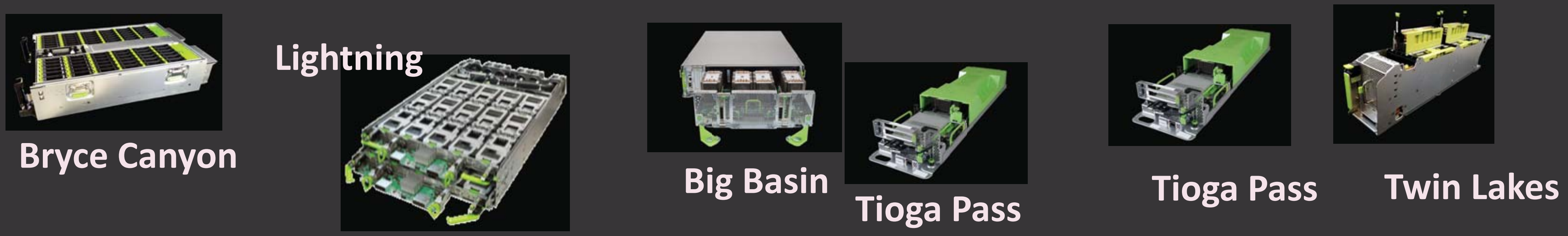
Fake Accounts Removed Proactively by Automated Systems Every Day

First, with Custom-Designed System Solutions

Facebook’s philosophy is to:

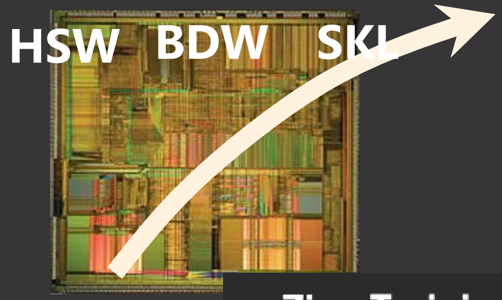
- Characterize and bucketize ML workloads of critical importance
- Custom-design server systems for the bucketized workloads

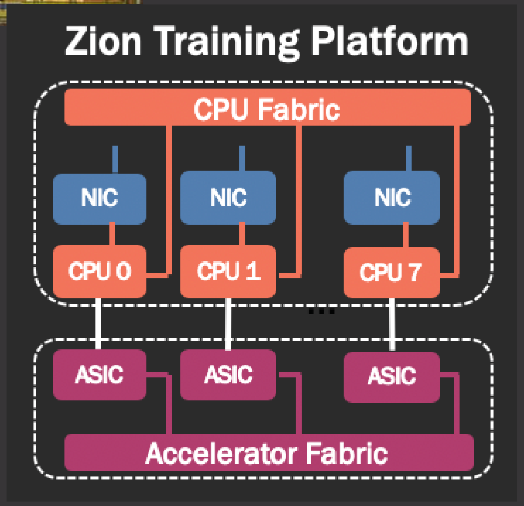


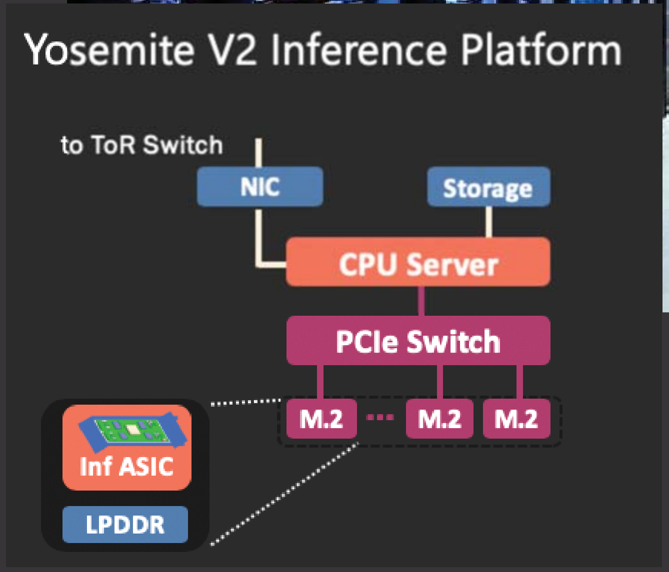


Highly Scalable Infrastructure

Scaling Up → Scaling Out







Outline

- Overview for Machine Learning @ Facebook
- Diversity of Machine Learning Workloads
- Neural Personalized Recommendation and System Implications
- Machine Learning Inference at the Edge
- Conclusion

Diversity in ML Models at Facebook

Support Vector Machines
SVM
- Facer

Gradient-Boosted Decision Trees
GBDT
- Sigma

Multi-Layer Perceptron
MLP
- News Feed
- Ads
- Search
- Sigma

Convolutional Neural Nets
CNN
- Facer

Recurrent Neural Nets
RNN
- Language Translation
- Speech Rec.
- Content Understanding

Hazelwood *et al.*, “Applied Machine learning at Facebook: A Datacenter Infrastructure Perspective”, *HPCA* 2018.

Diversity in ML Models at Facebook

Support Vector Machines — **SVM**
- Facer

Gradient-Boosted Decision Trees — **GBDT**
- Sigma

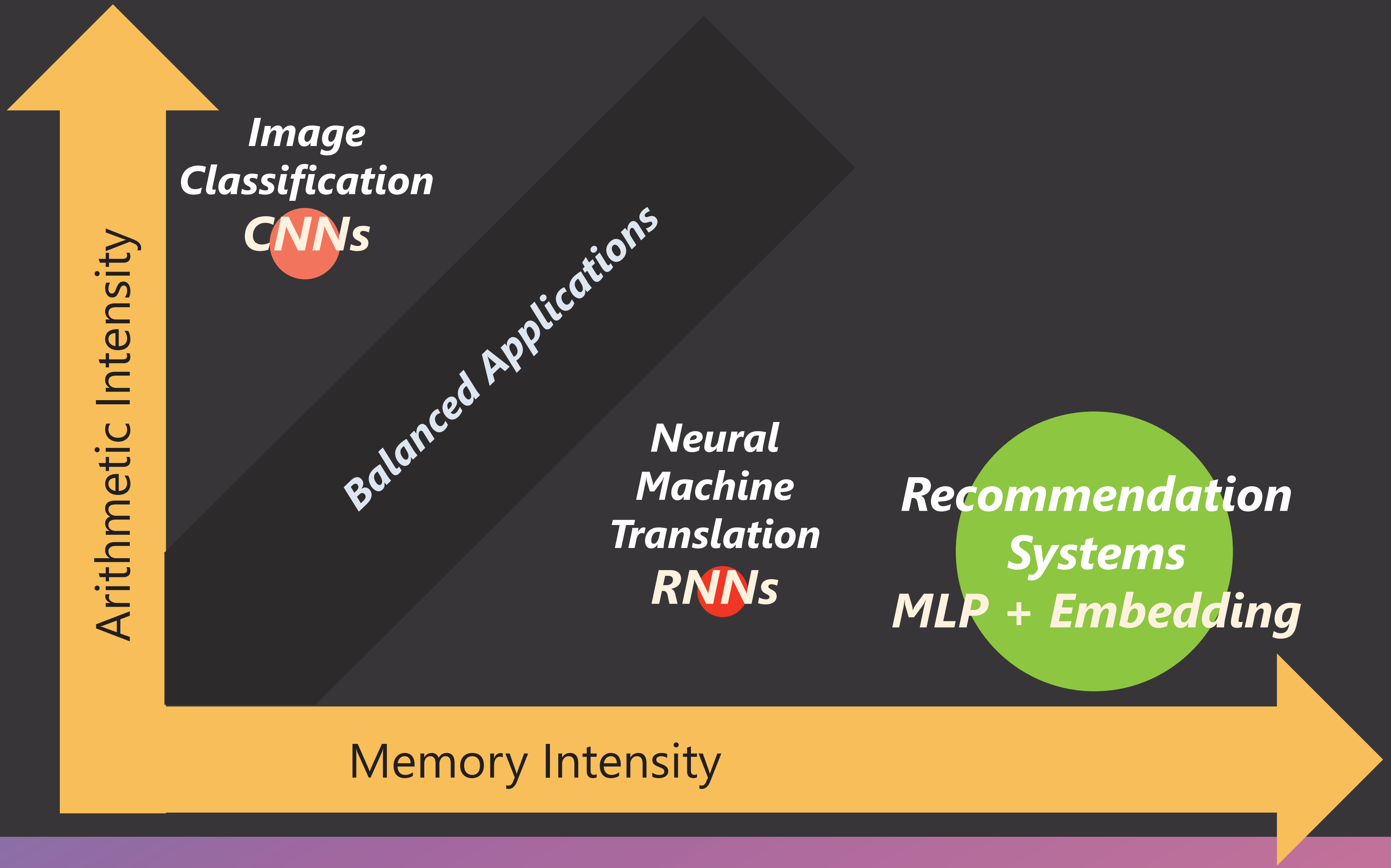
Multi-Layer Perceptron — **MLP**
- News Feed
- Ads
- Search
- Sigma

Convolutional Neural Nets — **CNN**
- Facer

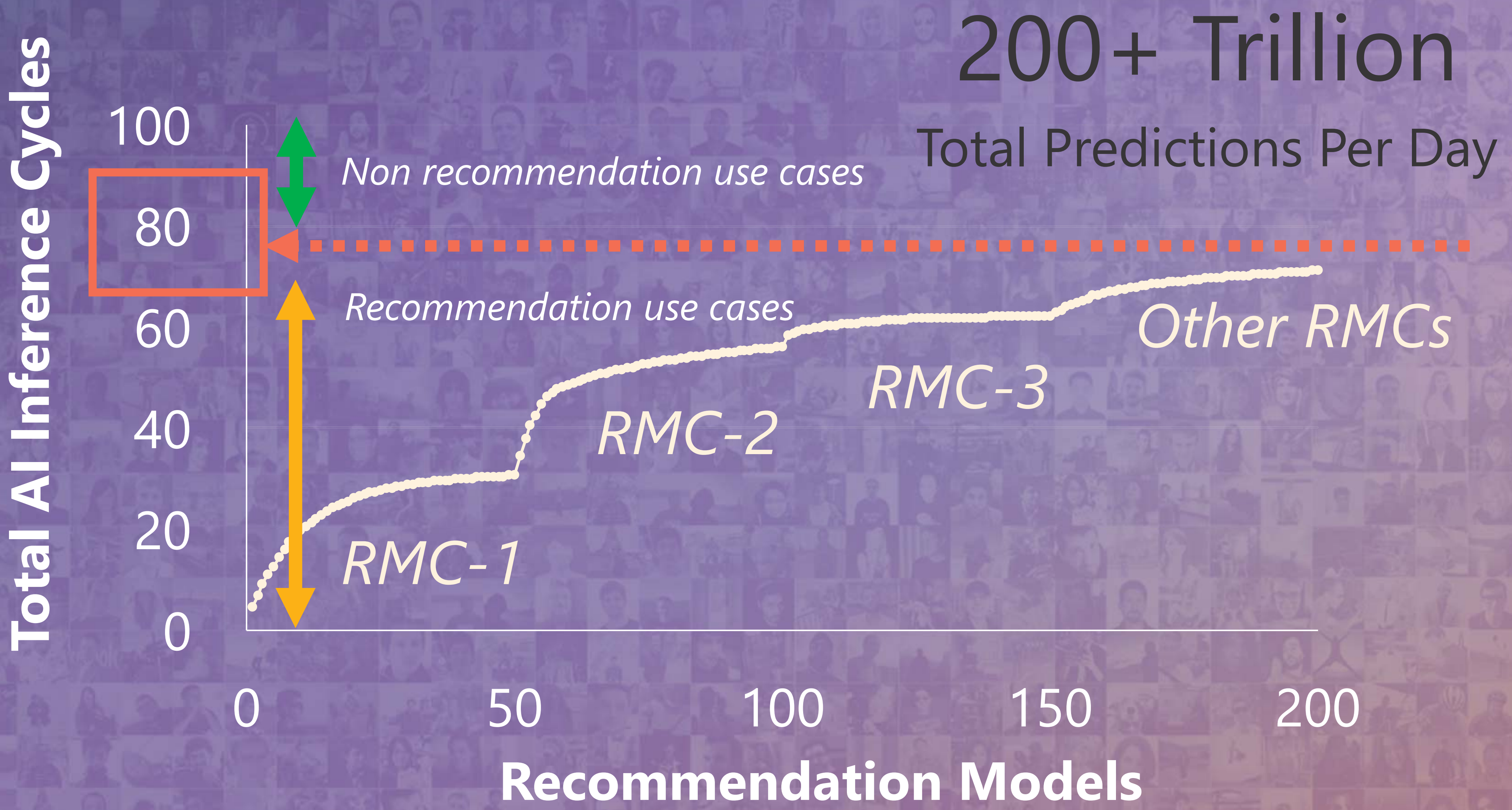
Recurrent Neural Nets — **RNN**
- Language Translation
- Speech Rec.
- Content Understanding

Hazelwood *et al.*, “Applied Machine learning at Facebook: A Datacenter Infrastructure Perspective”, *HPCA* 2018.

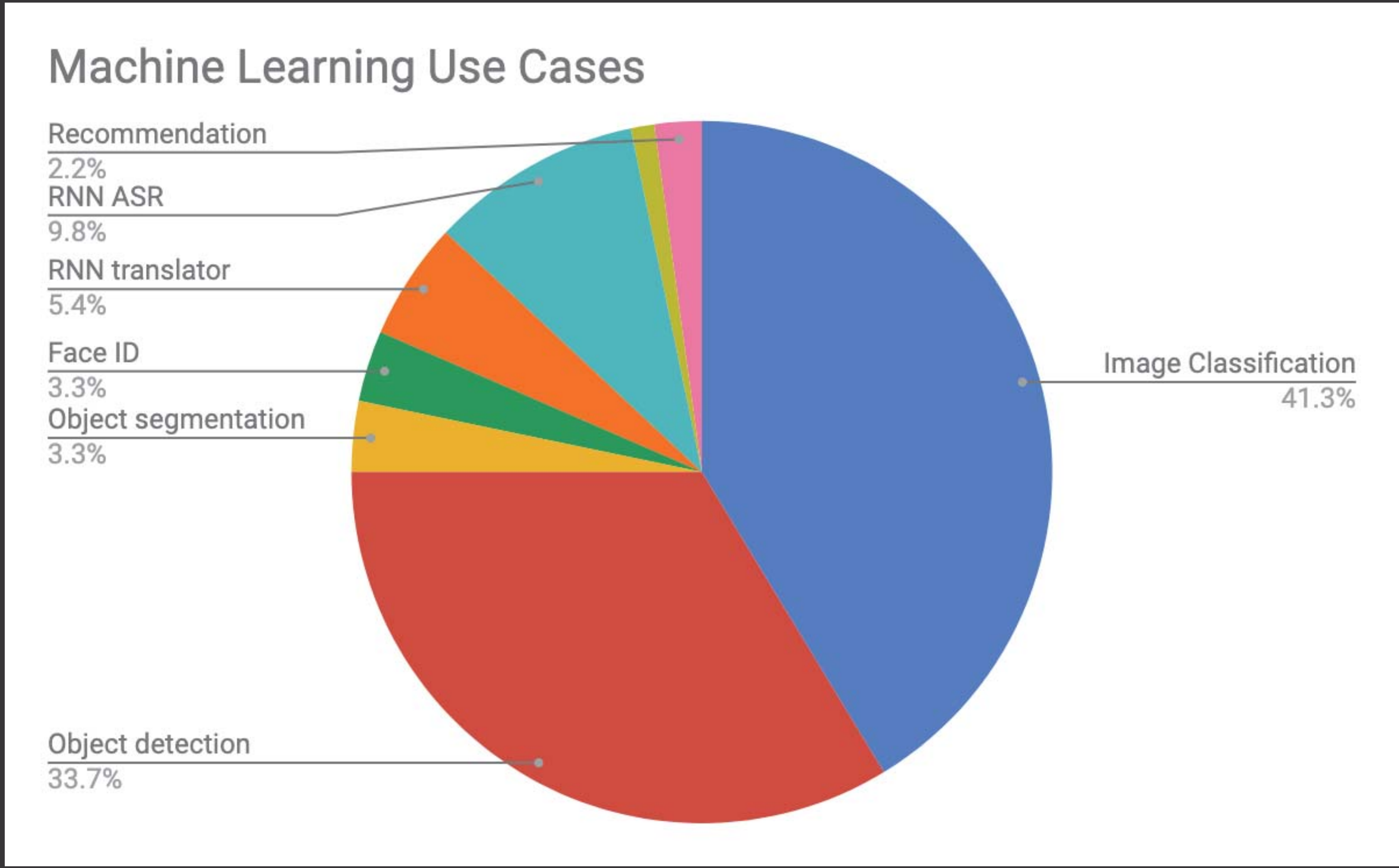
Diversity in DNN Use Cases



AI Inference Cycle Breakdown

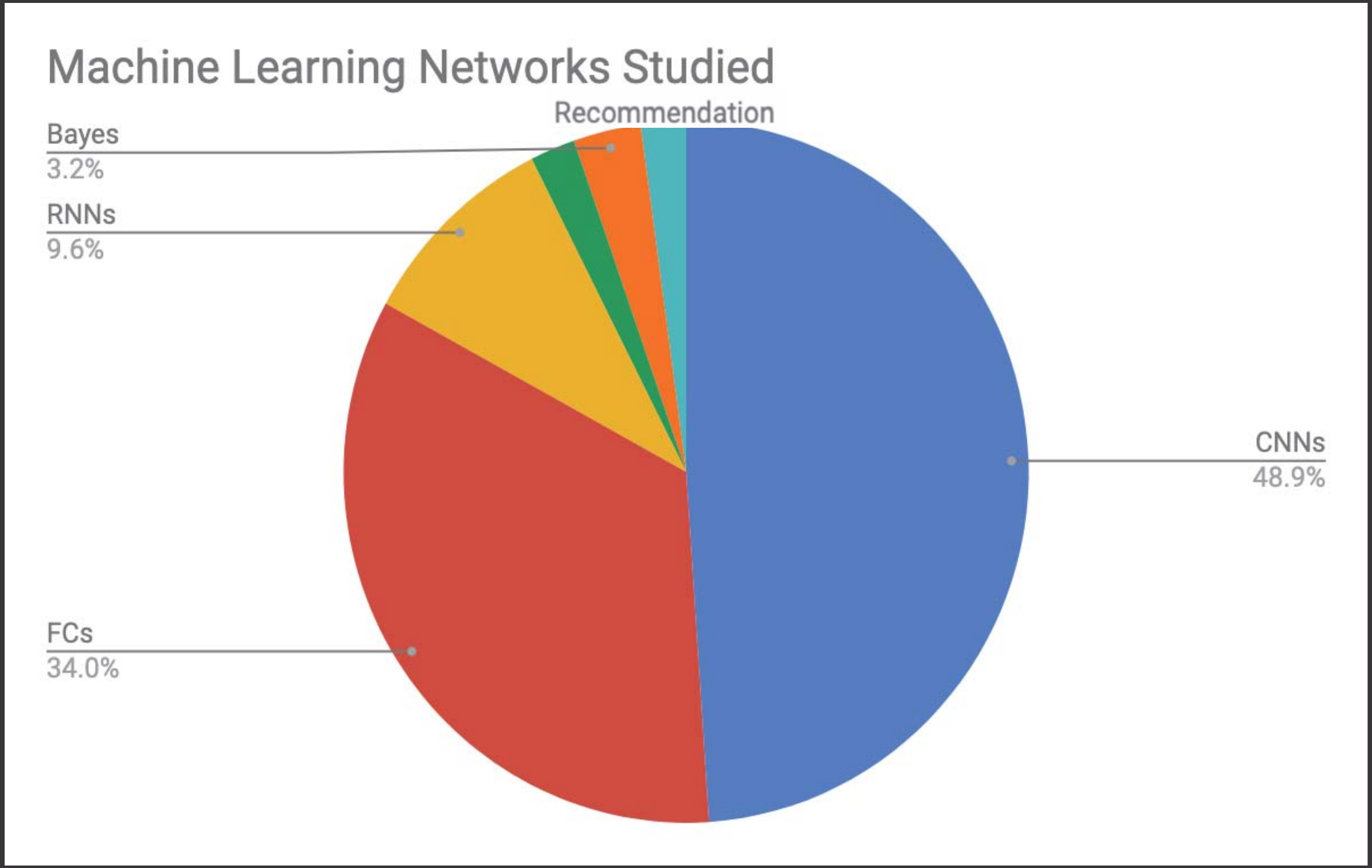


ML Topics of Interest by the Research Community



https://www.sigarch.org/deep-learning-its-not-all-about-recognizing-cats-and-dogs/

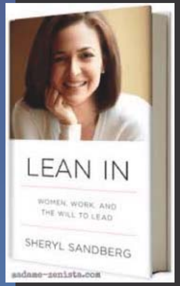
Modeling Techniques Studied by the Research Community



Neural Personalized Recommendation Systems

The Use Case Challenge

An Example of Recommendation



User/Dense Features

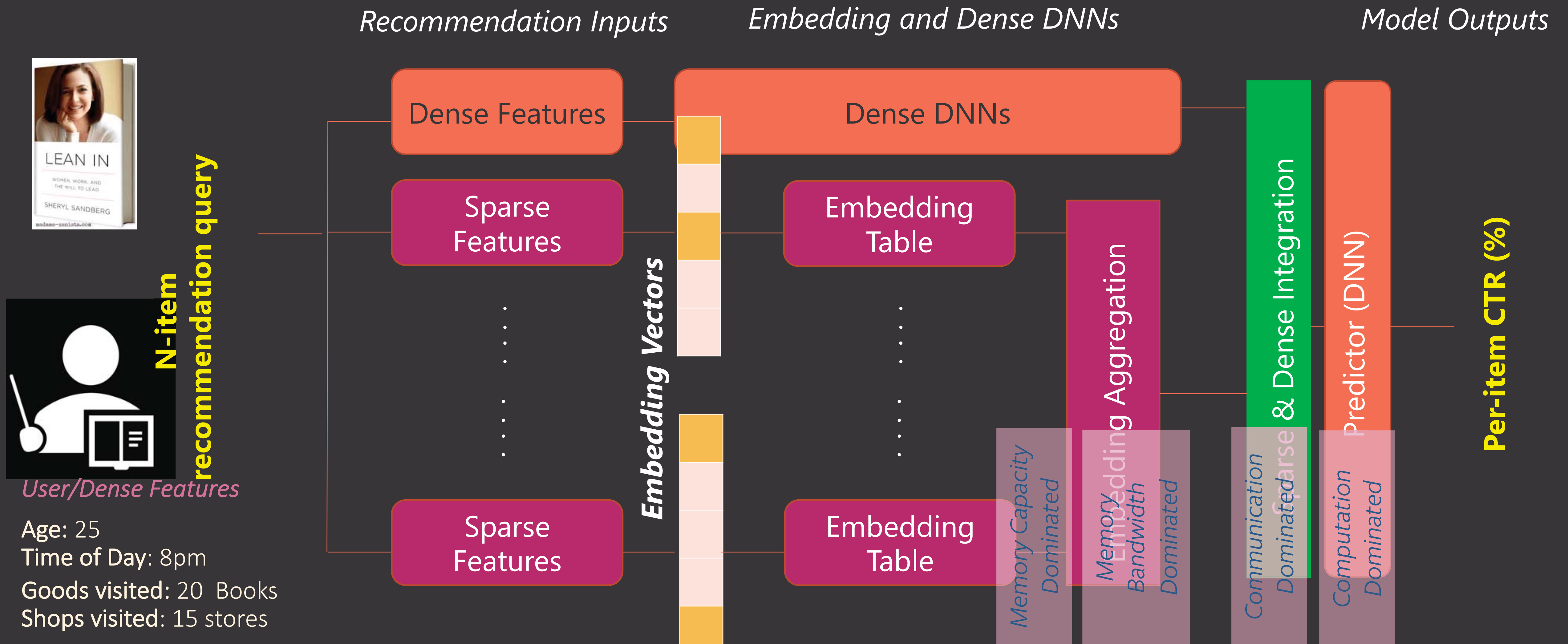
Age: 25
Time of Day: 8pm

Likelihood of Clicks

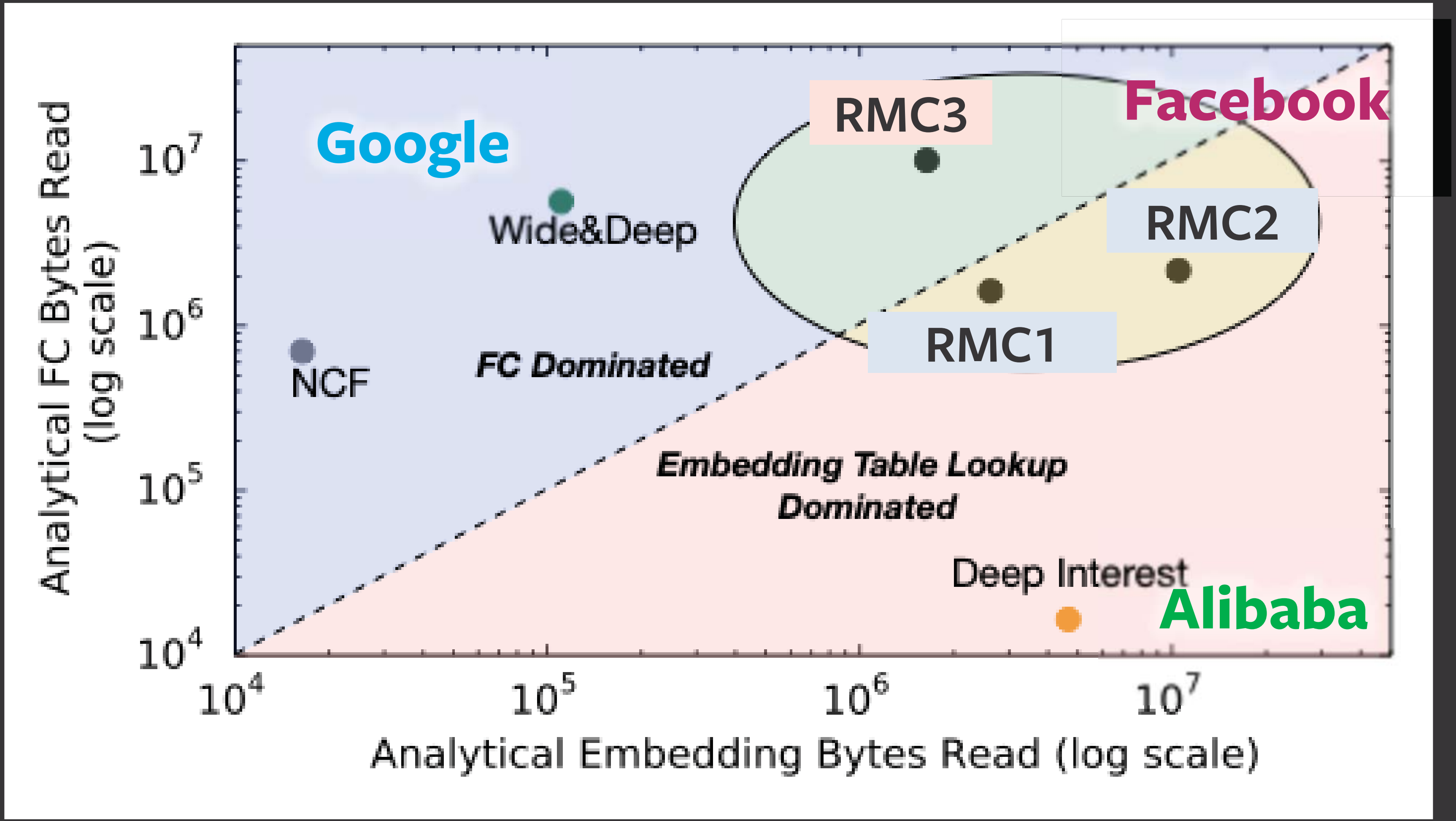
Categorial/Sparse Features

Goods visited: 20 Books
Shops visited: 15 stores

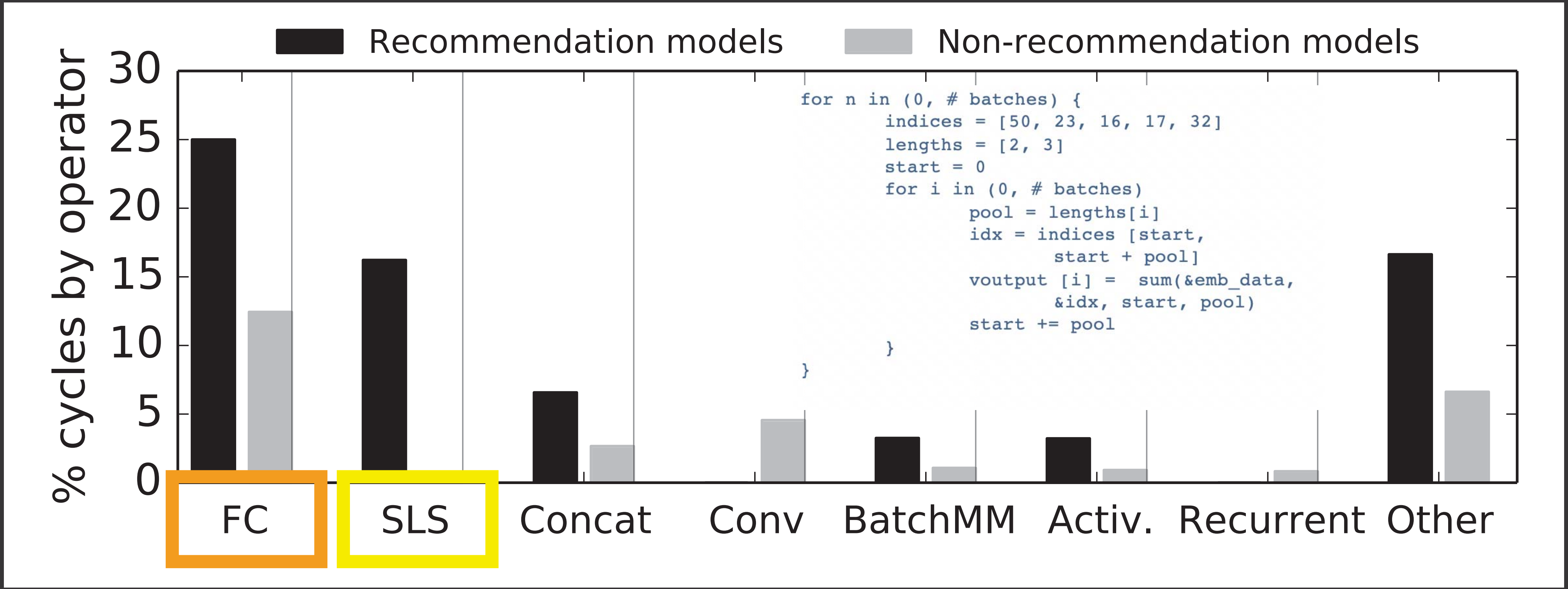
What is Deep Learning Personalized Recommendation?



Diversity in Recommendation Models

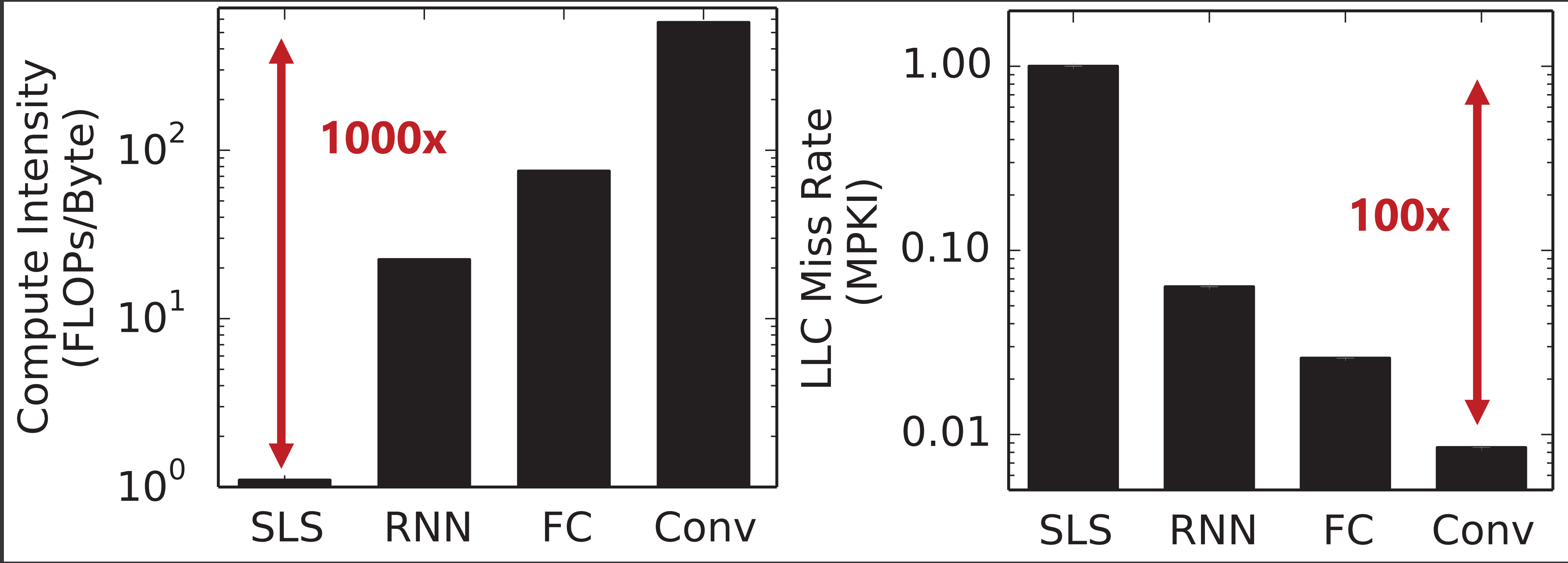


ML Operator Breakdown at Facebook Datacenter Fleet



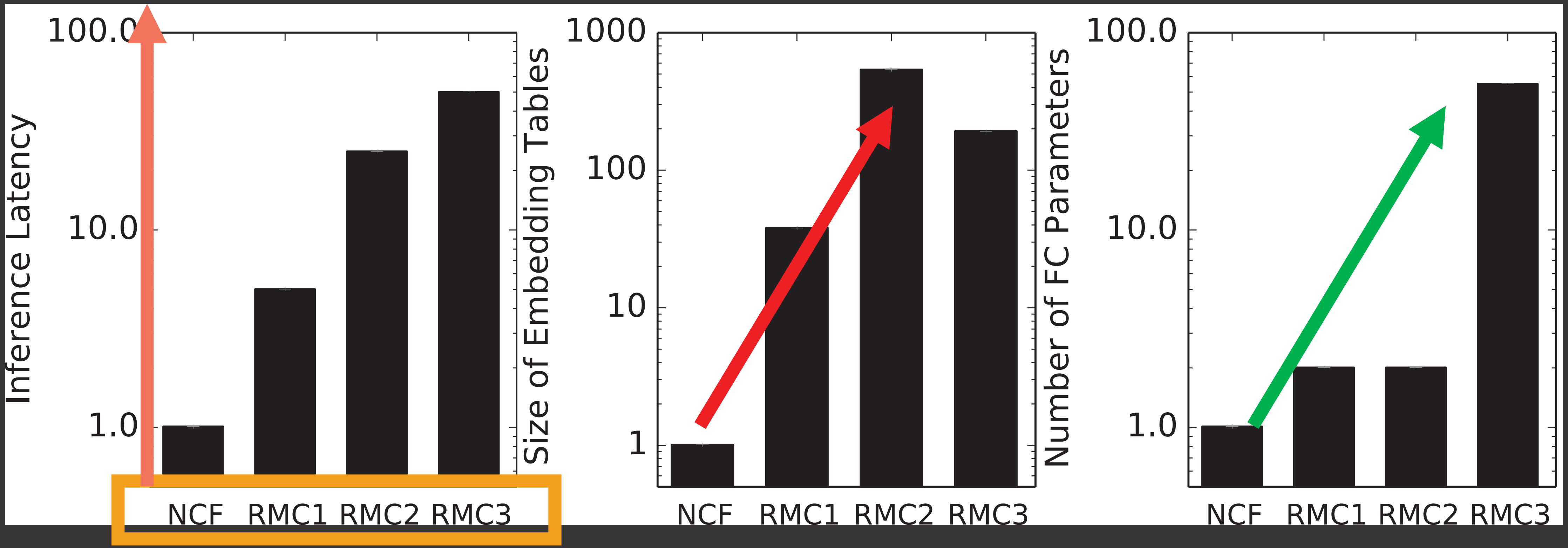
Gupta et al., "The Architectural Implications of Facebook's DNN-based Personalized Recommendation Systems," HPCA-2020.

Embedding Table Accesses Incur High LLC MPKI with Low Compute Intensity



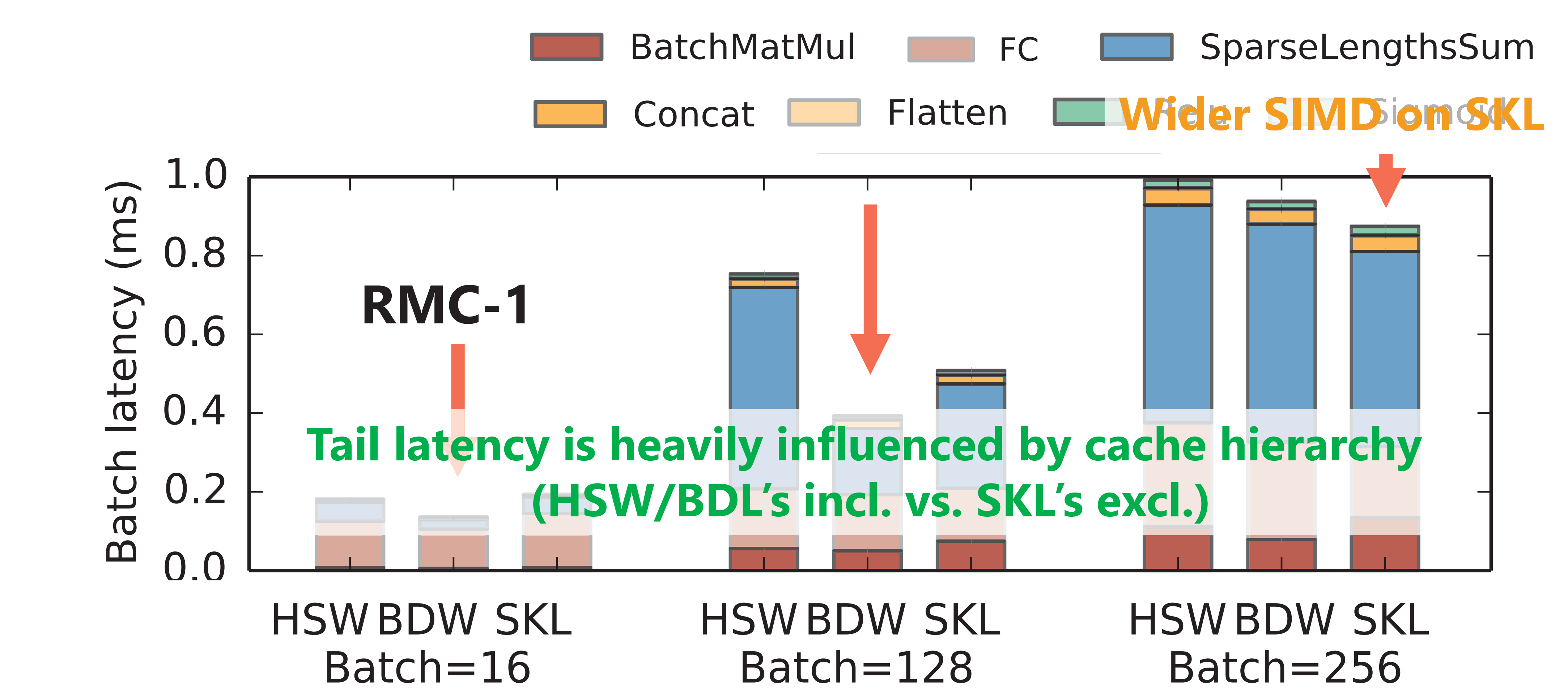
Gupta et al., "The Architectural Implications of Facebook's DNN-based Personalized Recommendation Systems," HPCA-2020.

Major Categories of Recommendation Models – RMC-1, RMC-2, RMC-3



** NCF from MLPerf v0.5 Training*

Lower Latency on SKL with Large Batching



Developing a Recommendation Benchmark for MLPerf Training and Inference

Carole-Jean Wu [1] Robin Burke [2] Ed Chi [3] Joseph Konstan [4] Julian McAuley [5] Yves Raimond [6] Hao Zhang [1]

Deep Learning Recommendation Model for Personalization and Recommendation Systems

a Mudigere, Hao-Jun Michael Shi*, Jianyu Huang,
o Park, Xiaodong Wang, Udit Gupta†, Carole-Jean Wu,
lgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu,
Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira,
jay Rao, Bill Jia, Liang Xiong and Misha Smelyanskiy
Hacker Way, Menlo Park, CA 94065
umov,dheevatsa}@fb.com

The Architectural Implications of Facebook's DNN-based Personalized Recommendation

Udit Gupta*, Carole-Jean Wu, Xiaodong Wang, Maxim Naumov, Brandon Reagen

David Brooks*, Bradford Cottel, Kim Hazelwood, Mark Hempstead, Bill Jia, Hsien-Hsin S. Lee, Andrey Malevich, Dheevatsa Mudigere, Mikhail Smelyanskiy, Liang Xiong, Xuan Zhang

Facebook Inc.
{carolejeanwu, xdwang}@fb.com

Exploiting Parallelism Opportunities with Deep Learning Frameworks

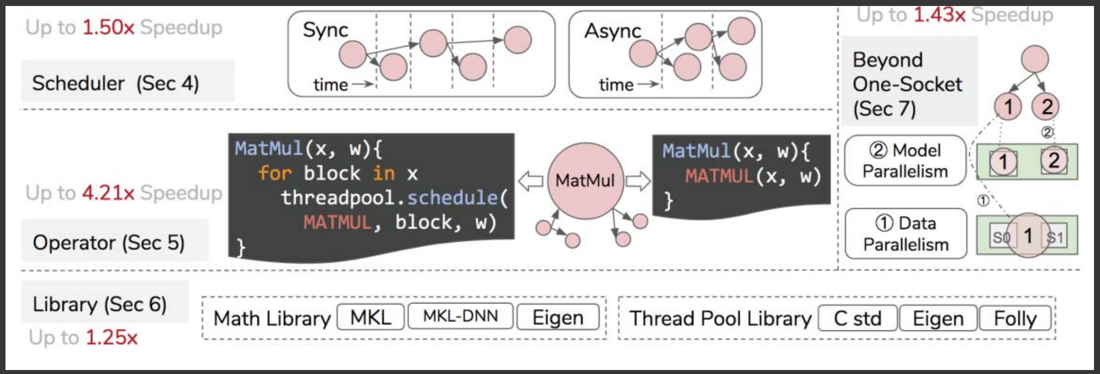
Yu Emma Wang
ywang03@g.harvard.edu
Harvard University

Carole-Jean Wu
carolejeanwu@fb.com
Facebook

Xiaodong Wang
xdwang@fb.com
Facebook

Kim Hazelwood
kimhazelwood@fb.com
Facebook

David Brooks
dbrooks@eecs.harvard.edu
Harvard University

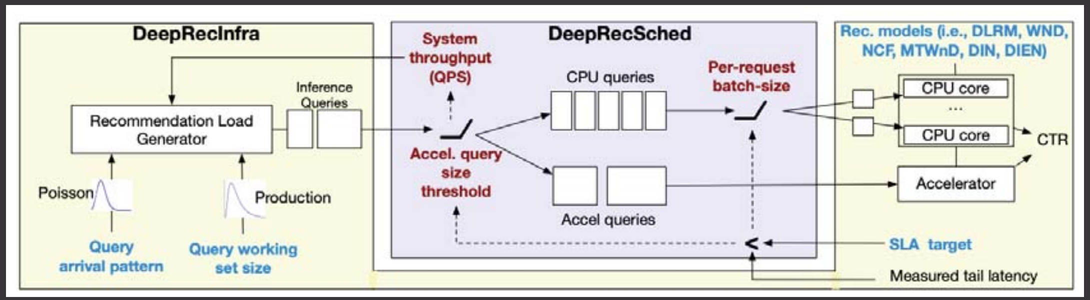


DeepRecSys: A System for Optimizing End-To-End At-scale Neural Recommendation Inference

Udit Gupta[*δ], Samuel Hsia[*], Vikram Saraph[δ], Xiaodong Wang[δ], Brandon Reagen[δ], Gu-Yeon Wei[*], Hsien-Hsin S. Lee[δ], David Brooks[*δ], Carole-Jean Wu[δ]

[*]Harvard University [δ]Facebook Inc.

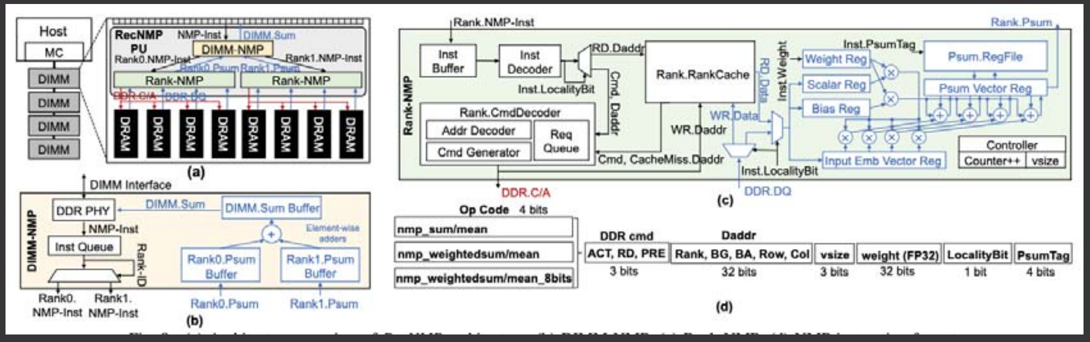
ugupta@g.harvard.edu carolejeanwu@fb.com



RecNMP: Accelerating Personalized Recommendation with Near-Memory Processing

Liu Ke, Udit Gupta, Carole-Jean Wu, Benjamin Youngjae Cho, Mark Hempstead, Brandon Reagen, Xuan Zhang

David Brooks, Vikas Chandra, Utku Diril, Amin Firoozshahian, Kim Hazelwood, Bill Jia, Hsien-Hsin S. Lee
Meng Li, Bert Maher, Dheevatsa Mudigere, Maxim Naumov, Martin Schatz, Mikhail Smelyanskiy, Xiaodong Wang



More on the DNN-based Recommendation Models

- Facebook Deep Learning Recommendation Model (DLRM)
  - https://github.com/facebookresearch/dlrm
- At-Scale Infrastructure Implication on Neural Recommendation Optimization
  - MLPerf Training and Inference Benchmark Suites

Outline

- Overview for Machine Learning @ Facebook
- Diversity of Machine Learning Workloads
- Neural Personalized Recommendation and System Implications
- Machine Learning Inference at the Edge
- Conclusion

Unique Challenges for Edge Inference

The **Diversity of Mobile Hardware and Software** is Not Found in the Controlled Datacenter Environment.

2	3	20+	20+	10+
MAJOR MOBILE OS	MAJOR GRAPHICS APIs	MAJOR CHIPSET VENDORS	MAJOR CPU UARCH	MAJOR GPU UARCH

How do we optimize system designs for real-time ML inference?

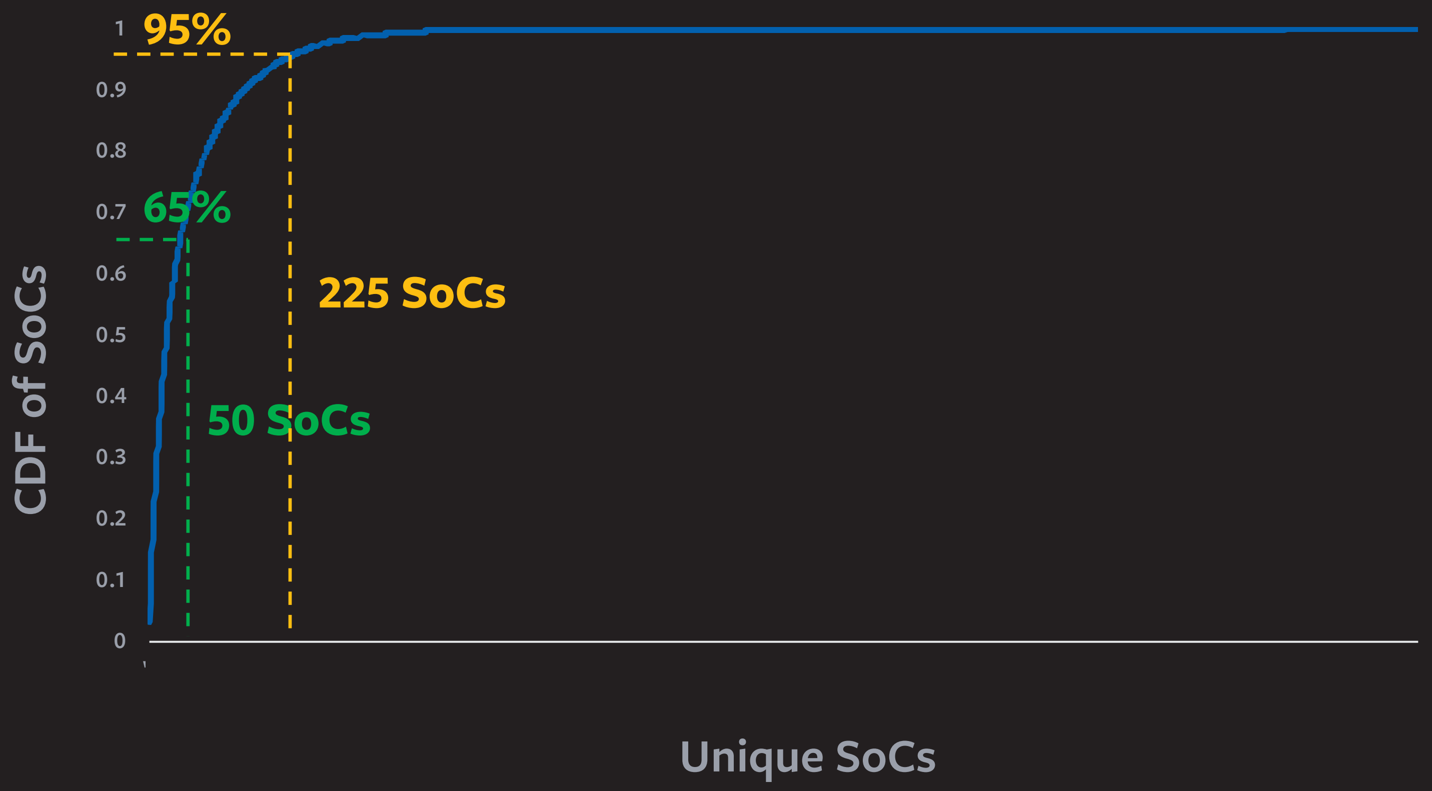
FRAGMENTED SMARTPHONE ECOSYSTEM POSES UNIQUE CHALLENGES FOR EDGE INFERENCE

Machine Learning at Facebook: Understanding Inference at the Edge. Wu et al. HPCA-2019.

Lay of the Land

FRAGMENTATION

Taking a Closer Look at Smartphones Facebook Runs on



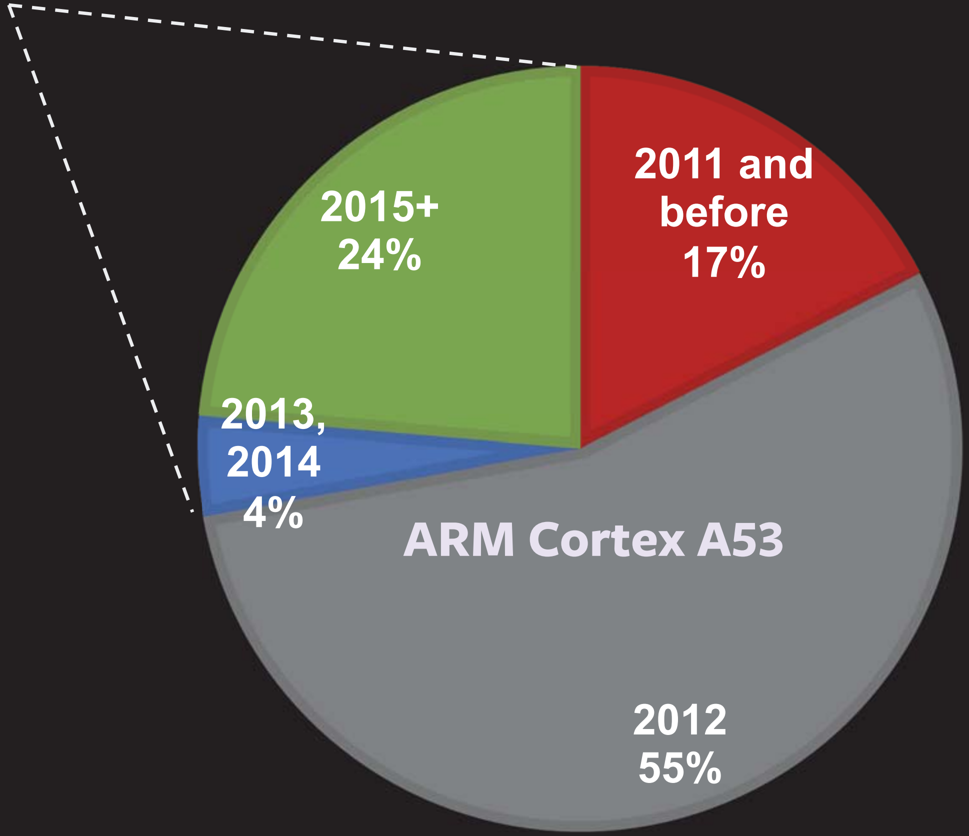
- Qualcomm Snapdragon
- Samsung Exynos
- MediaTek Helio
- HiSilicon Kirin et al.

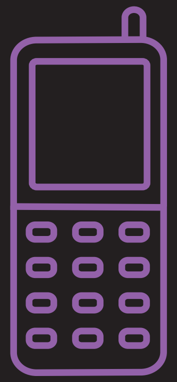
THERE IS NO STANDARD SOC TO OPTIMIZE FOR

Lay of the Land

FRAGMENTATION

In 2018, ~28% of SoCs Use CPUs Designed in 2013 or Later





72%

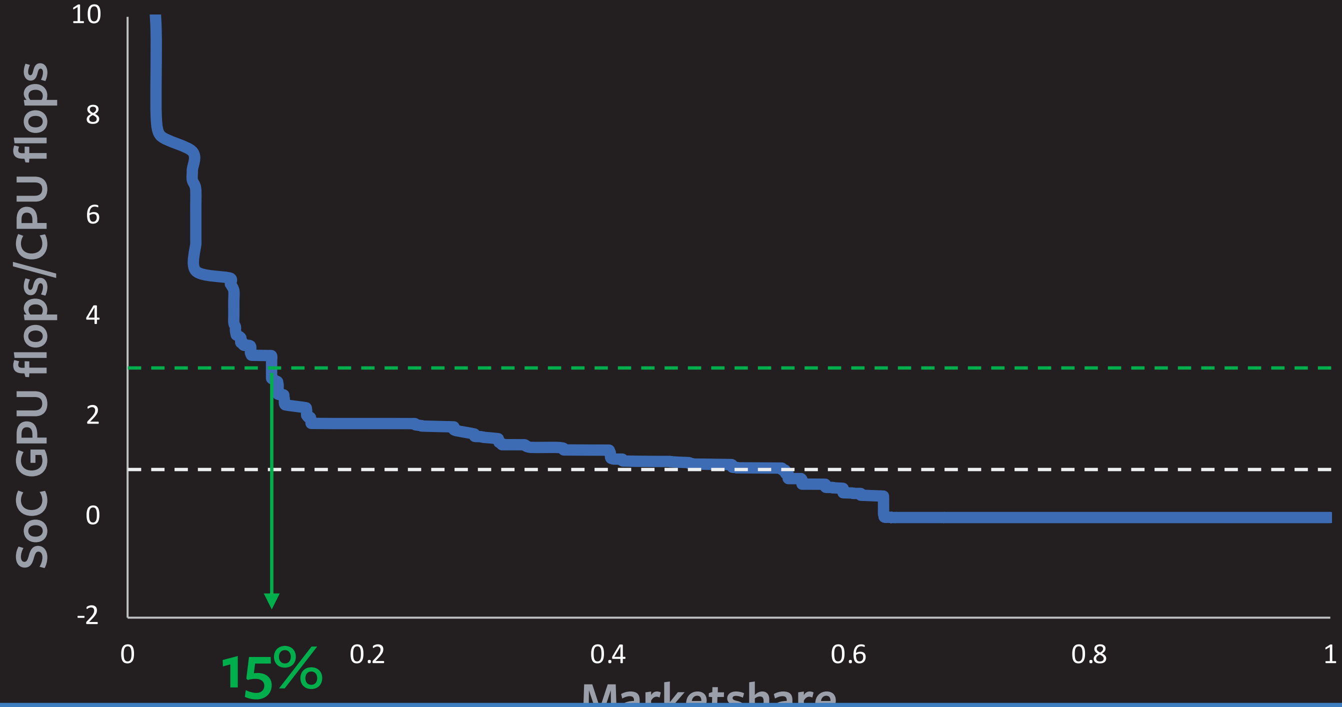
OF THE WORLD’S CELL PHONES ARE MORE THAN 7 YEARS OLD

MOBILE CPUS SHOW LITTLE DIVERSITY

Lay of the Land

PERFORMANCE

The Performance Difference between a Mobile CPU and GPU is Narrow



ON A MEDIAN SMARTPHONE, THE GPU PROVIDES AS MUCH THEORETICAL PEAK PERFORMANCE AS ITS CPU

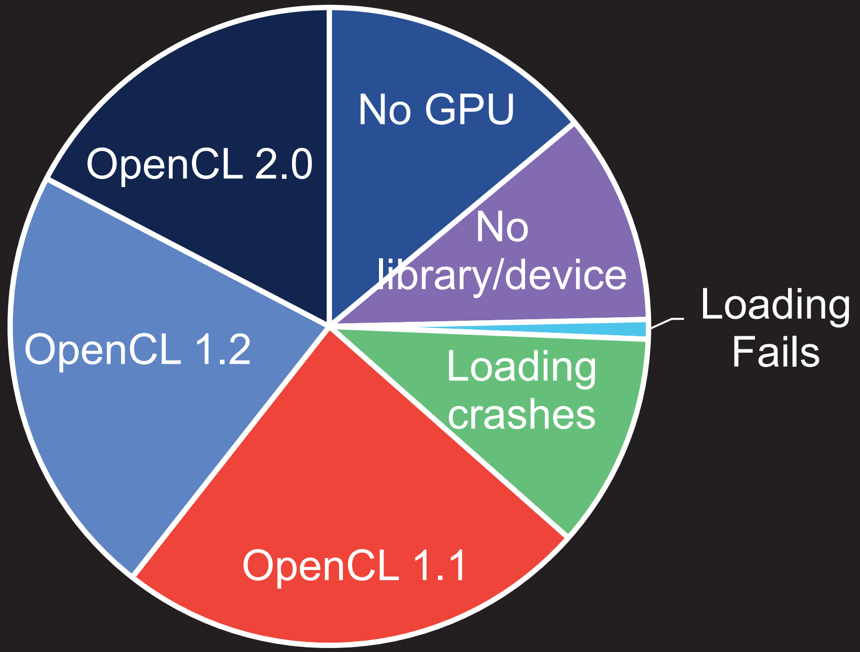
LESS THAN 15% SMARTPHONES HAVE A GPU THAT IS 3 TIMES AS POWERFUL AS ITS CPU

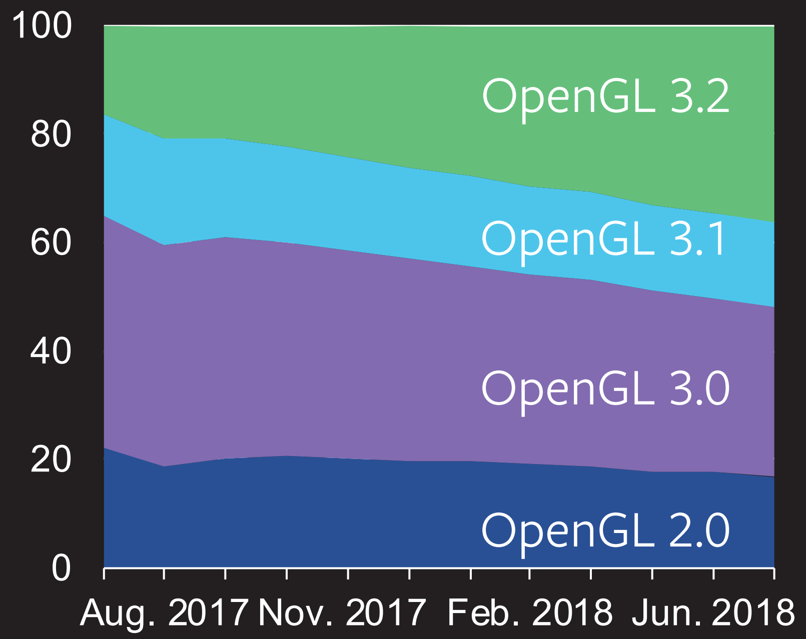
Lay of the Land

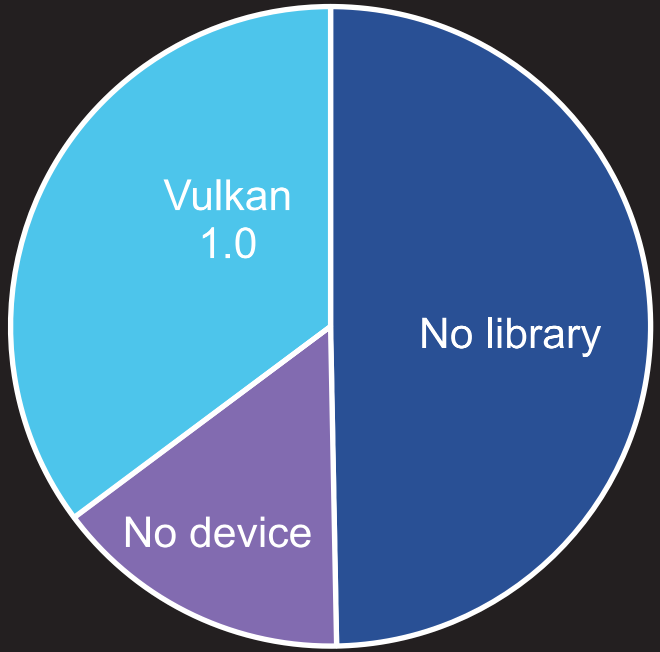
PROGRAMMABILITY

Programmability is a Primary Roadblock for Using Mobile Co-processors

- **OpenCL, OpenGL ES, Vulkan for Android GPUs**







ANDROID GPUS HAVE FRAGILE USABILITY AND POOR PROGRAMMABILITY WHILE IOS HAS BETTER SUPPORT WITH METAL

Quantitative Approach to Edge Inference Designs

State of the Practice for Mobile Inference is Using CPUs

FRAGMENTATION

- There are more than **2000+ different SoCs** but mobile CPUs show little diversity with ARM's Cortex A53 dominating the market

PERFORMANCE

- Performance difference between mobile **CPUs** and **GPUs** is narrow

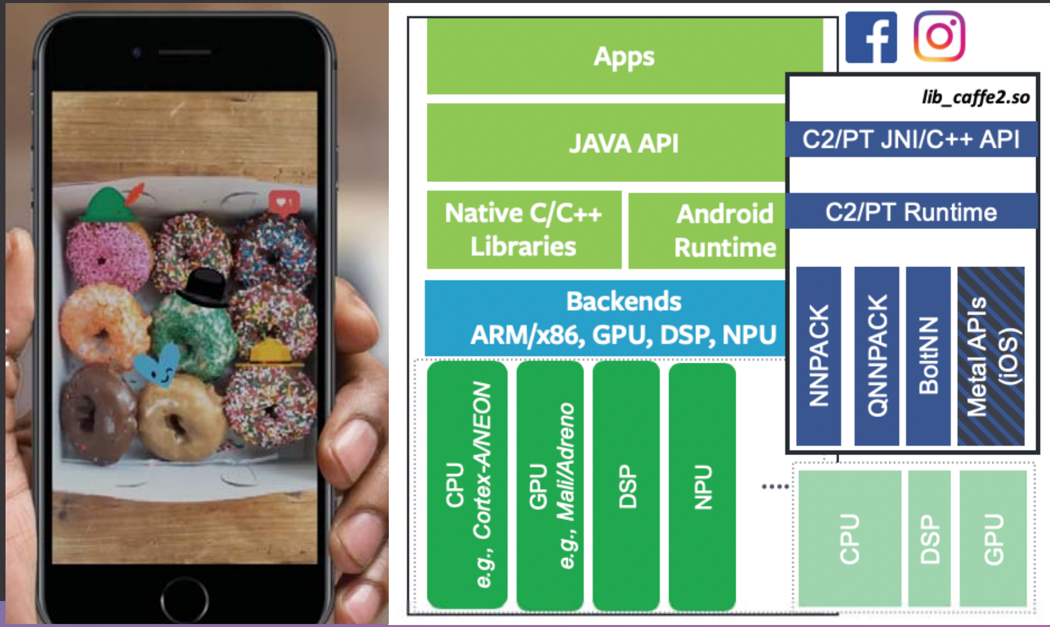
PROGRAMMABILITY

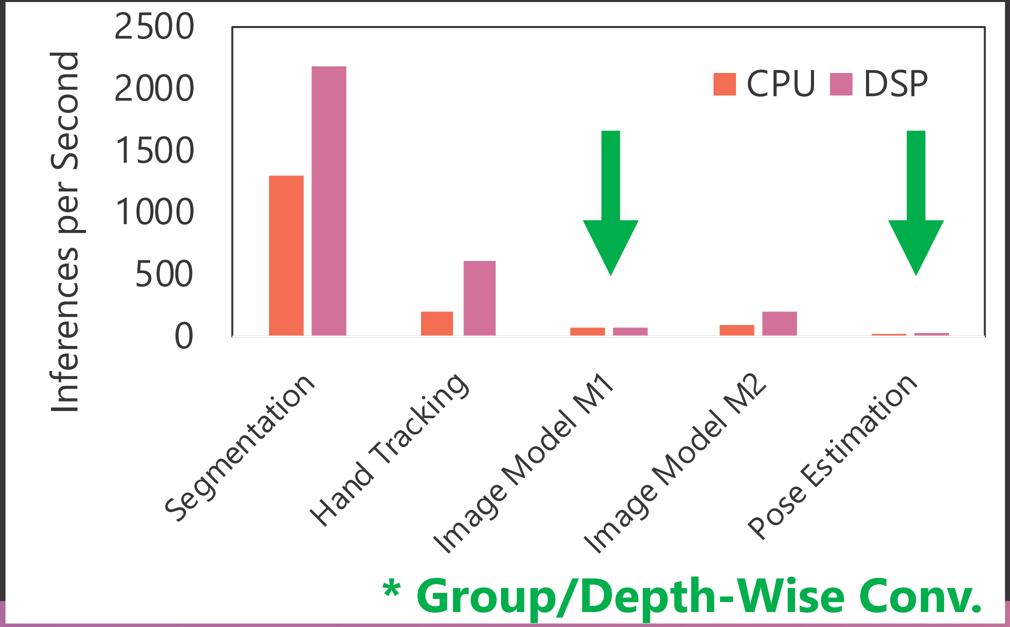
- Programmability is a major road block for **co-processors** (e.g. Android GPUs)

MOBILE INFERENCE OPTIMIZATION IS TARGETED FOR THE COMMON DENOMINATOR OF THE FRAGMENTED SOC ECOSYSTEM

More Detail on Inference at the Edge

- Machine Learning at Facebook: Understanding Inference at the Edge. *Wu et al. HPCA-2019.*
  - Horizonal integration for efficient mobile inference
  - Vertical integration for efficient AR/VR inference
  - Variability matters (not just in the datacenter)



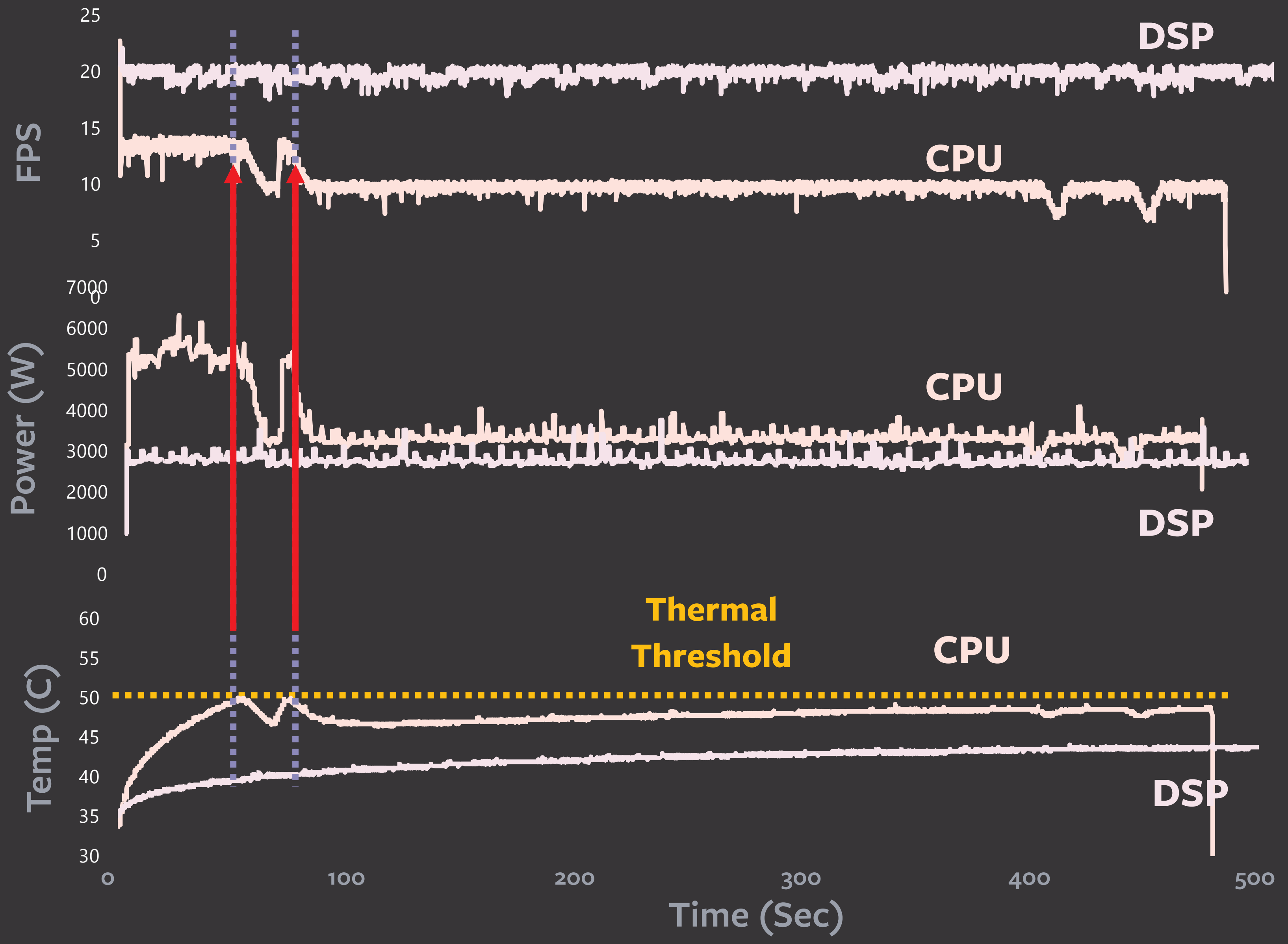


Vertical Integrated Systems

Making Inference on DSPs Leads to Consistent Performance

CPU thermal throttling causes sudden **FPS drop**

The primary reason for using co-processors and accelerators are for **lower power** and **more stable performance**



Outline

- Overview for Machine Learning @ Facebook
- Diversity of Machine Learning Workloads
- Neural Personalized Recommendation and System Implications
- Machine Learning Inference at the Edge
- Conclusion

At-Scale Infrastructure Challenges for Machine Learning

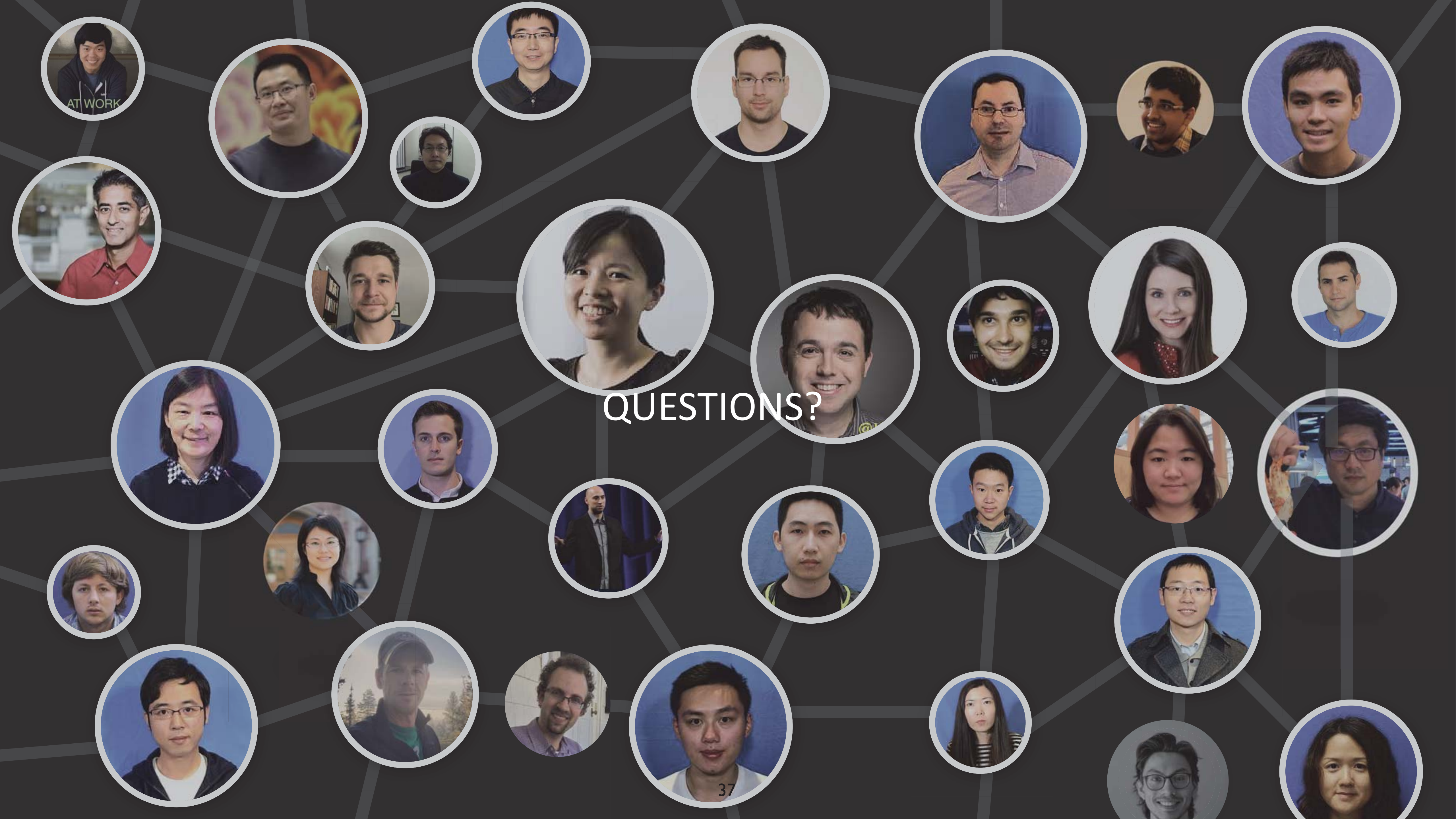
- **Diversity of ML Models in Facebook's Datacenter**
- **A Variety of Neural Personalized Recommendation Models Dominate AI Inference Cycles**
- **Legacy Devices Matter; Performance Differences at the Edge Are Huge**
- **Without solid performance analysis for ML models, we are in the dark**

DeepRecSys

MLPerf Benchmark Suite

It is important to consider full-picture and system effects for efficient, practical at-scale ML infrastructure designs

K. Hazelwood et al., "Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective," HPCA 2018.
C.-J. Wu et al., "Machine Learning at Facebook: Understanding Inference at the Edge," HPCA 2019.
U. Gupta et al., "The Architectural Implications of Facebook's DNN-based Personalized Recommendation," HPCA 2020.



QUESTIONS?