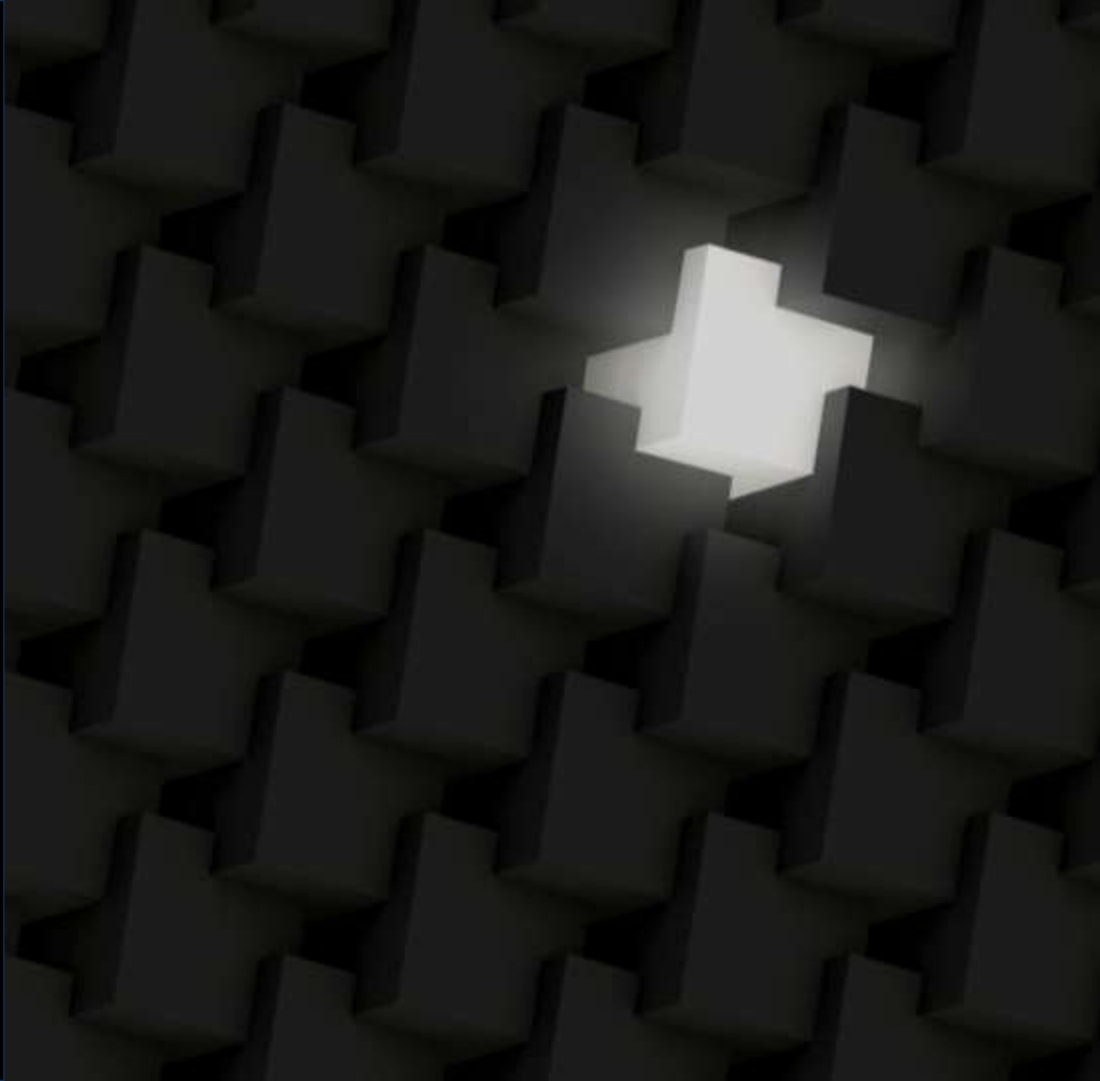
Carnegie Mellon University
Software Engineering Institute

RESEARCH REVIEW 2020

Topics in Advanced Computing: Promise and Challenges of Recommendation Systems for the DoD

John Wohlbier, Scott McMillan CMU SEI

Tze Meng Low, Elliot Binder CMU ECE



DM20-0946

Recommendation Systems Overview

- Concepts and foundations
- Applications to DoD and IC
- Facebook DLRM and MLPerf
- Advanced computing in the ETC
- CMU SEI advances in DLRM
- Impact

What is a Recommendation System?

Given your profile and the things you've liked in the past, what is the probability that you will "click through" on a recommendation?

- Netflix
- Amazon
- YouTube
- Spotify
- Facebook
- Twitter

"DNN-based personalized recommendation models comprise up to 79% of AI inference cycles in a production-scale data center."

Gupta, Udit, et al. "The architectural implications of Facebook's DNN-based personalized recommendation." 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2020.

The Idea Behind Recommendation Systems

Given a "user" and an "item" that the user has not interacted with, what is the probability that the user will click on the item?

User-item pairs with the highest predicted click-through rate are prioritized

The data is "sparse," i.e., any given user has interacted with very few items

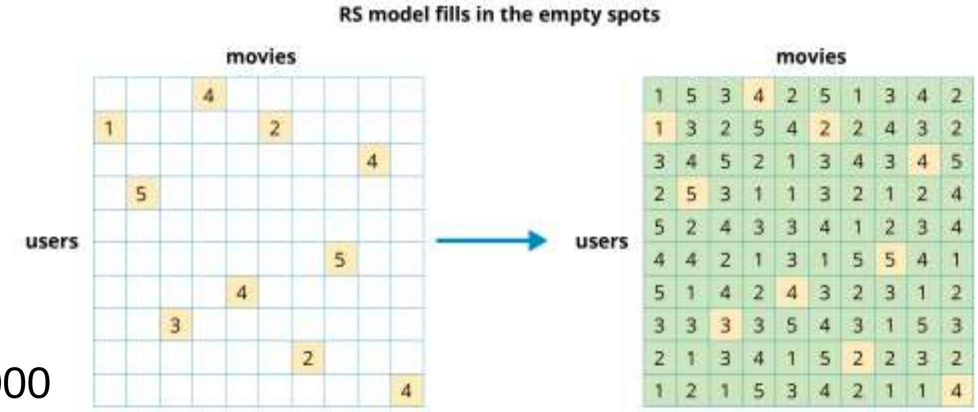
Sparsity example: Netflix Prize Dataset

- 17,770 movies
- 480,189 users

Ratings on scale of 1 – 5.

~100,000,000 total ratings

- ~20,000 x ~500,000 = ~10,000,000,000
- Sparsity: 100,000,000 / 10,000,000,000 ~ 1%



Recommendation Systems in the DoD and IC

Intelligence Analysis	Cybersecurity Analysis	Social Network Analysis
• Prioritizing documents when number of documents much greater than number of analysts • Guiding novice analyst searches using search paths of more experienced analysts	• Generating prioritized lists for defense actions • Detecting insider threats • Monitoring network security • Predicting cyber attacks • As an attack vector • Software vulnerability severity assessments	• Discovering fake news • Identifying malicious conversations

Recommendation Systems are Appearing in the JAIC

JAIC Mission Initiatives

- Joint Warfighting Operations
- Warfighter Health
- Business Process Transformation
- Threat Reduction and Protection
- Joint Logistics
- Joint Information Warfare

Kitware Inc. developed and demonstrated **Interactive Query Refinement** with intel imagery. Actively developing this capability and migrating to Project Maven.

Operationalizing AI for Predictive Maintenance (H-60 T700 Engines)

- Train an AI that provides results to users who can quickly approve/reject the results
- Rapidly train the AI to improve performance
- Unsupervised data exploration to generate "candidate questions" that a user may want to ask the AI
- Use model to recommend future questions

Contact: Dr. Juan Vasquez, AFRL ACT3 Product Development Director

MLPerf

Community-wide effort to develop benchmarks for evaluating vendor hardware that represent real-world problems

- 70+ companies including:

AMD	NetApp	Facebook	Baidu	NetApp	Microsoft	VMWare
Google	Lenovo	Dell	Cisco	IBM	Intel	Qualcomm

- 10 universities and research institutes including:

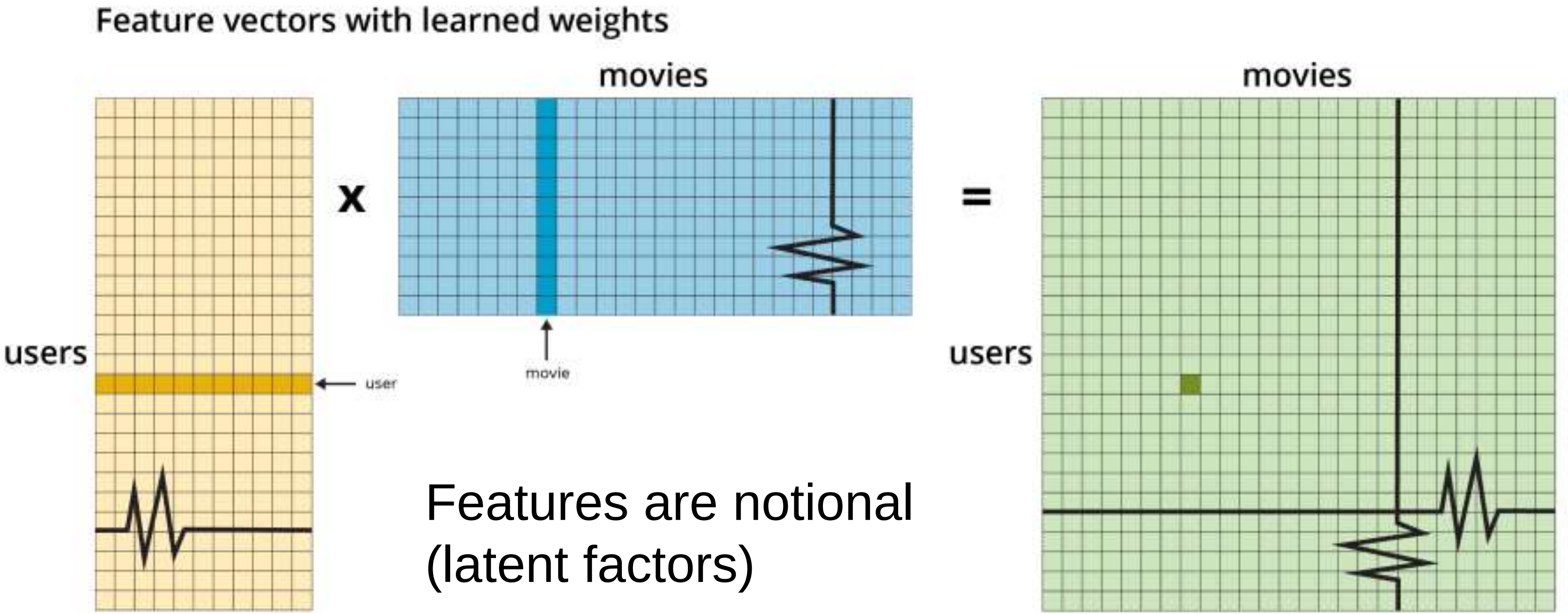
Harvard University	University of Minnesota	University of Illinois, Urbana Champaign	University of California, Santa Cruz
Stanford University	University of Toronto	University of Texas, Austin	University of California, Berkeley

DoD-relevant benchmarks:

- Image classification and object detection
- Natural language processing
- Recommendation systems

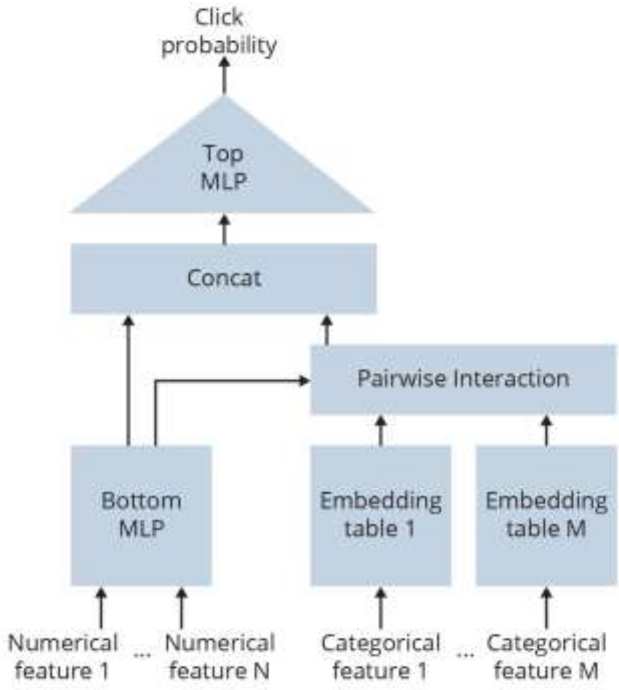
Facebook's Deep Learning Recommendation Model recently added

Feature vectors and latent factors



Features are notional
(latent factors)

Will a User Click on an Ad?

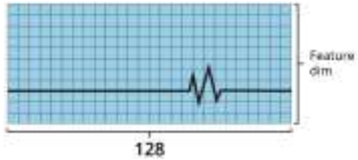


MLP = multilayer percepteron (neural network)

Click-through rate prediction

Users and products represented by **continuous** and **categorical** features

- User represented by a latent factor vector
- Categorical features described by an embedding matrix
  - Different numbers of categories:
    - new, used, in original box
    - sports, music, theater, movies, news, cuisine, …
    - “category” for each individual website
  - 26 Categorical features



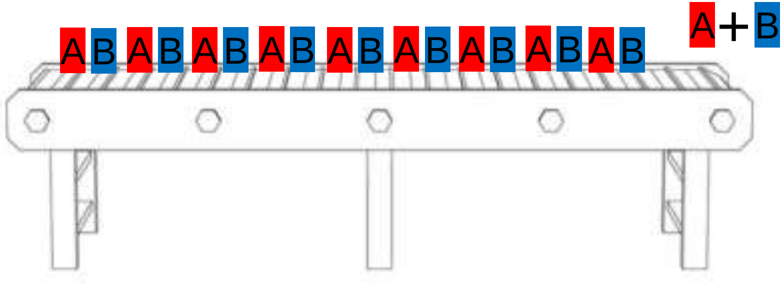
Facebook released Deep Learning Recommendation Model (DLRM) May 31, 2019 https://arxiv.org/abs/1906.00091

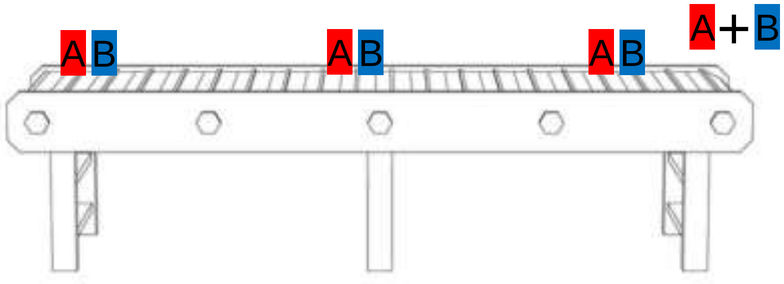
Advanced Computing and DLRM: Relevant ETC Projects

Research Areas in Advanced Computing	Big Learning Benchmarks	Spiral AI/ML	Quantum Computing	DARPA SDH	DARPA DSSoC
Parallelism					
Data-level Parallelism	✓	✓	✓	✓	✓
Model-level Parallelism	✓	✓	✓	✓	✓
Interlayer Parallelism	✓	✓	✓		
Intralayer Parallelism	✓	✓	✓		
SIMD/SIMT Parallelism	✓	✓	✓	✓	✓
Specialized Processing Units					
Vector Cores	✓	✓	✓	✓	✓
Tensor Cores	✓	✓	✓	✓	✓
Application-Specific Integrated Circuits			✓	✓	✓
Data Motion	✓	✓	✓	✓	✓

What is Data Motion?

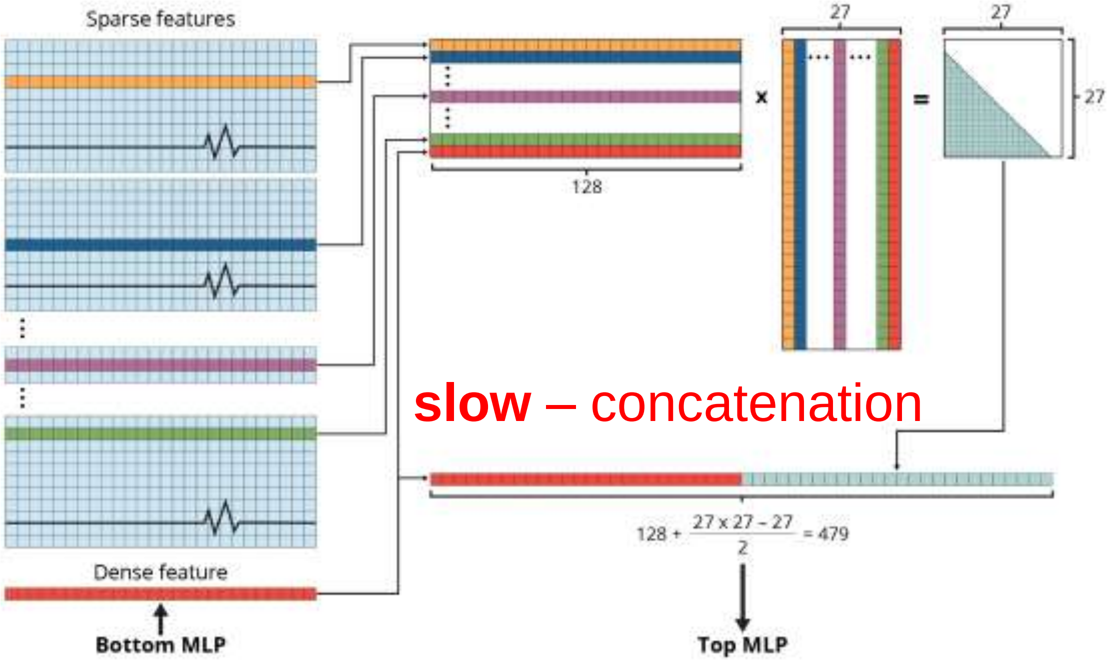
- The most expensive part of any calculation
  - "expense" – time and energy
- Values moved between memory spaces
- Types of memory boundedness
  - **Bandwidth bound** – data pipe is full
    - Can process data much faster than it is delivered
    - Dense, structured workloads (computer vision)
  - **Latency bound** – data pipe is not full
    - Spends time waiting for data to arrive
    - Workloads with random access to data
- Math is ***fast,*** data motion is **slow**





DLRM Piece Parts – Data Motion

slow – DRAM Access

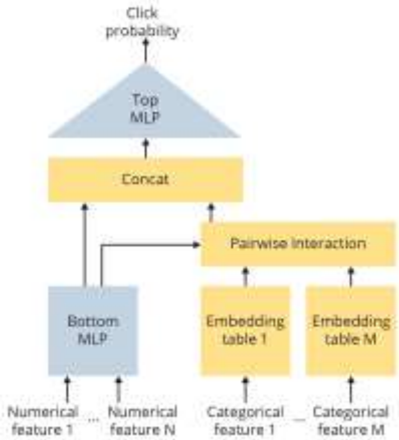


slow – concatenation

fast – Vector Math

fast –Vector Math

fast – Vector Math



CMU SEI Contributions: Spiral AI/ML

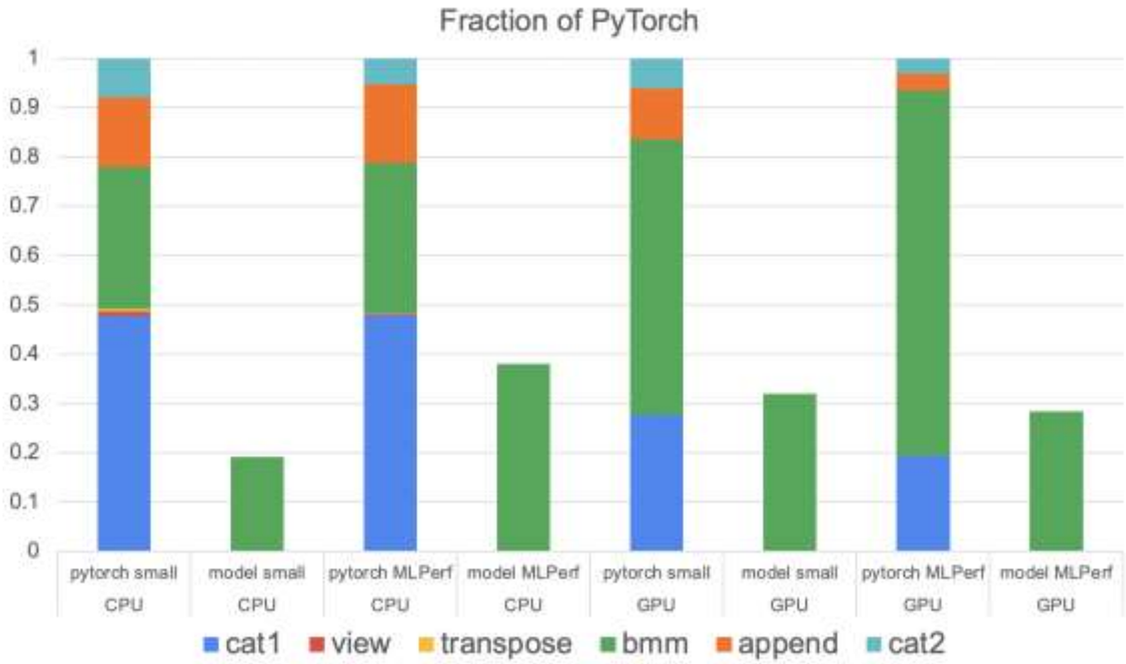
- CMU ECE Prof. Tze Meng Low, student Elliot Binder
- Low's group develops hardware performance models to write optimal code for various platforms
  - Models incorporated into Spiral (Franchetti, CMU ECE) to automatically generate optimal code
- AI models are overwhelmingly implemented in Python frameworks such as PyTorch and Tensorflow
  - Python front ends link to high performance, hardware specific back ends
  - High-level abstractions introduce performance tradeoffs
- Compare performance of model-driven, hand-tuned code with vendor-submitted results to MLPerf

CMU SEI Contributions: Spiral AI/ML (cont.)

- Exploit knowledge of memory systems and frameworks to minimize data motion
  - Block data to make most efficient use multi-way set associative caches
- Eliminate unnecessary framework-induced overheads
  - Fuse operations
- Loops determine data motion
  - Interplay between vector sizes and cache sizes determines optimal ordering
- Effect of optimizations applied to both CPU and GPU implementations
- Present results at conferences to show the best possible performance to the community

CMU SEI Contributions: Spiral AI/ML (cont.)

- Up to five times faster results
  - "bmm" = batch matrix multiply
  - Other components are data motion



Improving Recommendation Systems: Impact on DoD

Financial savings

- Back of the envelope – commercial
  - Hyperscale data center market in 2025: **~$100B**
  - ~10% of data center time spent on recommender systems: **~$10B**
  - 2x faster model would save **~$5B**
- DoD FY21 AI budget proposal: **$841M**
  - DoD will spend **~$100M** on inference and training in coming years
  - Savings with these techniques: **~$10M**

Advanced Computing in the Emerging Technology Center

John Wohlbier

Scott McMillan

Annika Horgan

Jason Larkin

Daniel Justice

Tze Meng Low

CMU ECE

Elliot Binder

CMU ECE

DARPA

- Software Defined Hardware
- Domain Specific System on Chip

Spiral

- Spiral AI/ML
- Spiral Graph

Quantum

- Quantum Advantage Evaluation Framework
- Quantum versus Classical
- Near Term Quantum Computing for Software Verification and Validation

References

[1] Gupta, Udit, et al. "The architectural implications of facebook's DNN-based personalized recommendation." 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2020.

[2] Naumov, Maxim, et al. "Deep learning recommendation model for personalization and recommendation systems." arXiv preprint arXiv:1906.00091 (2019).

[3] Gadepally, Vijay N., et al. "Recommender systems for the department of defense and the intelligence community." Lincoln Laboratory Journal 22.1 (2016).

[4] K.B. Lyons, “A Recommender System in the Cyber Defense Domain,” master’s thesis no. AFIT-ENG-14-M-49, Air Force Institute of Technology Graduate School of Engineering and Management, Wright-Patterson Air Force Base, 2014.

[5] P. Thompson, “Weak Models for Insider Threat Detection,” Proceedings of SPIE, vol. 5403: “Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense,” 2004, pp. 40–48.

[6] T.A. Lewis, “An Artificial Neural Network-Based Decision Support System for Integrated Network Security,” master’s thesis no. AFIT-ENG-T-14-S-09, Air Force Institute of Technology Graduate School of Engineering and Management, Wright-Patterson Air Force Base, 2014.

[7] Polatidis, N., Pimenidis, E., Pavlidis, M., & Mouratidis, H. (2017, August). Recommender systems meeting security: From product recommendation to cyber-attack prediction. In International Conference on Engineering Applications of Neural Networks (pp. 508-519). Springer, Cham.

References (cont.)

[8] Cai, H., & Zhang, F. (2019). Detecting shilling attacks in recommender systems based on analysis of user rating behavior. Knowledge-Based Systems, 177, 22-43.

[9] You, D., Vo, N., Lee, K., & Liu, Q. (2019, November). Attributed Multi-Relational Attention Network for Fact-checking URL Recommendation. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 1471-1480).

[10] Cheryl Pellerin. Project Maven to deploy computer algorithms to war zone by years end. US Department of Defense, 21, 2017.

[11] The JAIC. The JAICs business process transformation mission initiative delivers, 2020.

[12] Kim, D., Park, C., Oh, J., & Yu, H. (2017). Deep hybrid recommender systems via exploiting document context and statistics of items. Information Sciences, 417, 72-87.

[13] Karlsson, L., Bideh, P. N., & Hell, M. (2019, October). A Recommender System for User-Specific Vulnerability Scoring. In International Conference on Risks and Security of Internet and Systems (pp. 355-364). Springer, Cham.

[14] Yang, Z., Sun, Q., Zhang, Y., Zhu, L., & Ji, W. (2020). Inference of Suspicious Co-Visitation and Co-Rating Behaviors and Abnormality Forensics for Recommender Systems. IEEE Transactions on Information Forensics and Security, 15, 2766-2781.