Harvard John A. Paulson
School of Engineering
and Applied Sciences

AI in Assisting Classical Disciplines: Three Suggestions

專家怎麼使用AI協助解決問題: 三點建議

H. T. Kung
Harvard University

北科大菁英講座

Taipei, Taiwan

November 28, 2019

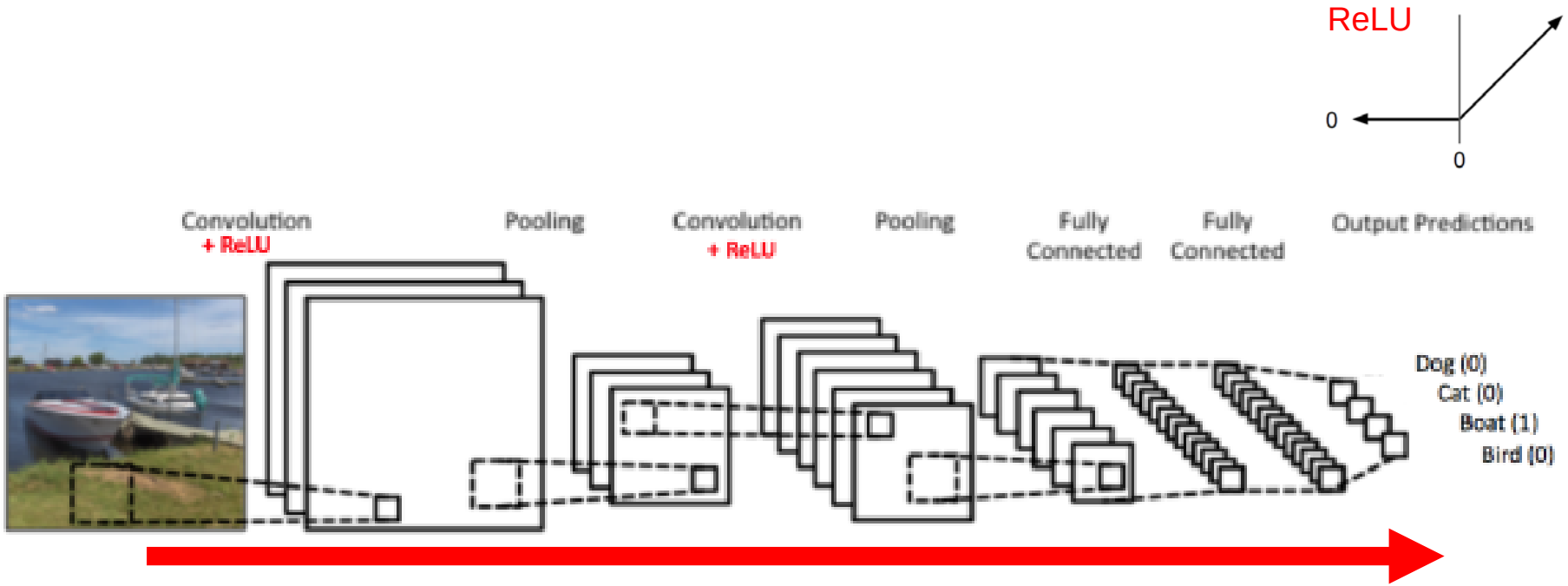
AI and Its Impacts

AI programming is driven by data and learning

AI applications are smart, in addition to automation

- "AI is more profound than **electricity** or **fire**“ --- Google CEO
- "The AI **renaissance**"
- "AI is **transformative**" --- for every aspect of our society
- "AI is **revitalizing** fields" --- e.g., computer architecture, sensing, finance, etc.

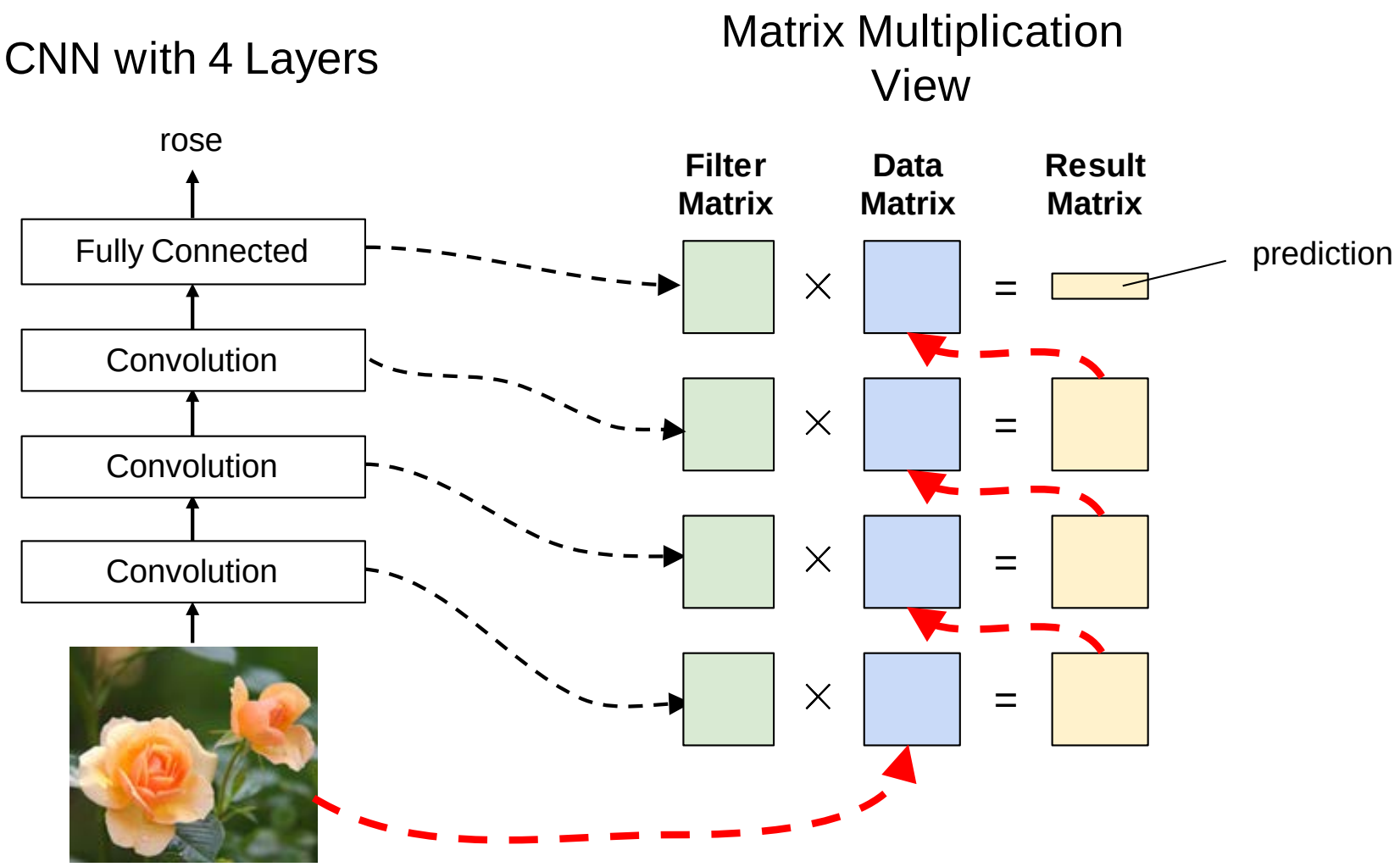
Convolutional Neural Network (CNN): A Workhorse in Deep Learning



(Diagram from text ooks)

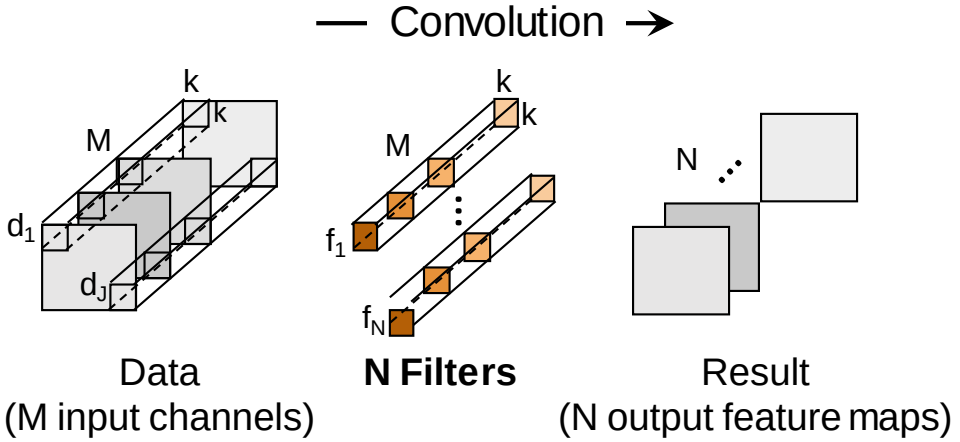
- The computation amounts to a large number of multiplier-accumulator (**MAC**) operations
- Good news: It is mainly **matrix multiplication**, highly optimizable

CNN Feedforward Pass as Series of Matrix Multiplications



Matrix Multiplication in Each Convolutional Layer

Computation of a convolutional layer

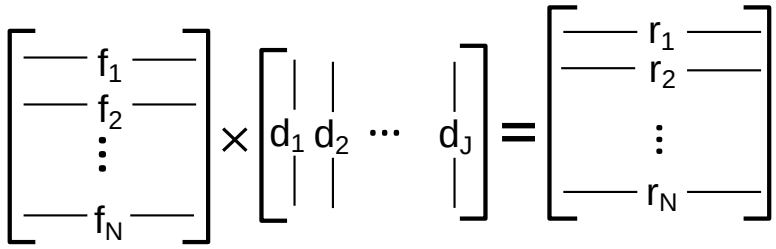


Data
(M input channels)

N Filters

Result
(N output feature maps)

Equivalent matrix multiplication



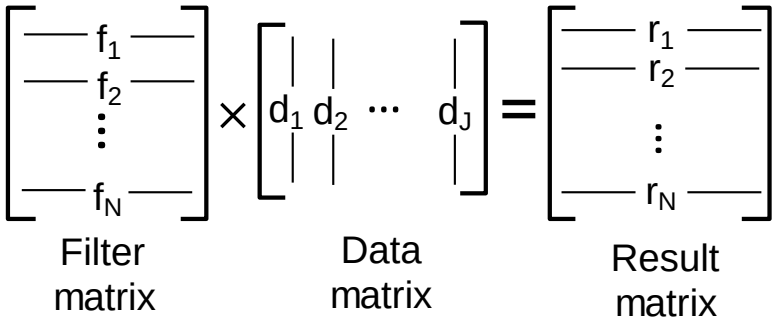
Filter matrix

Data matrix

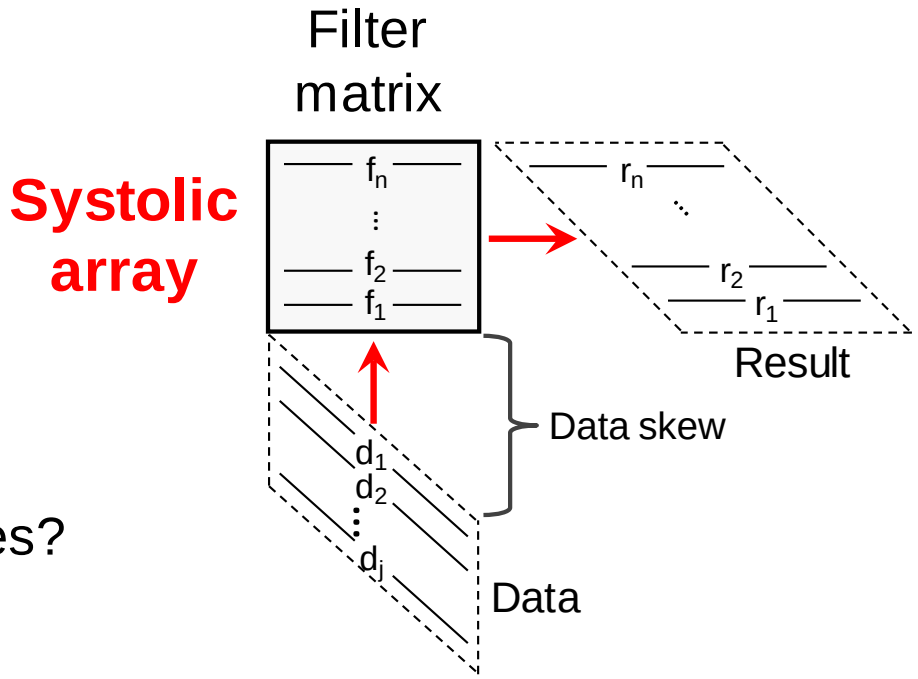
Result matrix

Use of Systolic Array for Efficient Matrix Multiplication

Matrix multiplication



Systolic array Implementation



[Kung and Leiserson 1979] VLSI Processor Arrays

[Kung 1982] Why Systolic Architectures?

High efficiency due to: (1) regular design, (2) data flow architecture and (3) memory access reduction

Why Systolic Architecture (Kung 1982)

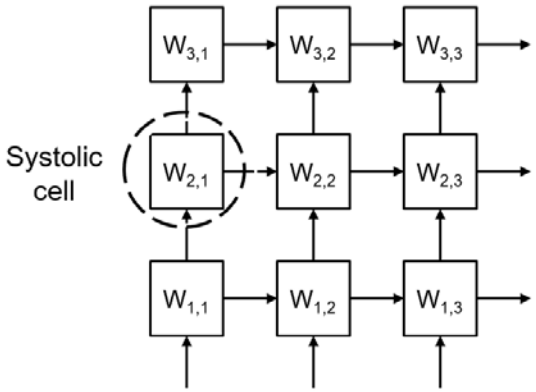
Systolic arrays are efficient in parallel processing:

(1) dataflow architecture

(2) regular layout of processing elements

(3) uniform and **local** inter-processor communication

(4) minimum I/0



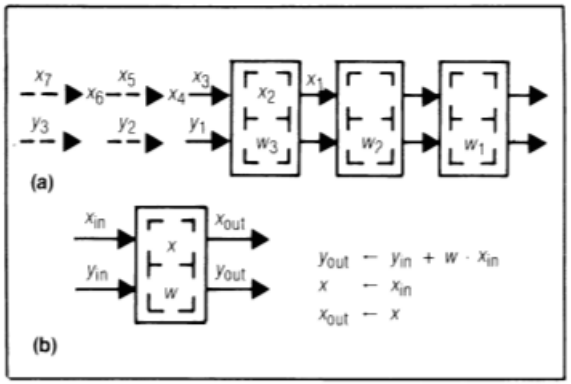


Figure 9. Design W2: systolic convolution array (a) and cell (b) where w_i's stay and x_i's and y_i's both move systolically in the same direction but at different speeds.

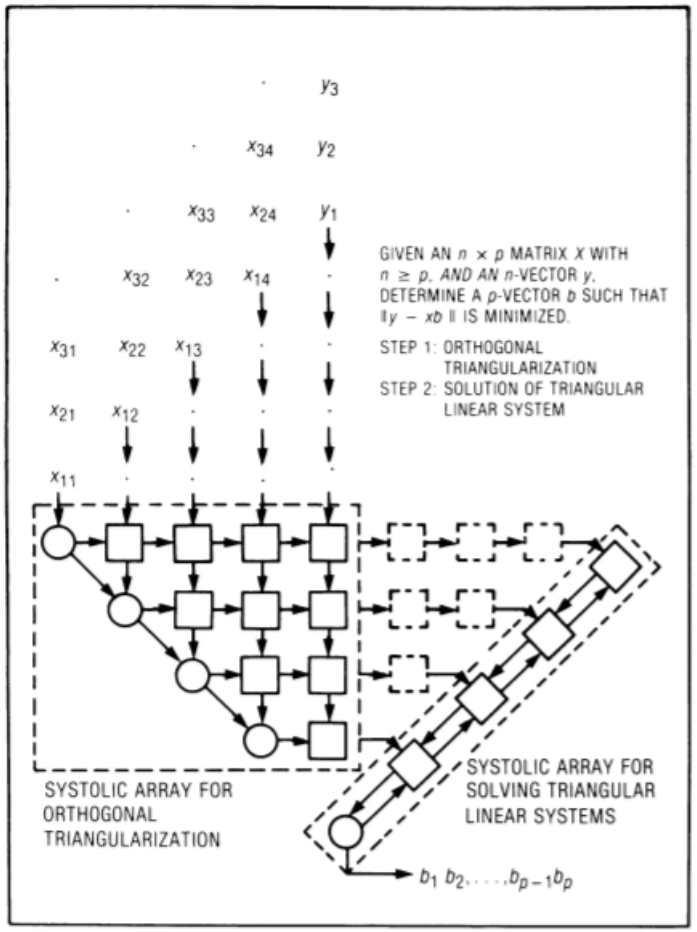
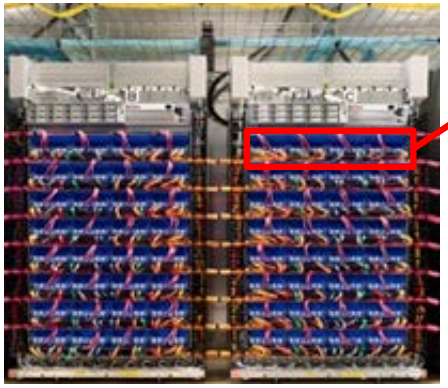
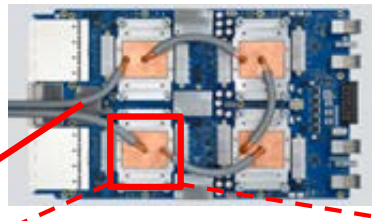


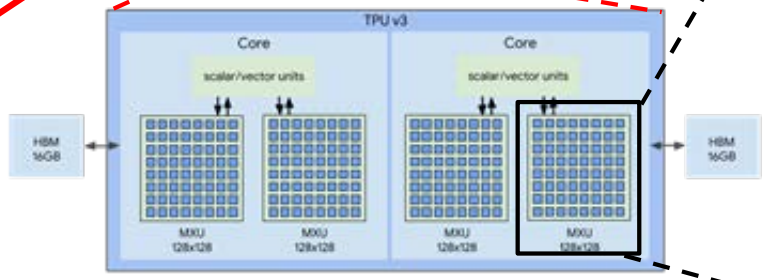
Figure 12. On-the-fly least-squares solutions using one- and two-dimensional systolic arrays, with $p = 4$.

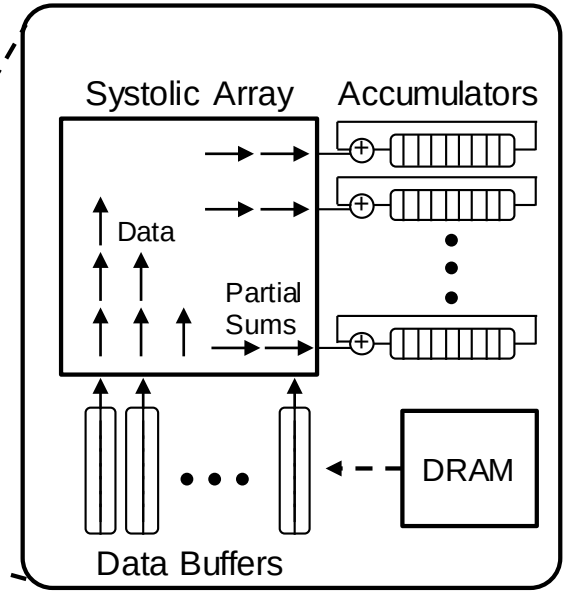
Google Tensor Processing Unit (TPU): Use of Many Systolic Arrays, 2017

TPU v3 in Google Cloud







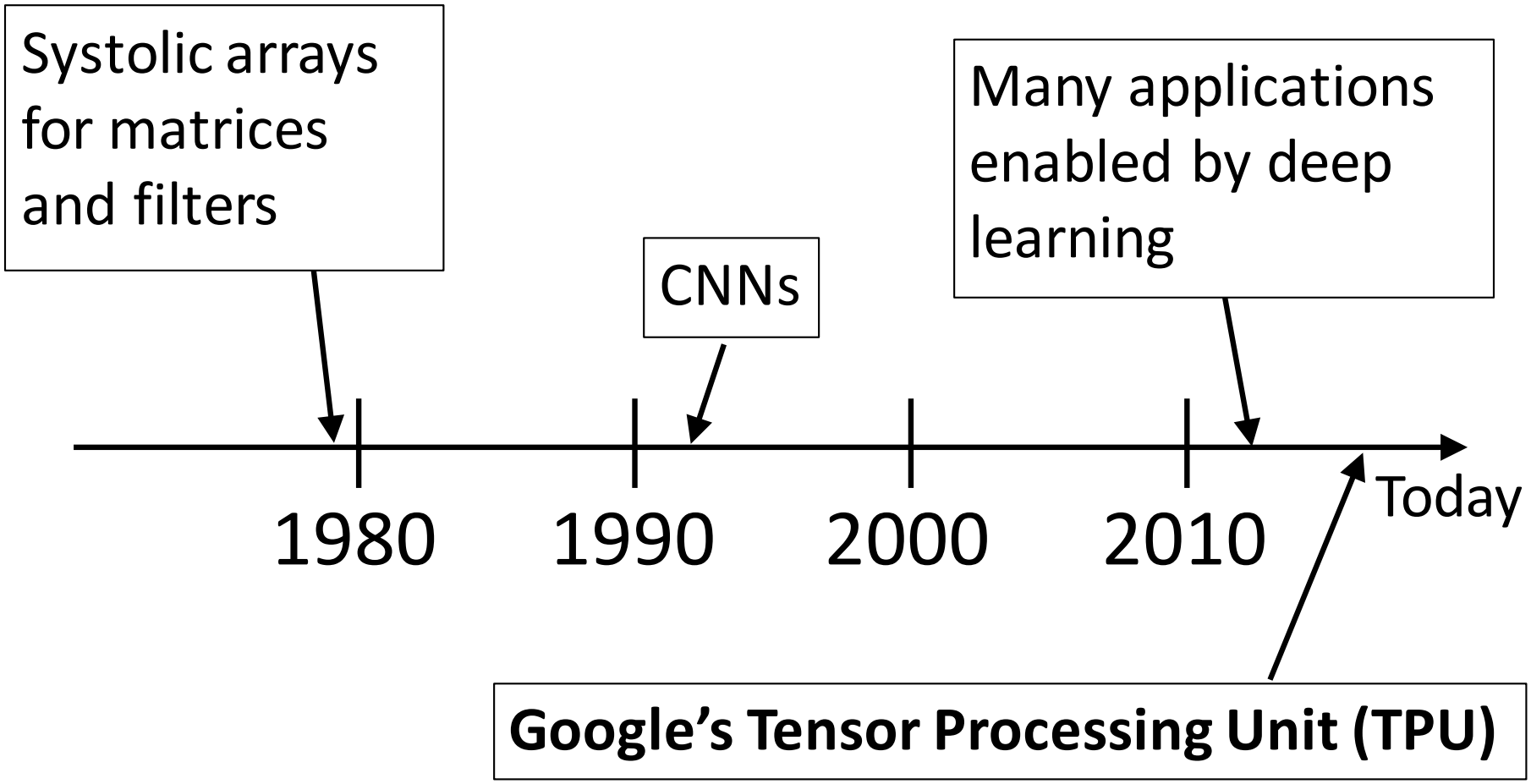


https://cloud.google.com/tpu/docs/system-architecture

Recent **Systolic Array** Based AI Chip/Processor Developments Reported in Open Literature

- Products
  - Alibaba's Cambrian: Shidiannao
  - Bitmain's Sophon BM1680 AI chip
  - Ceva's NeuPro
  - **Google's TPU**
  - Xilinx XDNN FPGA Architecture for AI Inference
- R&D efforts
  - "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," 2017

Back to the Future



1985 Movie “Back to the Future”

Go back thirty years for the future

Moral of the story:

AI is revitalizing many fields; your knowledge and experience may have a new perspective and present new opportunities

You may not want to retire early

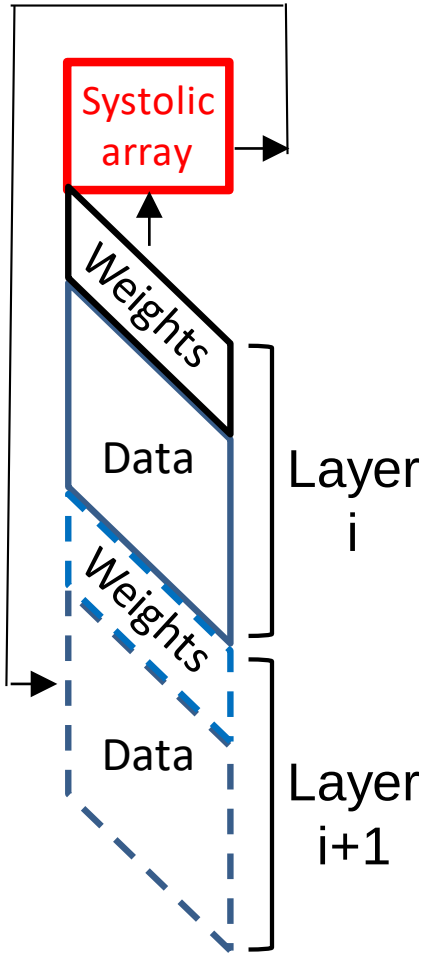
Harvard's 19-layer CNN Inference for ImageNet on a 170MHz FPGA Chip (McDanel, Zhang, Kung, Dong [ICS 2019])

Systolic array based FPGA: **120** images/sec per watt at **2ms** inference latency while achieving state-of-the-art **51%** top-1 accuracy for ImageNet

3x improvements in both **energy efficiency** and **latency**, compared to best prior FGPA implementations, while achieving a similar classification accuracy

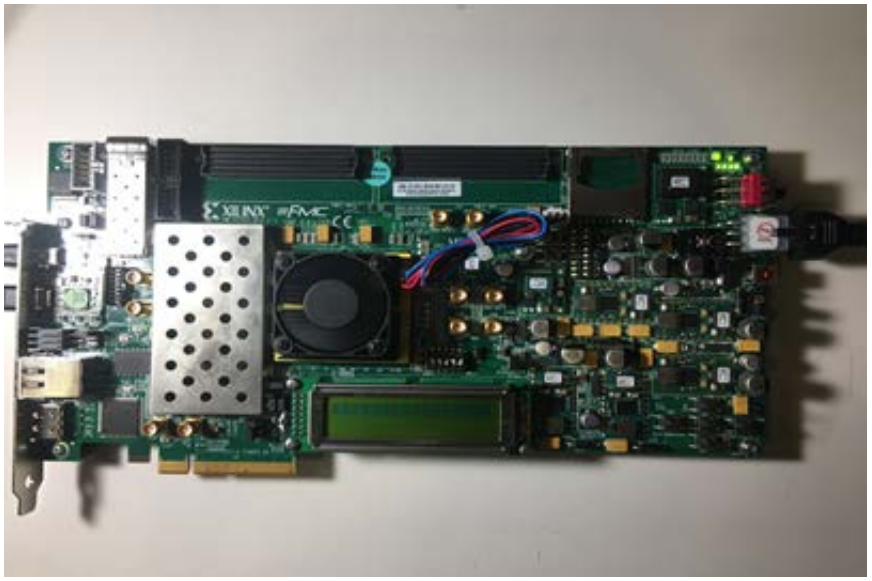
(note that 52% accuracy is the state-of-the-art of mobile CNNs for the ImageNet benchmark)

How? Rest of the presentation will explain



All on-chip

FPGA Implementation of a 128x64 Systolic Array for Inference in Kung's Lab at Harvard



- Xilinx XC7VX485T board
- Total hardware resources: Lookup Table LUT (303600), Flip-Flops (607200), BRAM (1030, each 36Kb)

McDanel, Zhang, Kung, Dong [ICS 2019]

Rest of the Presentation: Sharing Personal Experience in AI Applications and Practice

- Taiwan AI Academy (台灣人工智慧學校)
  - 6K students trained in the past two years; working towards 10K students for end of 2020
- Chang Gung Memorial Hospital (長庚紀念醫院)
- Traditional manufacturing companies
- High-tech companies in Hsinchu
- Governments and agencies
- Start-ups

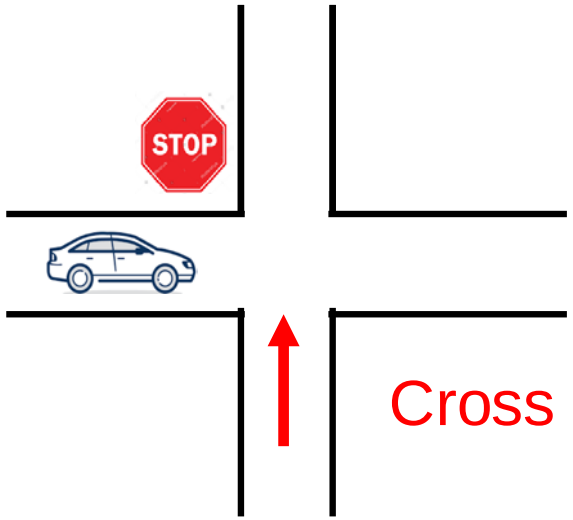
AI in Assisting Classical Disciplines

Using AI in assisting classical fields such as precision manufacturing and health care is of great interest

- Compare new AI-assisted methods with traditional methods
- Domain experts and AI professionals collaborate and learn from each other

- However, success here is a **tall order** (艱鉅的任務)

Why Is It a Tall Order?



Cross or not cross the street?

- A good method is to check if the car driver ran stop signs in the recent past
- It is an individual-level prediction (**specific driver**); statistics from insurance companies (e.g., juvenile has high car accidence rates) will not help
- It is actually a small data problem (AI assisted by on-car IoT devices could help)

Individual-level Prediction as Opposed to Population-level Statistics

Population-level statistics is useful in prediction for a group of subjects

However, it is individual-level prediction that AI is expected to deliver

For example, in patient treatment or machine tuning, we want to make **individual decision** on the specific subject at hand

A Common Theme in Individual-level Prediction: Learning **Delayed Outcome**

Given a subject A, we know its state A1 at time T1 and want to predict its state A2 at some later time T2

A1 at time T1 → A2 at time T2

Ex 1: Treat a stroke patient in emergency room:

- Based on an x-ray image at T1, predict what we may classify on a 3D CT scan image at T2, two hours later

Ex 1: Placement and route a chip:

- Based on a circuit netlist at T1, predict routing performance at T2, 10

Unfortunately, Direct Adoption of Textbook Machine Learning Methods **Seldom** Yields Competitive Results in Real-world Applications Where Prediction Accuracy and Reliability of the Technology Are Important

It is fortunate in the sense that otherwise high school kids could have taken over our jobs!

For Individual-level Prediction, We Need Specifics about the Subject as Well as Context ("Meta Data")

- **Specifics** of the subject at hand, such as the patient's age and sex, medical history and lifestyle
- **Context** about the subject, such as support environment for a patient, interactions with other factors (e.g., other drugs), cleanness of the manufacturing equipment, etc.

If one would use rules to predict outcomes, there would be too many rules to manage

Infinitely Diminishing Dataset

- To train a model, we need enough labeled **good data**, which are correct, up-to-date, balanced, not biased, and safe (no malicious content)
  - Spoiling a specific prediction seems to be much easier than making good inference in general
- But, most importantly, the data must be **relevant** to the specific subject and context at hand
- The intersection of data satisfying a expanding number of conditions could be **exponentially small**
- What should we do in this case?

There Has Been "Magic": Domain Experts Can Capture Cues and Make Context-aware Decisions

- Domain experts can often notice **important signal** in data and make sensible decisions for the given context
- In response to the same sensor reading, an expert may subscribe to different treatments depending on the **situation**
  - Turn the control knob in the same direction but at a different speed

Collaboration between Domain Experts and AI Professionals: Necessary But Challenging

- Most of us now believe that AI-assisted approaches can potentially have a huge impact on a domain (no organization can afford ignoring this)
- But domain experts cannot really keep up with rapid advances in AI (can anyone these days?)
- The IT department with AI know-how may now have to take on a **new role** of providing AI assistance to domains
- But hand-holding assistance cannot scale

Suggestion I: Develop Domain-specific AI Software and Tools (Infrastructure)

- Infrastructure such as libraries is essential in deployment of AI models and algorithms
  - E.g., MATLAB has made linear algebra easy to use
- Domain-specific AI software and tools
  - Physics-based regularization to prevent model overfitting
  - Incorporation of analytic models and simulators
    - E.g., Apple's gaze detection work is largely assisted by the UnityEyes simulator
  - Automatic model parameter tuning for nonlinear transformers
    - E.g., SVM kernel tricks
  - Data acquisition and processing
  - Interpretation of results

Suggestion II: Establish Collaboration Frameworks between Domain Experts and AI Professionals

Standardize collaboration **processes** and provide **templates**

- Topic definition
- Approaches to adopt
- Expectation setting
- Agreements on mutual responsibilities, schedule, progress review, ownership, and IP
- Deployment in **workflow**, etc.

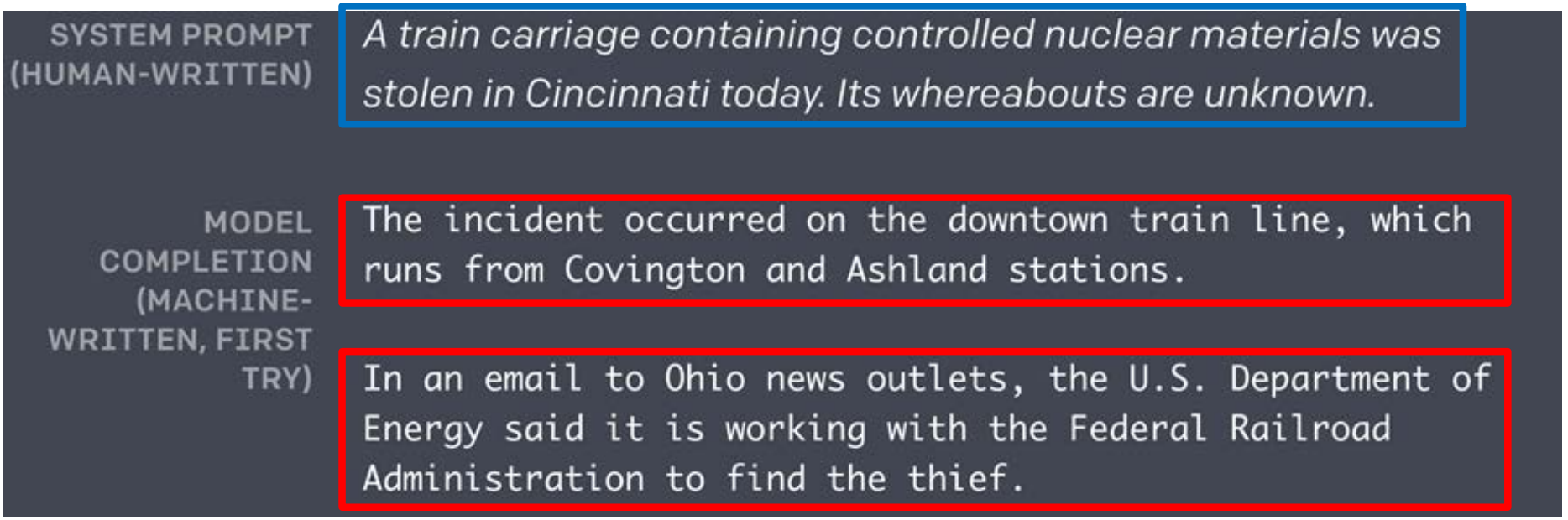
- Developing such collaboration frameworks is a **new agenda** for industry; it requires careful design, e.g., lining up **incentives** of each party
- Having workable frameworks is critical for **sustaining** and **scaling** collaboration
- Ideally, we could develop an **assembly line** for the scalable development of AI applications

Suggestion III: Develop Novel Domain-specific Approaches

- Any successful AI-assisted approaches must be grounded in **physical insights** and **logical reasoning**
- Domains usually have **abundant unlabeled data** but at present AI is not using them well
- Based on domain insights and reasoning, we may, for example, come up **domain-specific methods** to exploit these unlabeled data

To Illustrate, Consider “Deep Fakes”

Genuine sentence



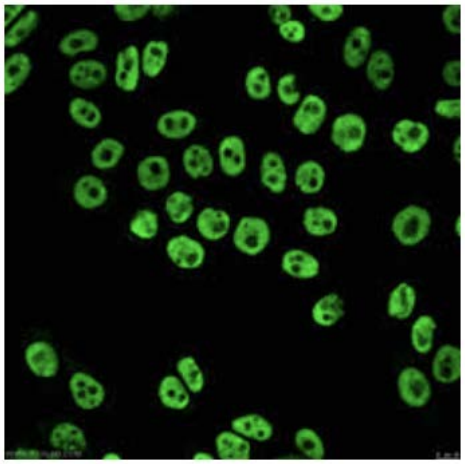
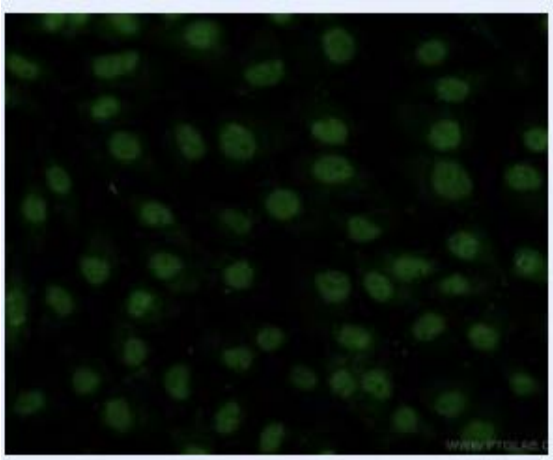
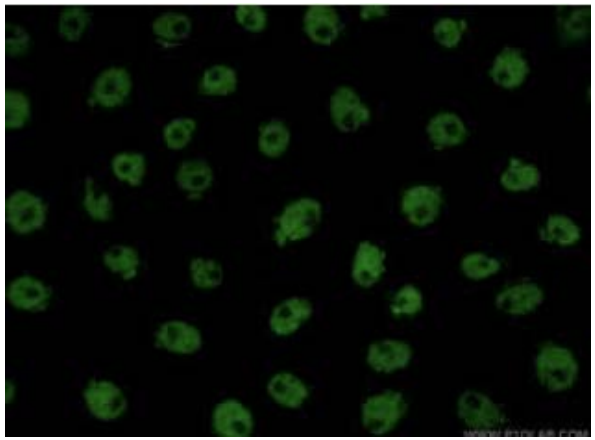
Generated sentences

“Politicians may want to consider introducing penalties for the misuse of such systems”

Generated by GPT-2 from OpenAI, 2019

Self-supervising Learning Using Unlabeled Data

- In natural language processing (NLP), such as "deep fake", we learn semantics of texts from texts themselves by filling in blanked out texts/sentences
- Such self-supervising learning may apply to other applications. E.g., can we learn to classify cell images below, by themselves **without human labeling**?

General-purpose Feature Extractor Obtained by Self-supervising Learning

- In NLP, we have seen success of using self-supervising to learn a large word embedding model using billions of parameters (e.g., BERT)
- We may then use this pre-trained model as a general-purpose feature extractor for various tasks such as language translation and topic classification

Hyper-Dimensional General-purpose Feature Extractor (e.g., Dimensionality 10,000)

- We could randomly map datapoints into a hyper-dim space ("reservoir") to ease their separation
- A large reservoir could behave like a large pre-trained feature extraction for various tasks
- While large in their dimension and size, these hyper-dim feature extractors are simple in their computation processes (lots of dot products)
- Emerging **analog computing**, which can compute a dot product for feature matching in one clock cycle, could be applicable here

Exploiting Bit-Level Sparsity for Efficient Hardware Implementation of Dot Products

- To illustrate, consider a number 95
  - Binary representation: 01011111
    - $95 = 2^6 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0$
  - Powers-of-two expression
    - $95 = 2^6 + 2^5 - 2^0$
- Multiplication is less costly now:
  - $95 = 2^6 + 2^5 - 2^0$
  - $23 = 2^4 + 2^3 - 2^0$
  - $95x23 = (2^6 + 2^5 - 2^0) \times (2^4 + 2^3 - 2^0)$

    9 additions of exponents only!

Harvard John A. Paulson
School of Engineering
and Applied Sciences

TOODL:
Transformer on Optical Delay Loop

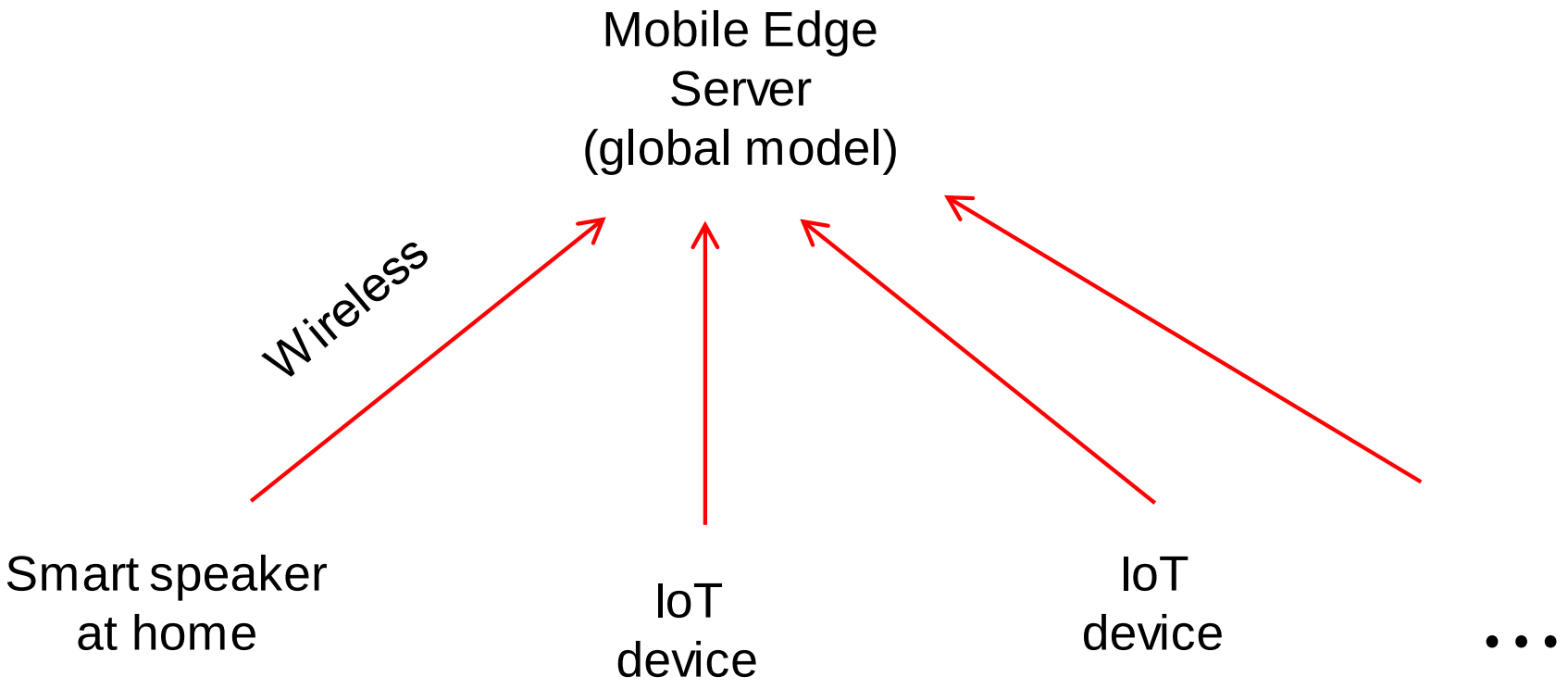
H. T. Kung

Harvard University

DARPA PEACH Program Review Meeting

October 17, 2019

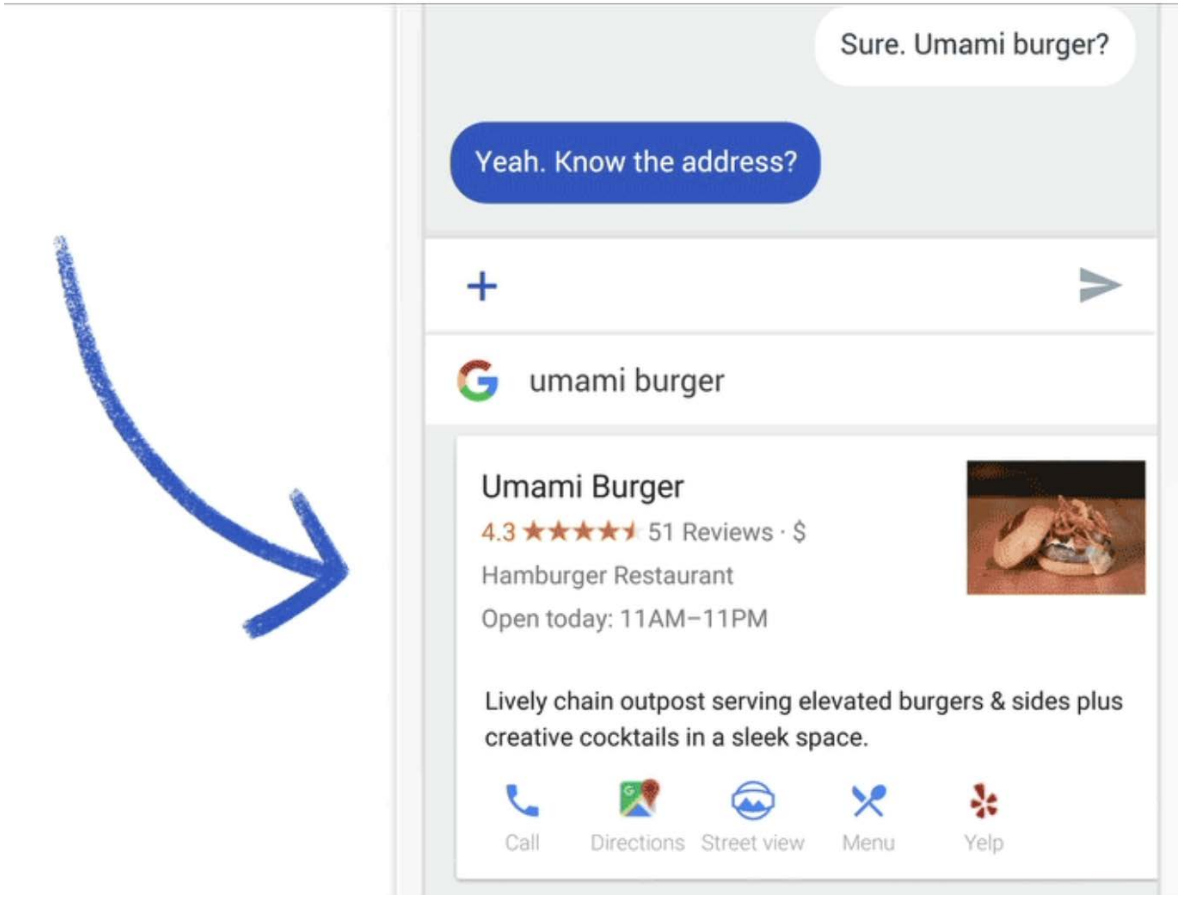
Federated Learning



Global model is continuously updated with local models, reflecting learning at devices with local data

Advantages: data privacy and personalization

Web Query Suggestion on Gboard

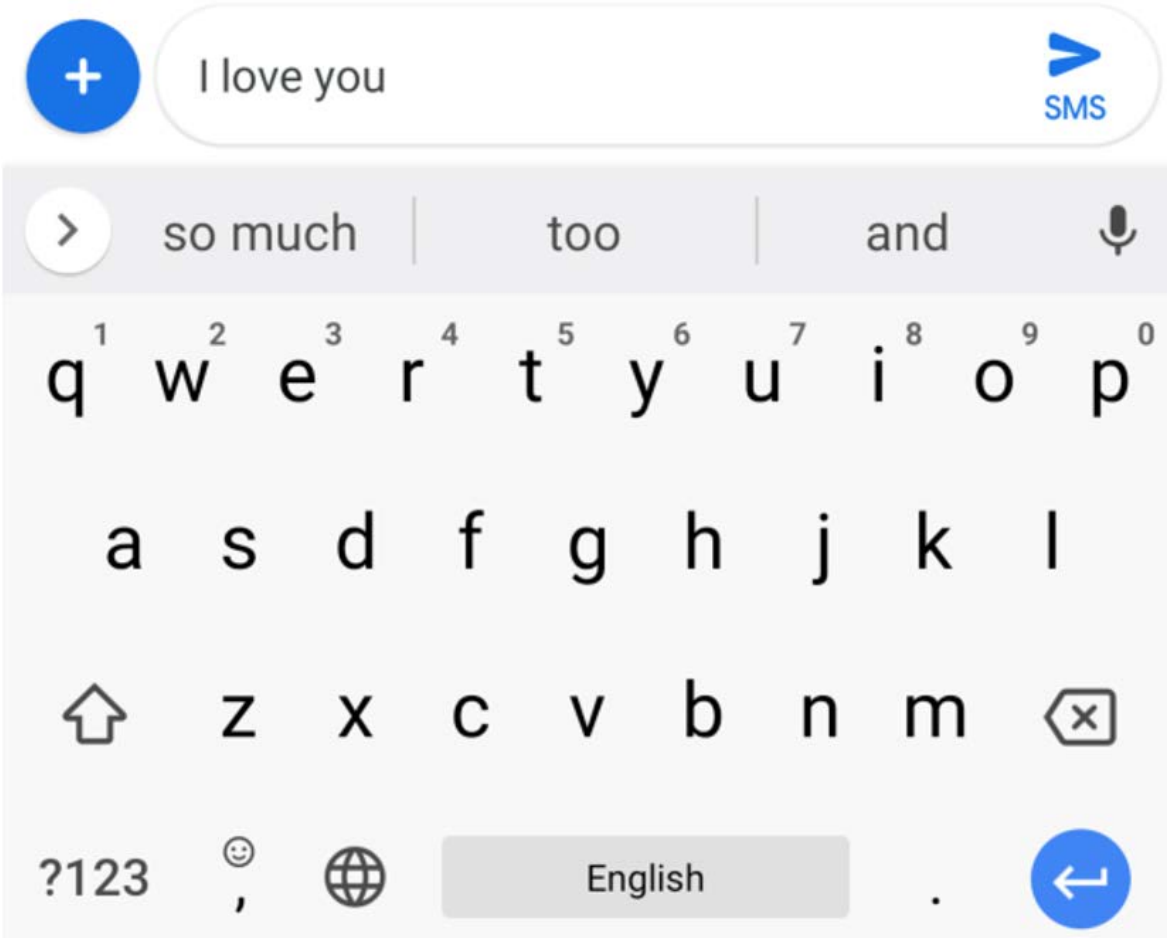


Traditionally, this is a server side application

We could use federated learning instead

- User keeps and manages **personal** data (can erase data!)
- Can leverage behavior learned from **other users**
- Low-latency, energy-efficient and accurate **local** inference (e.g., low-shot learning with imprinted weights)

Next Word Prediction on Gboard



Collaborative Model Learning

- **Federated learning**: Global model update based on model updates from local users
- **Kaggle competition**: Contestants competing to increase the accuracy of their models on a fixed test dataset
- **Blockchains**: Distributed ledger based on the Proof-of-Work (PoW) protocol to maintain a globally consistent order of transactions

Can we have a collaborative model learning over peer-to-peer blockchain? It will be a **decentralized** system which offers **incentives** to users who participate in model improvements (see next slide)

How To Hide Test Labels from Peer Verifiers in the Blockchain-based Model Collaboration?

- We have developed a *learnable* **Distance Embedding for Labels** (DEL) function (Teerapittayanon and Kung [IEEE Blockchain 2019]), specific to the test label vector for the classifier in question
- DEL embeds the label vector (e.g., dimensionality **10,000** for MNIST) inferred by a classifier in a low-dimension space (dimensionality **256**) where its distance to the test label vector is approximately preserved
- DEL allows peers to verify model quality without revealing to them test labels

Conclusion

To advance individual-level prediction, we suggest:

I. Develop domain-specific AI **software** and **tools**
II. Establish collaboration **frameworks** and **templates** between domain experts and AI professionals
III. Devise novel **domain-specific approaches**, such as feature extractors using unlabeled data and collaborative model learning

These efforts are **fundamental**. They could redefine traditional disciplines

New methodologies such as **AI app deployment assembly** could have a huge potential; let's **grab** these opportunities