The bridge to possible

AI/ML for Network Engineers

John Capobianco *Technical Leader*

Patrick LeMaistre *Solutions Engineer*

Dave Zacks *Distinguished Engineer*

BRKENT-2209

#HighBitRate

Cisco Webex App

https://ciscolive.ciscoevents.com/ciscolivebot/#BRKENT-2209

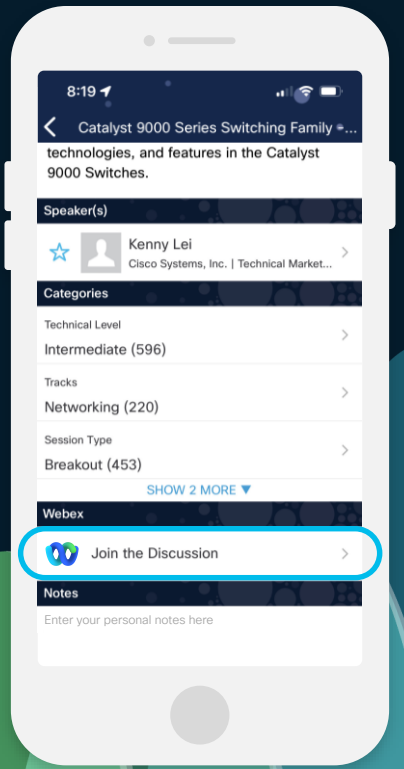
Questions?

Use Cisco Webex App to chat with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 7, 2024.



Agenda

- **Introduction**

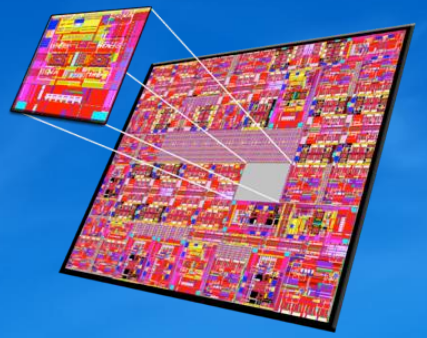
  AI/ML, Generative AI,
  Neural Networks, Transformers …

- **What is Cisco doing with AI/ML?**
  - Using AI/ML in Solutions
  - Building Networks for AI/ML

- **On-Prem AI Solutions**
  - Open Source LLMs w/ UCS and more
  - Demo Time!
- **Leveraging LLMs for Networking**
  - LLMs Deeper Dive
  - Working with LLMs
  - Demo Time!

- **Wrap-up**

by way of **Introduction ...**



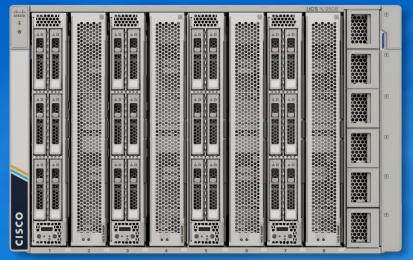
I am a **Distinguished Engineer** in the Network Transformation team, and have been with Cisco for 24 years.

I work primarily with large, high-performance Enterprise network architectures, designs, and systems. I have over 39 years of experience with designing, implementing, and supporting solutions with many diverse network technologies.

I have a strong background in, and focus on, customer requirements, and integrating these into the products and solutions Cisco builds.
I have a special interest in Flexible Hardware, Fabrics, Assurance and ML/AI.

Dave Zacks
Distinguished Engineer
dzacks@cisco.com

by way of **Introduction ...**



I am a **Compute Solutions Engineer** in Cisco's Cloud Infrastructure and Software Group. I have been with Cisco for 20+ years focused on the areas of AI, high performance computing, cloud, mass-scale infrastructure and mobility.

I am the author of the AI Toolkit for UCS X-Series and C-Series which can be found on GitHub at: www.github.com/pl247/ai-toolkit

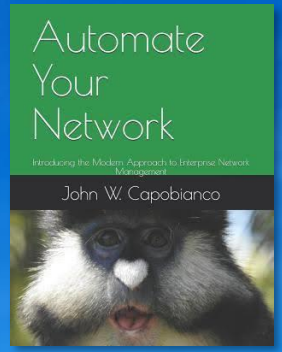
Prior to Cisco, I was co-founder and Vice President of Network Operations for Escape Communications, and hold a patent pertaining to Data Center Ethernet.

Patrick LeMaistre
Solutions Engineer
plemaist@cisco.com

by way of **Introduction ...**

... and for you Tolkien geeks out there ...

John Capobianco and Tom Bombadil – brothers separated at birth? ☺



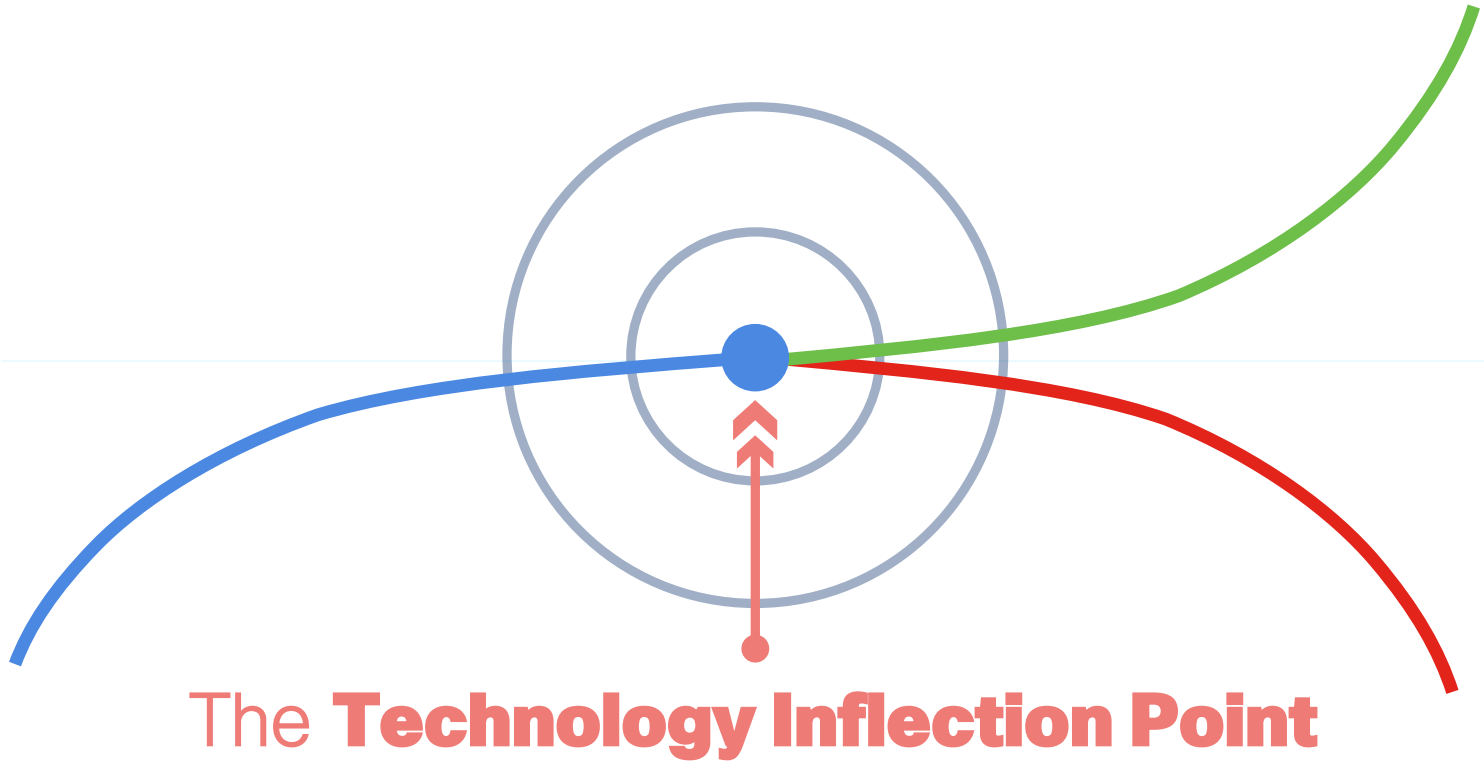
I am a **Technical Leader** in the Cisco Machine Learning / AI team. I have 20+ years experience in the IT industry, and most recently before Cisco worked as a network architect for the Canadian Parliament.

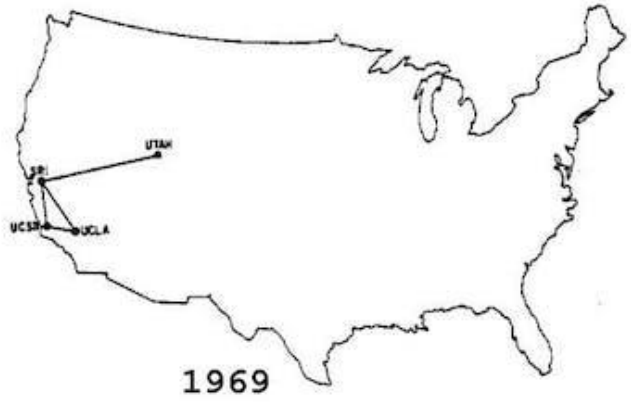
I have previously presented on both network automation (at Cisco Live 2015), Ansible (2015 – 2019), and last year on AI and LLMs at Cisco Live and other events worldwide.

I have a self-published book “Automate Your Network”, available on Amazon (2019).

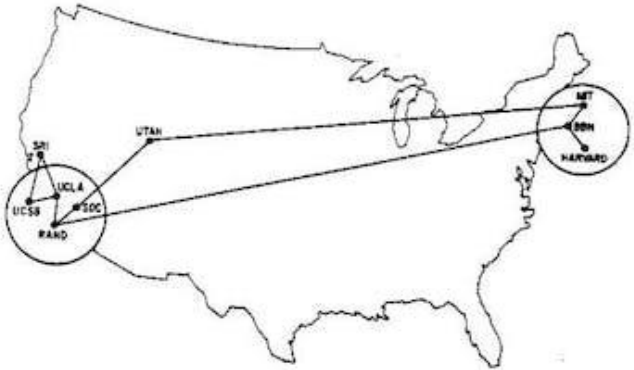
John Capobianco
Technical Leader
johcapob@cisco.com

Rapid Change doesn’t come in normal times

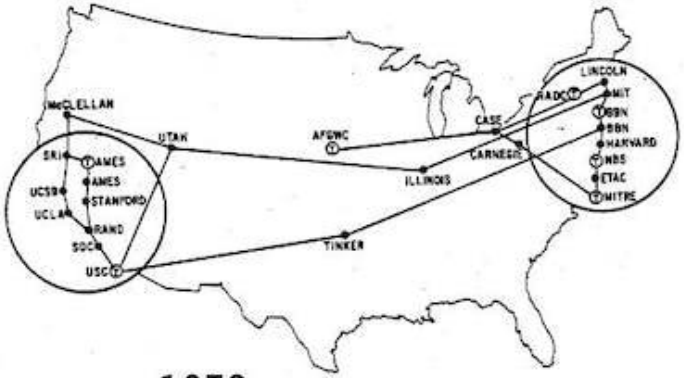




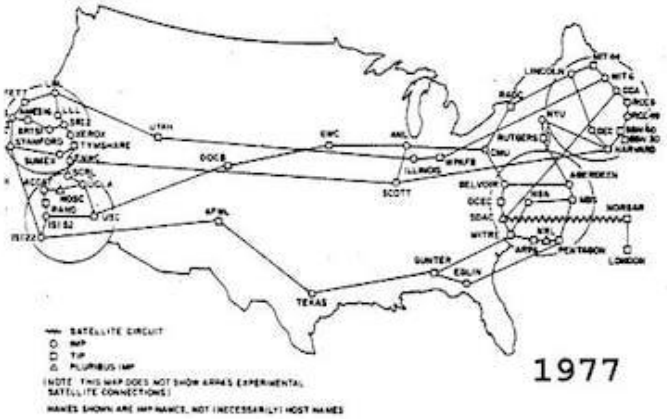
1969



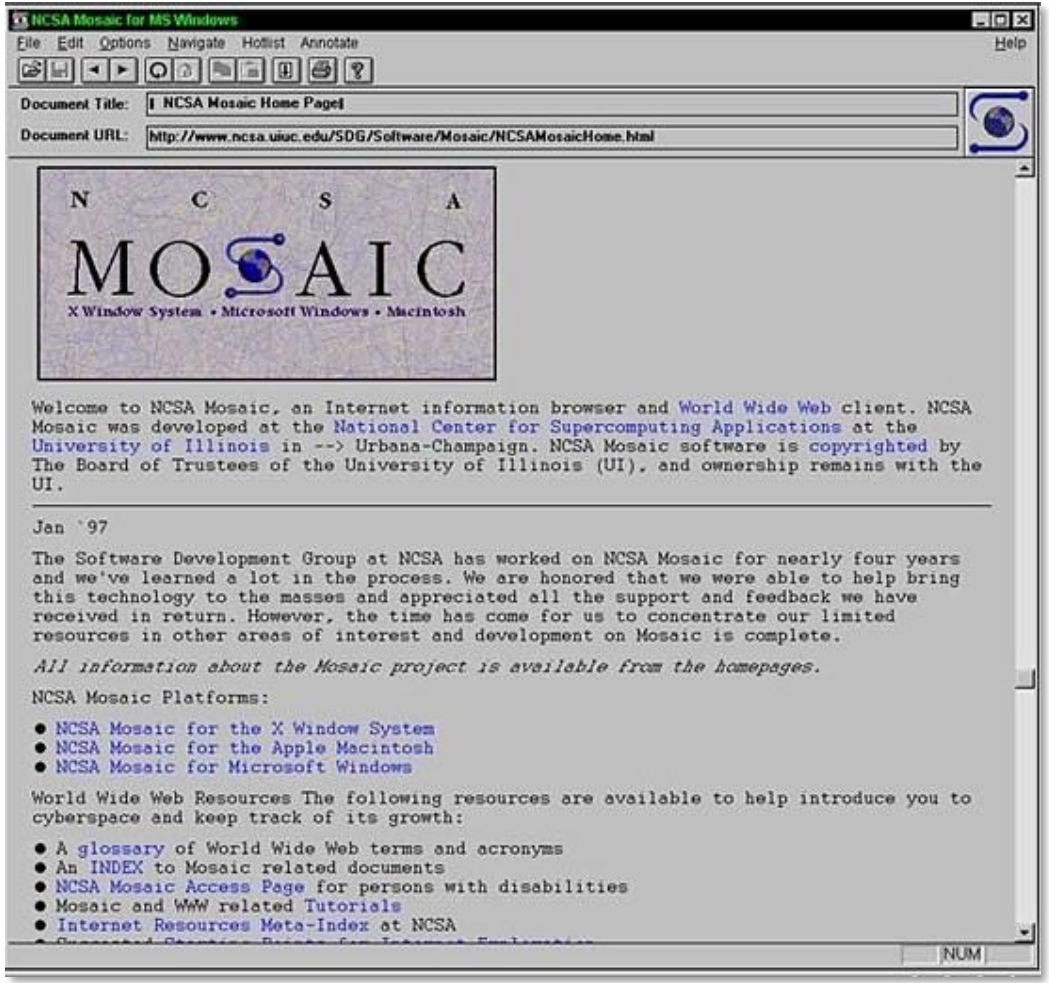
1970



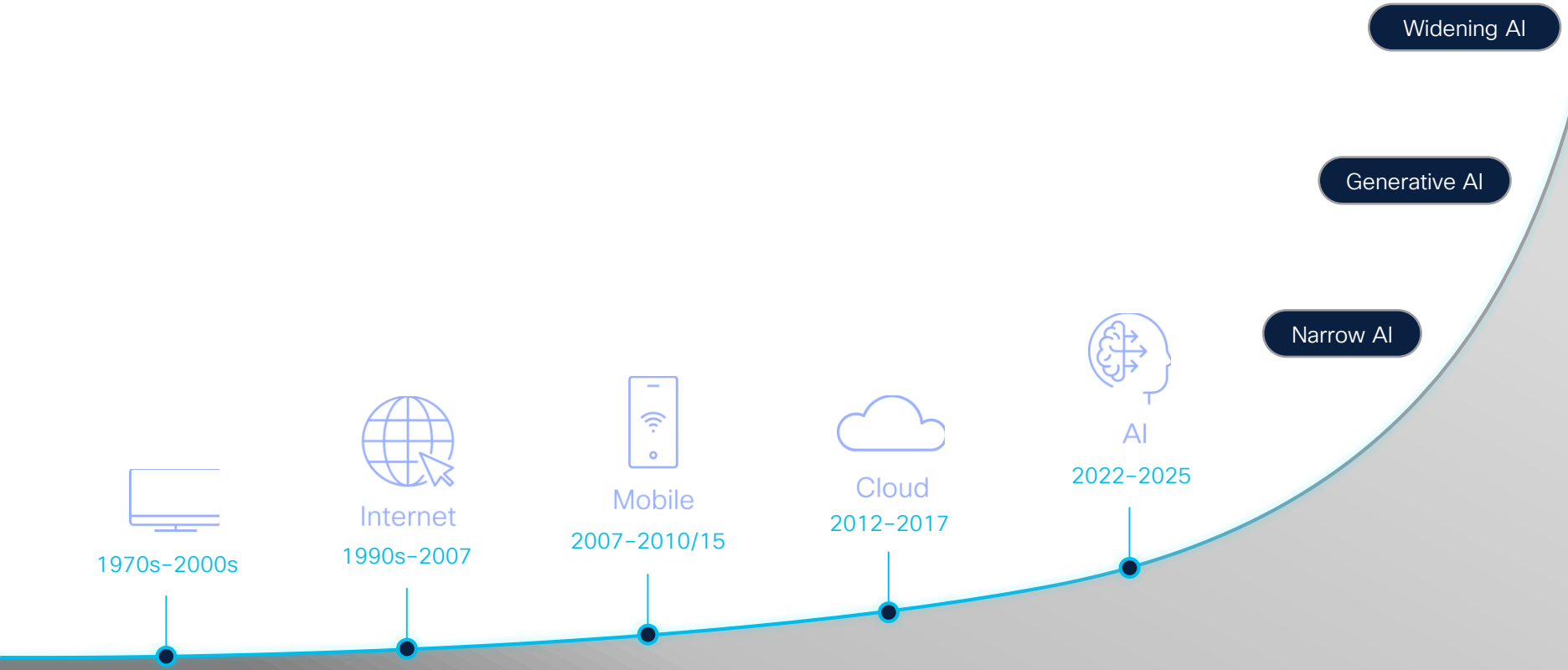
1972



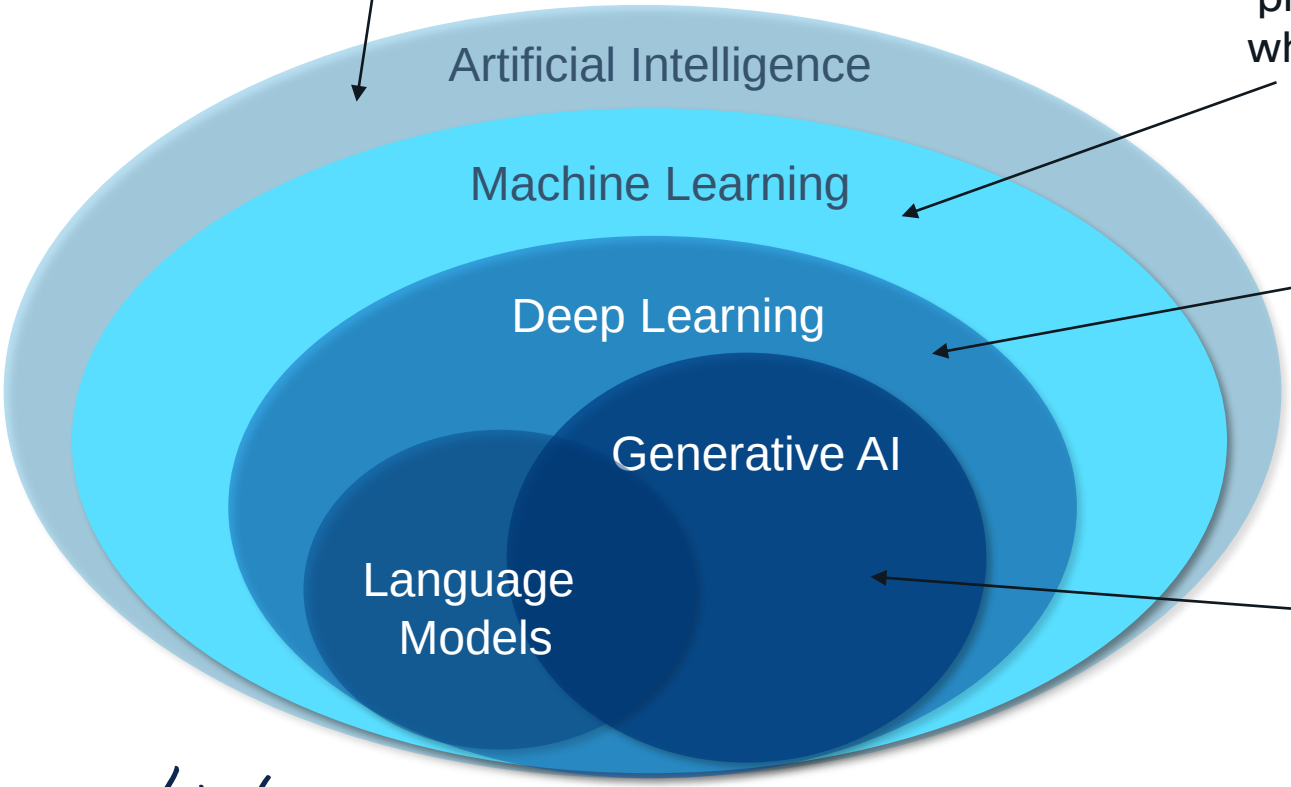
1977



Generations of Technology Inflection Points



AI encompasses the whole world of ML and Deep Learning



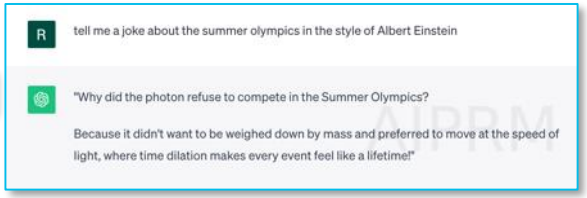
An AI technology where the rules are not set in the program, but are learned while the program is used

A form of ML that uses Neural Learning Networks

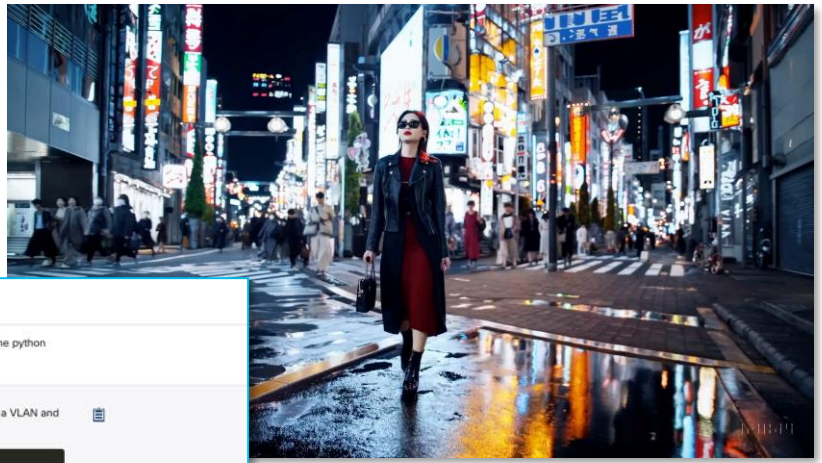
A powerful mechanism that allow neural networks to learn language and generate content

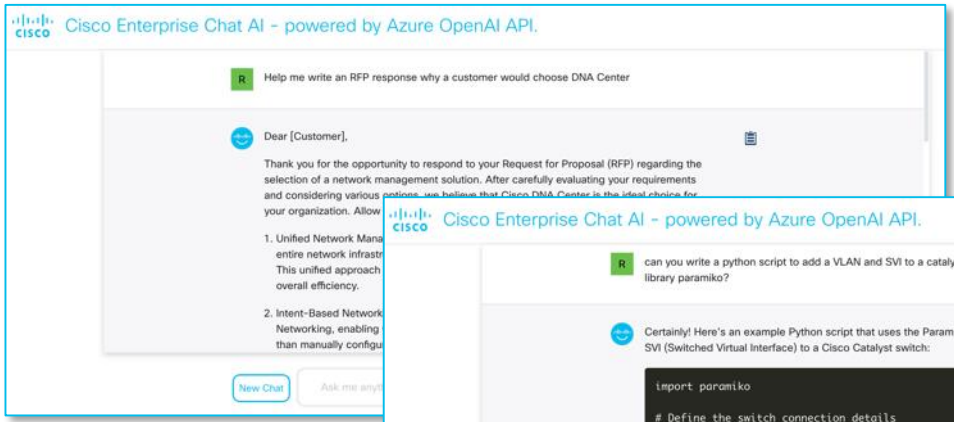
A Snapshot of GenAI Capabilities …

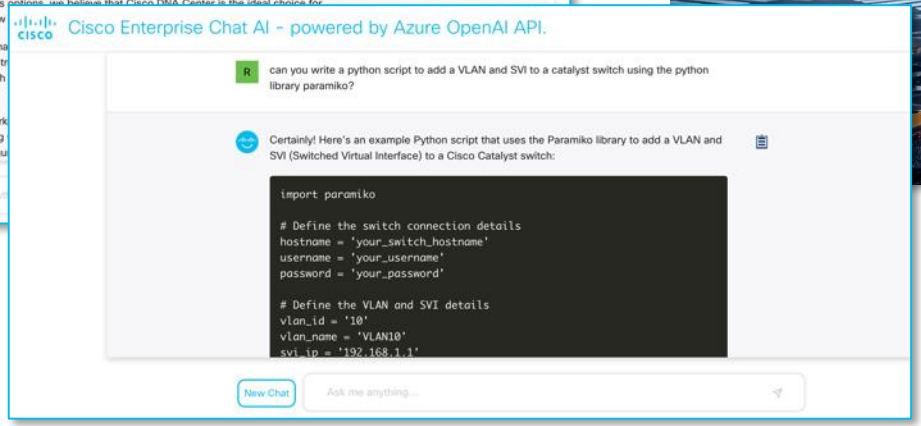
Casual / Fun



Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.







BUSINESS – Answering RFPs, Writing code …

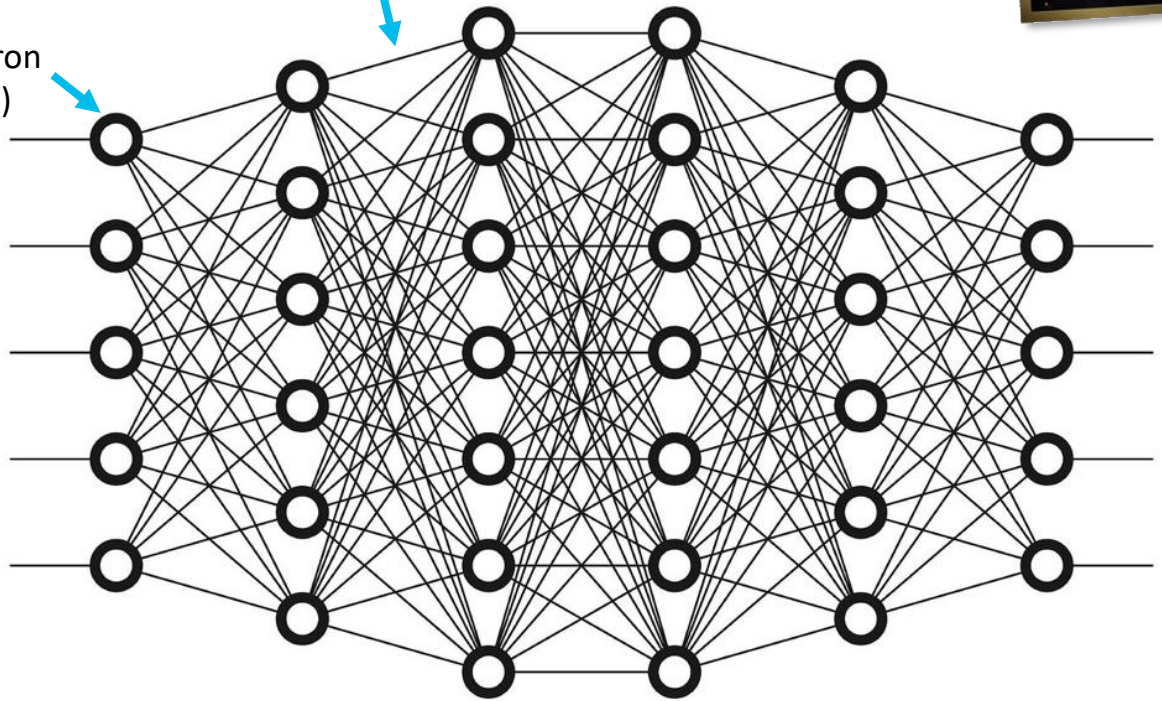
MULTIMEDIA – Text, Pictures, Video, …

Why is this happening **now?**

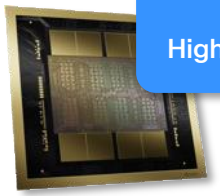
Scale

Parameter
(Synapses)

Perceptron
(Neuron)



NVIDIA
Blackwell



Advances in Silicon –
High-density, High-performance GPUs

NVIDIA Flagship Accelerator Specification Comparison

	B200	H100	A100 (80GB)
FP32 CUDA Cores	A Whole Lot	16896	6912
Tensor Cores	As Many As Possible	528	432
Boost Clock	To The Moon	1.98GHz	1.41GHz
Memory Clock	8Gbps HBM3E	5.23Gbps HBM3	3.2Gbps HBM2e
Memory Bus Width	2x 4096-bit	5120-bit	5120-bit
Memory Bandwidth	8TB/sec	3.35TB/sec	2TB/sec
VRAM	192GB (2x 96GB)	80GB	80GB

Geoffrey Hinton – the "Godfather" of Deep Learning

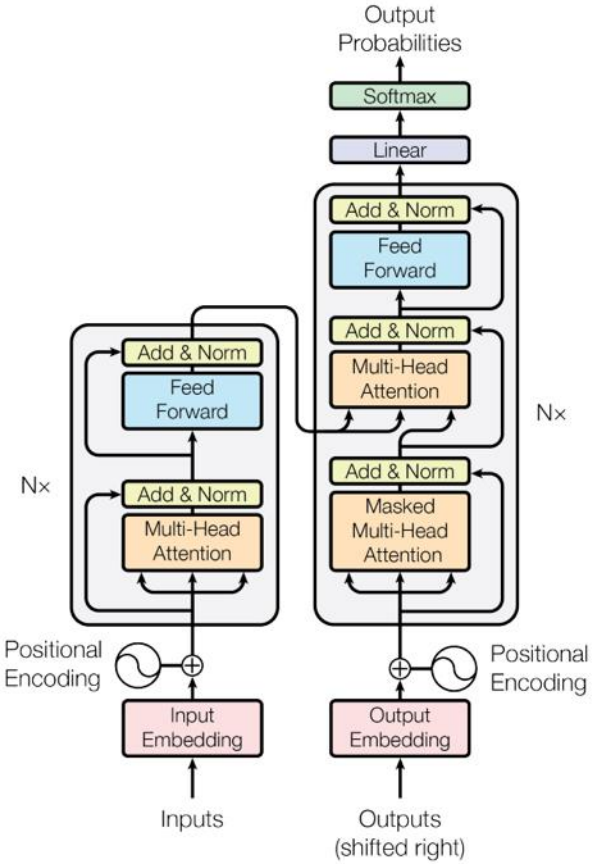


Figure 1: The Transformer - model architecture.

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

arXiv:1706.03762 **[cs.CL]**

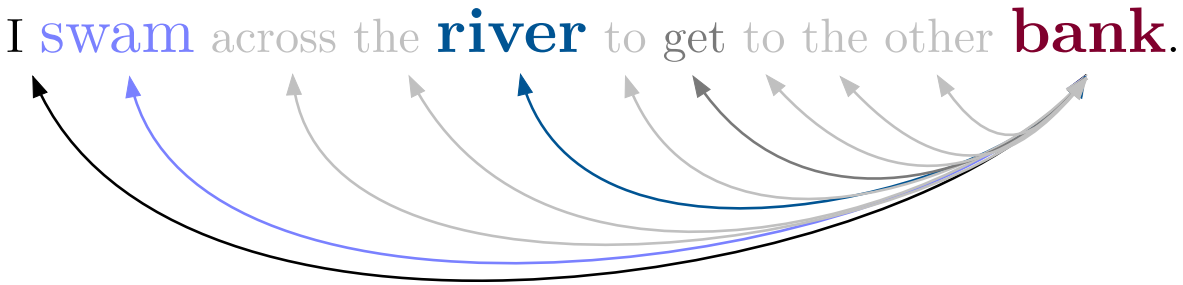
Attention Mechanism – Overview

Foundation for Transformer Models

You have no problem interpreting "bank" in the following sentence:

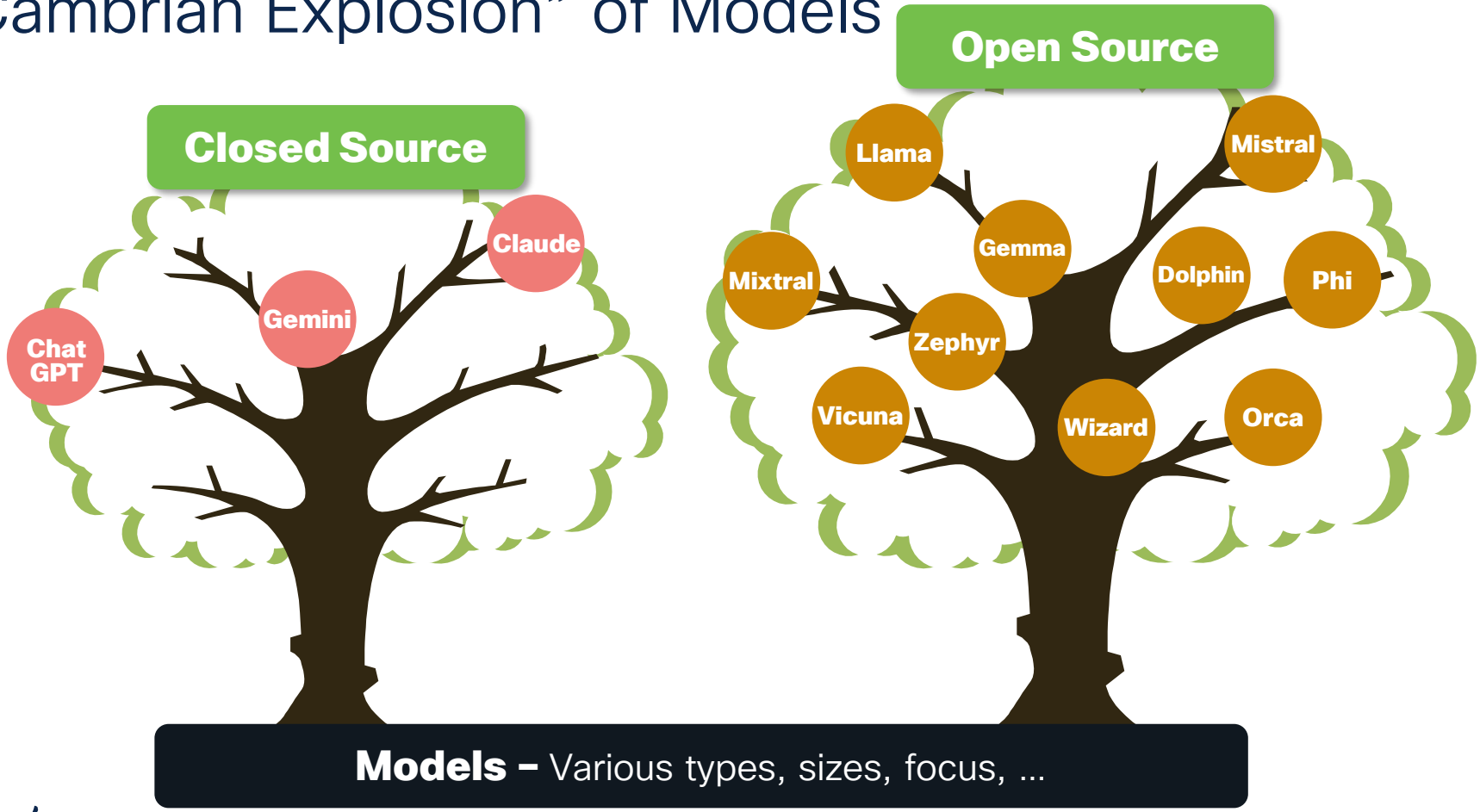
"I swam across the river to get to the other bank."

A machine needs some help…



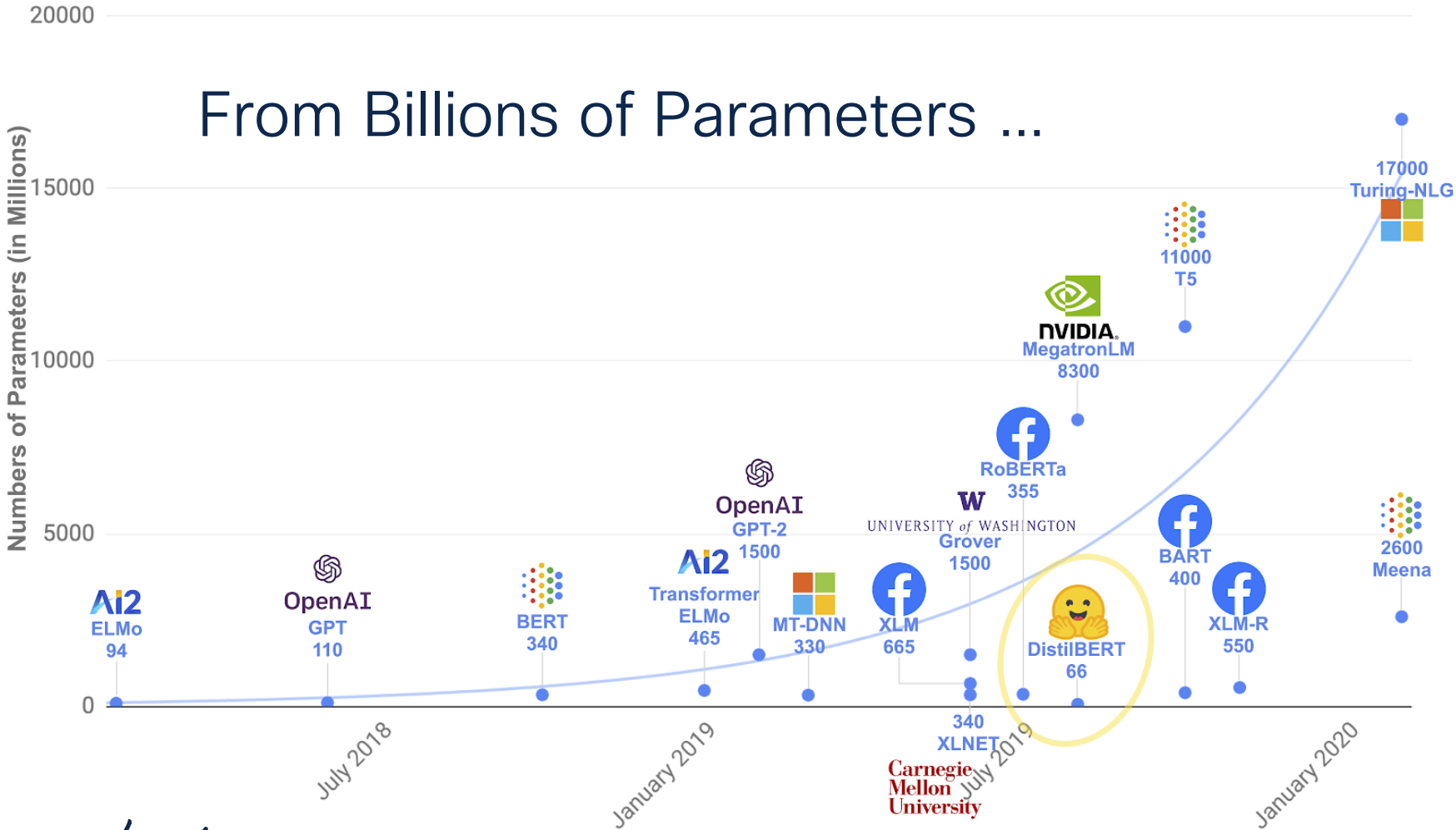
The goal of the attention mechanism is to add **contextual information** to words in a sentence.

“Cambrian Explosion” of Models

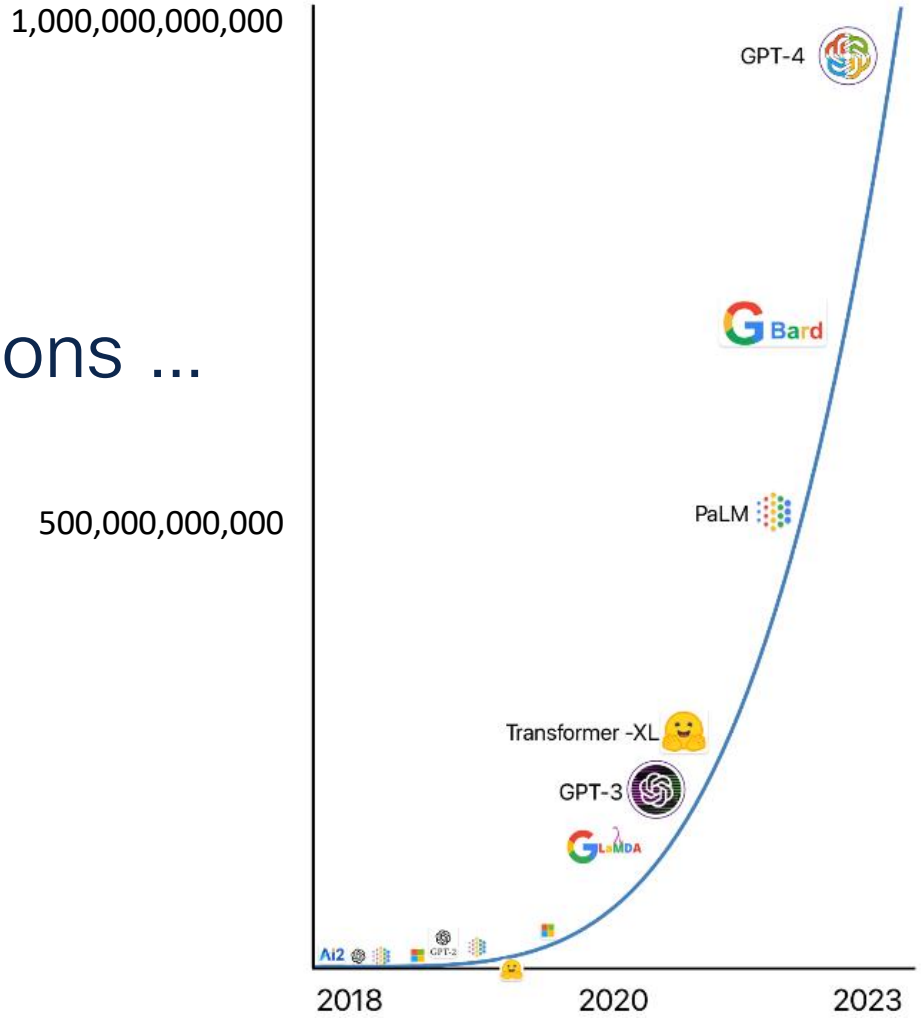


Models – Various types, sizes, focus, ...

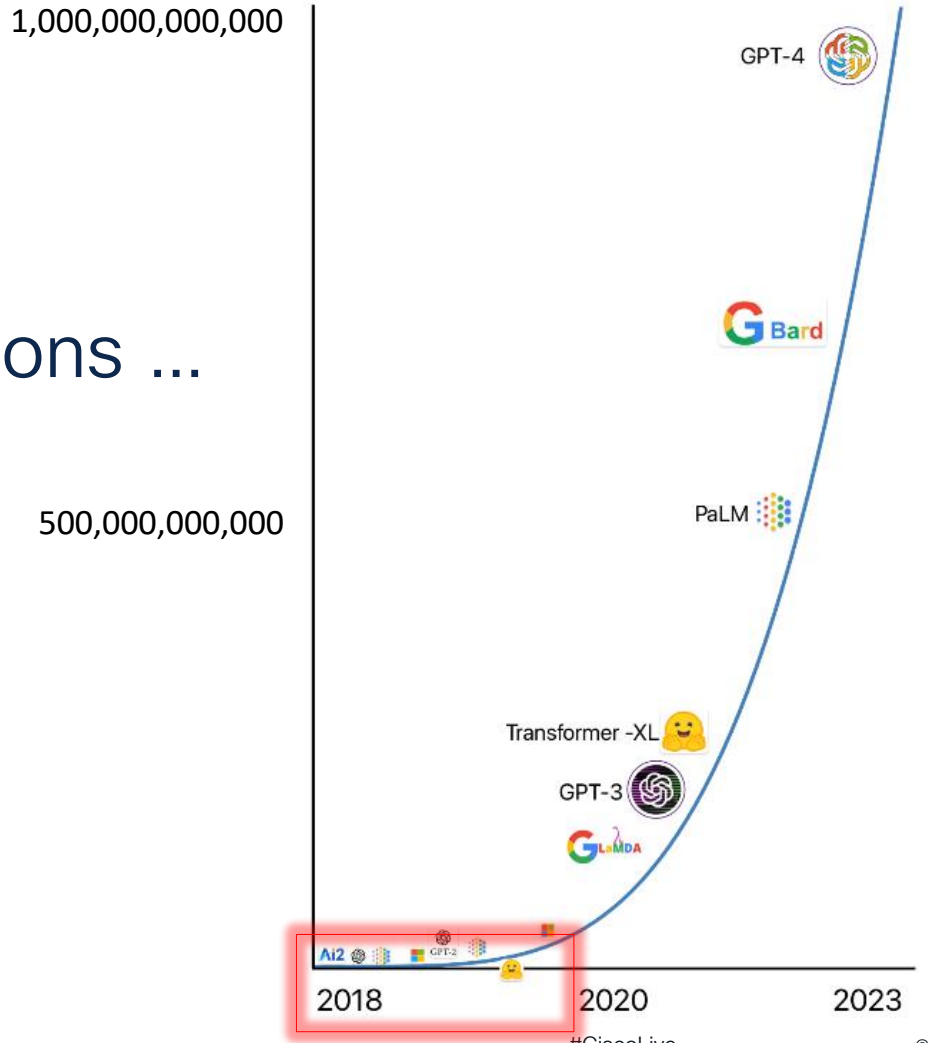
From Billions of Parameters ...



To Trillions ...



To Trillions ...

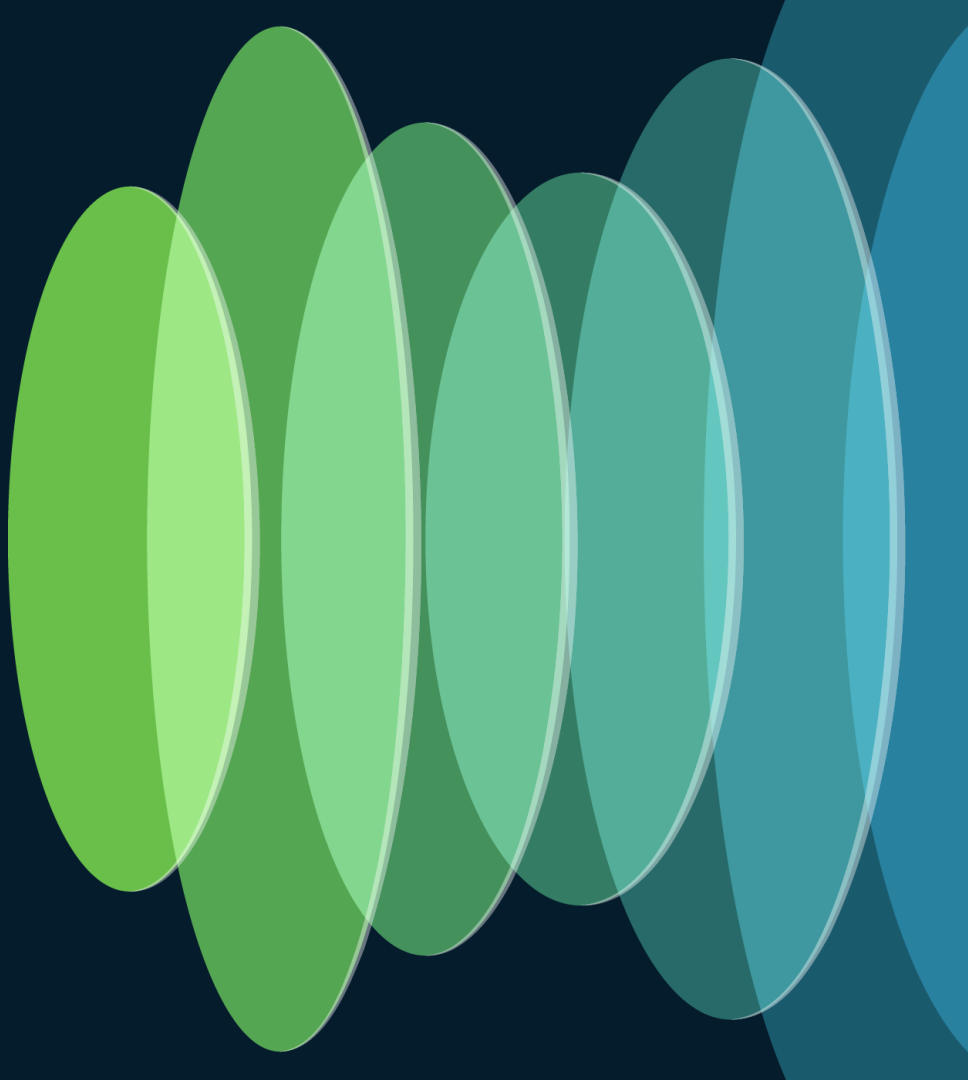


FUN FACT!

The human brain contains
86 billion neurons, and over
100 trillion synaptic connections

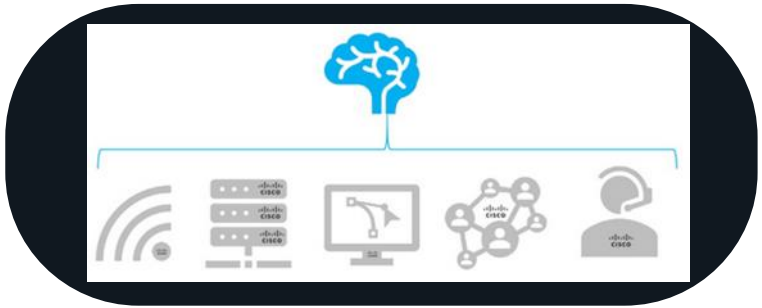
Overview –

Where is Cisco Using AI/ML in Products Today?

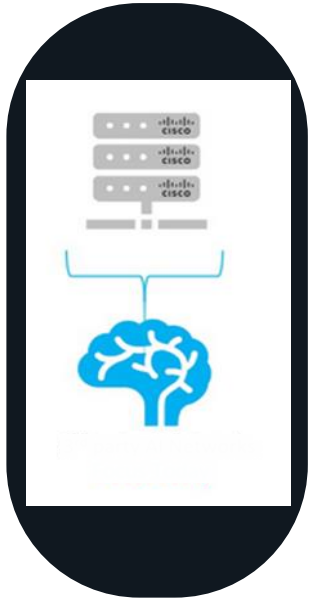


Artificial Intelligence and Cisco

AI in Cisco –
AI to improve products

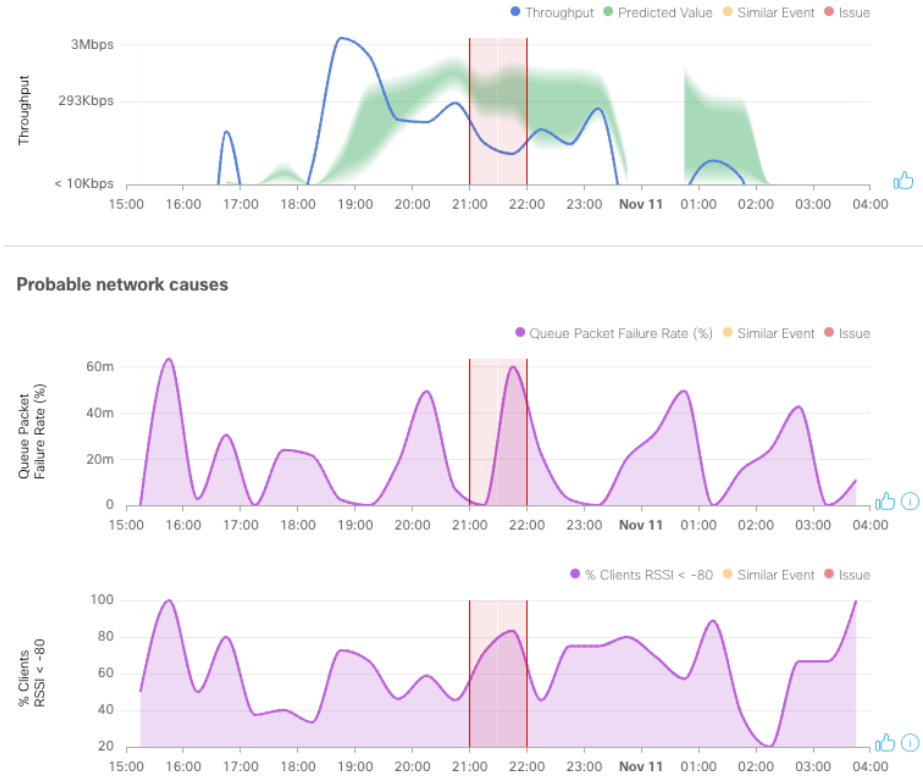


AI on Cisco –
Products to improve AI



AI-Powered Wireless Analytics

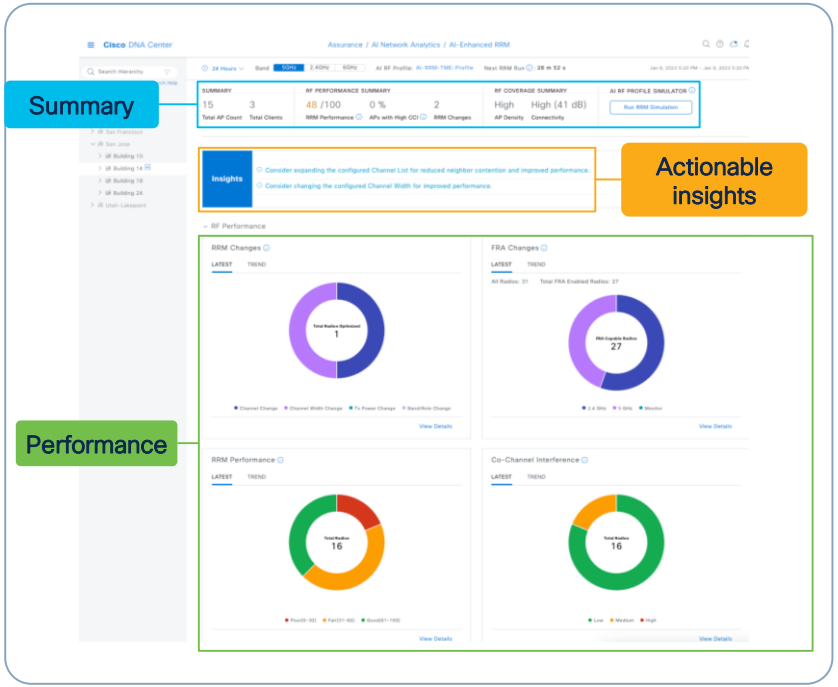
AI-Driven Anomaly Detection
Find + Root Cause Complex Issues



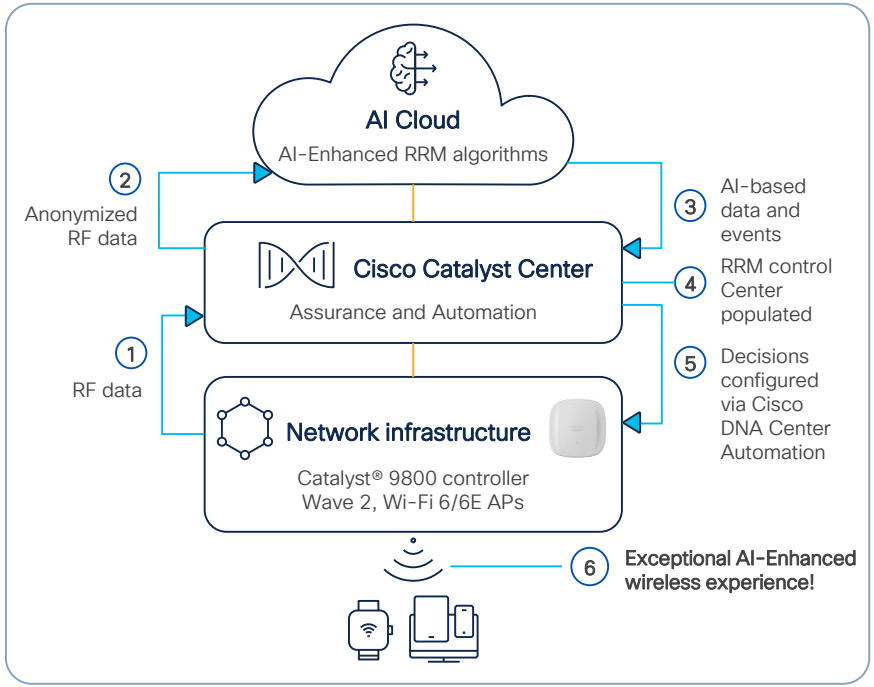
AI-Enhanced RRM

Catalyst AI-Driven RRM solution

Deep RF visibility & advanced control

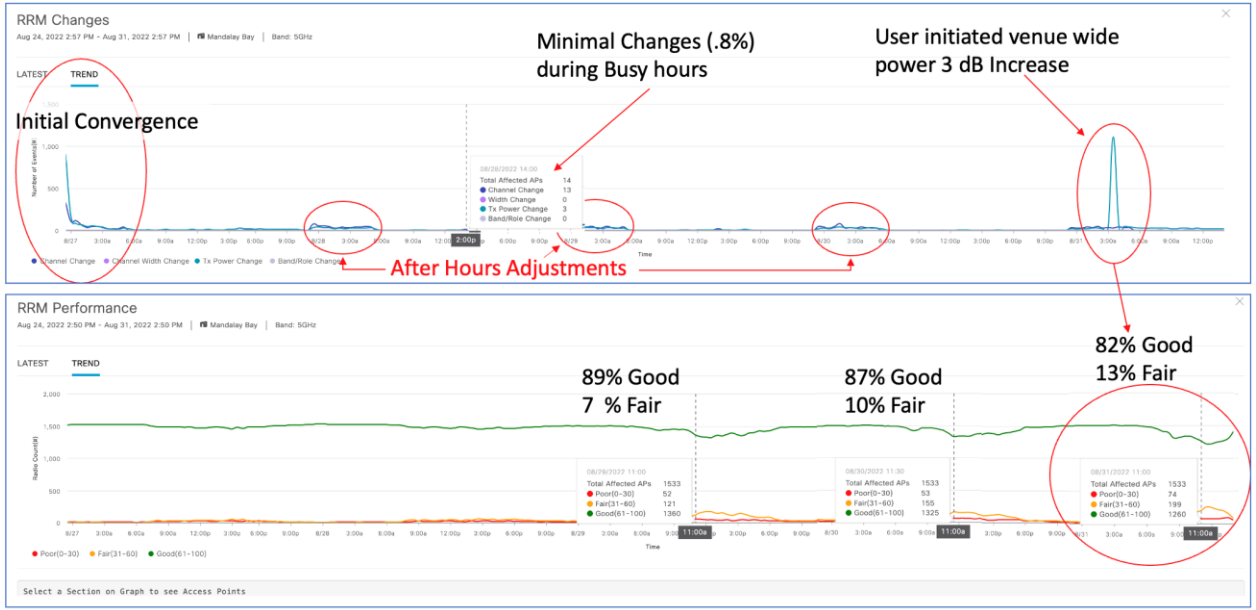


Proactive optimizations for all deployment sizes



AI-Enhanced RRM in Action
Cisco Impact 2022

- Initial Convergence ~3 Hours
- Changes made at Night
- Health stayed above 85% (very good with load)
- Manual Changes made last day, easy to spot the decrease in efficiency



Predictive Internet

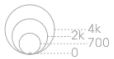
SLA Violations Across the World
and how much Predictive Networks can help

0
CUSTOMERS

0
COUNTRIES

2
CITIES

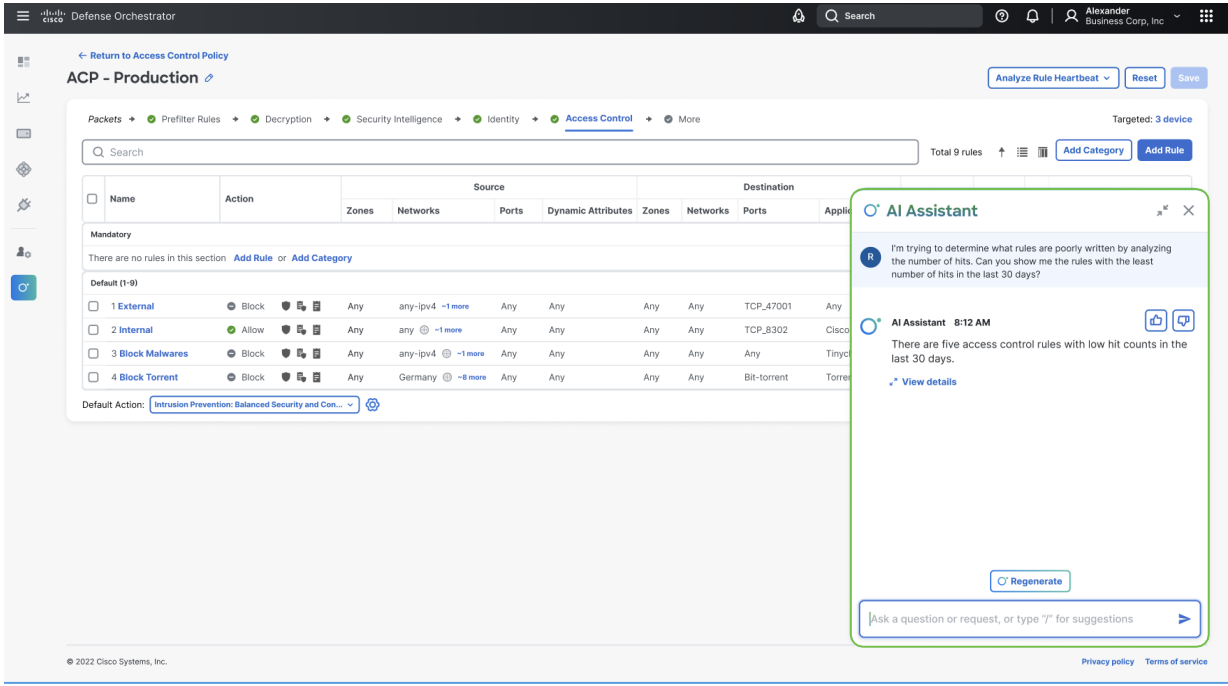
redictive
etworks



Cisco Live Melbourne (Dec 2023)
AI Policy Assistant

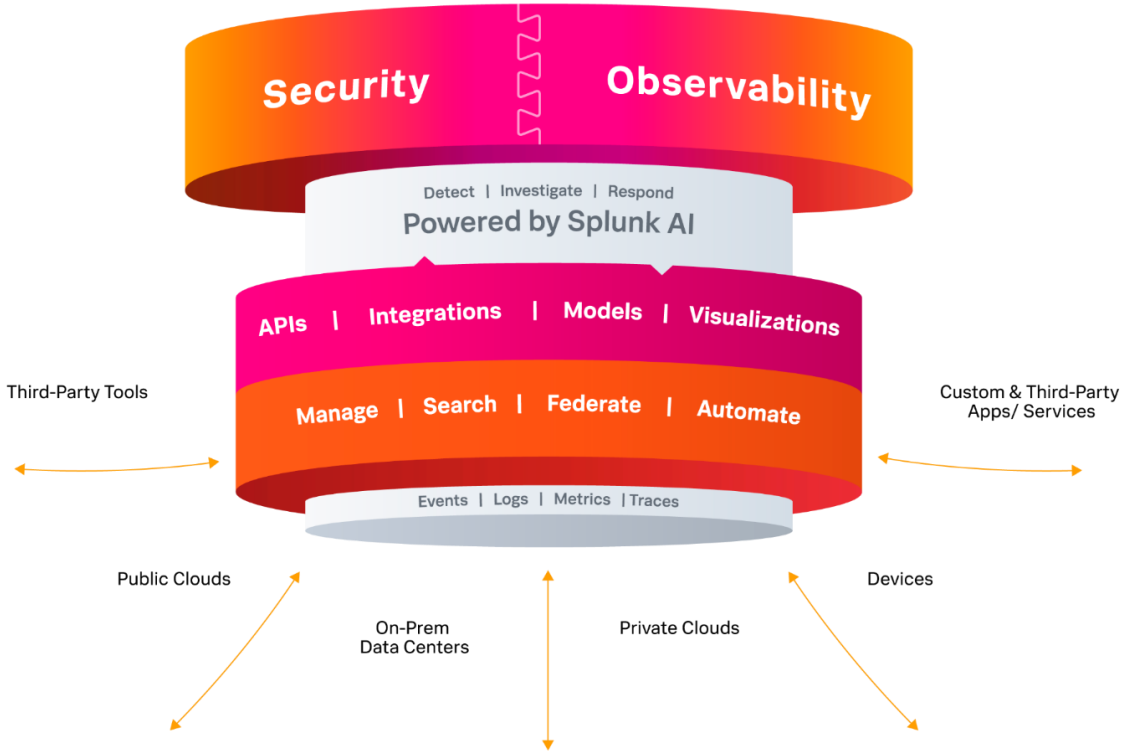
Policy Assistant

- Reduces security policy complexity
- Optimizes policy for efficacy & efficiency

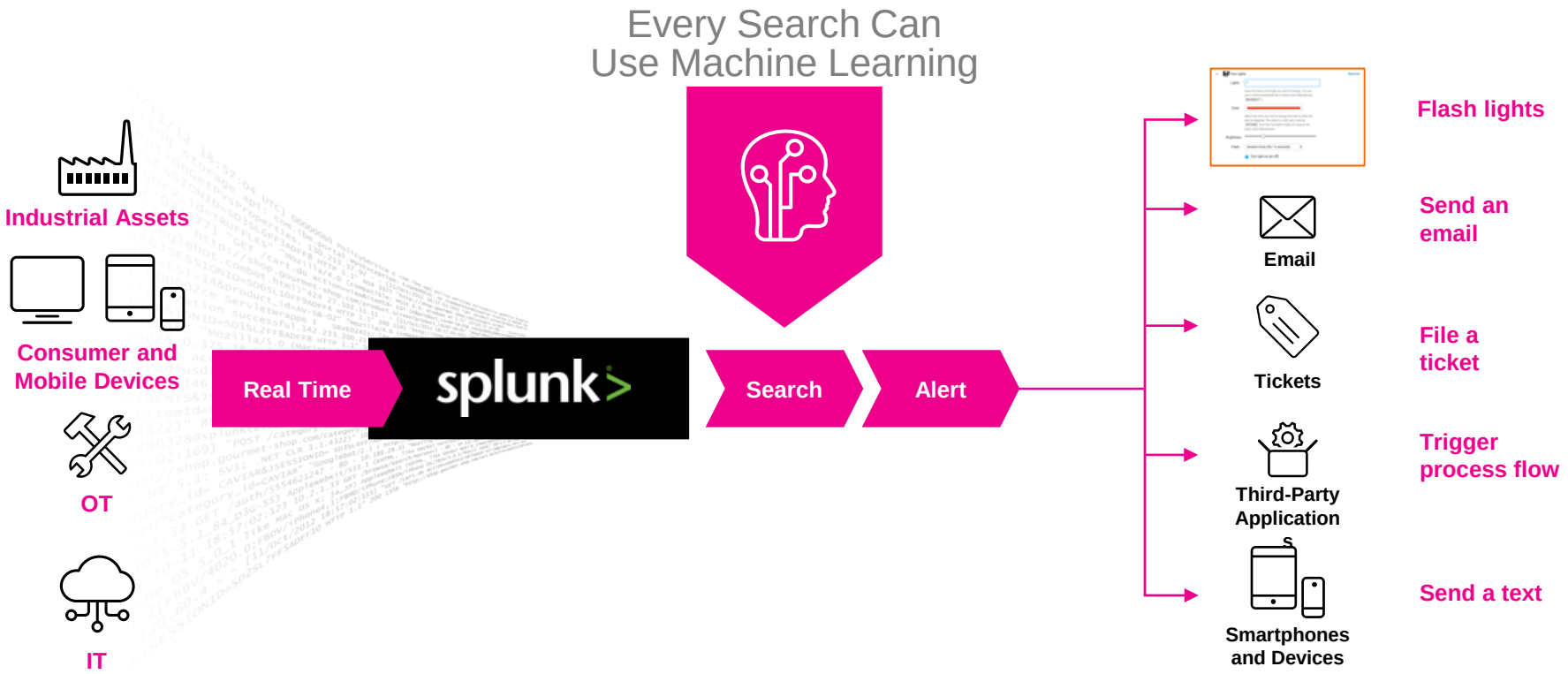


Unified Security & Observability Platform

Powered by Splunk AI



Splunk Machine Learning Tool Kit (MLTK)

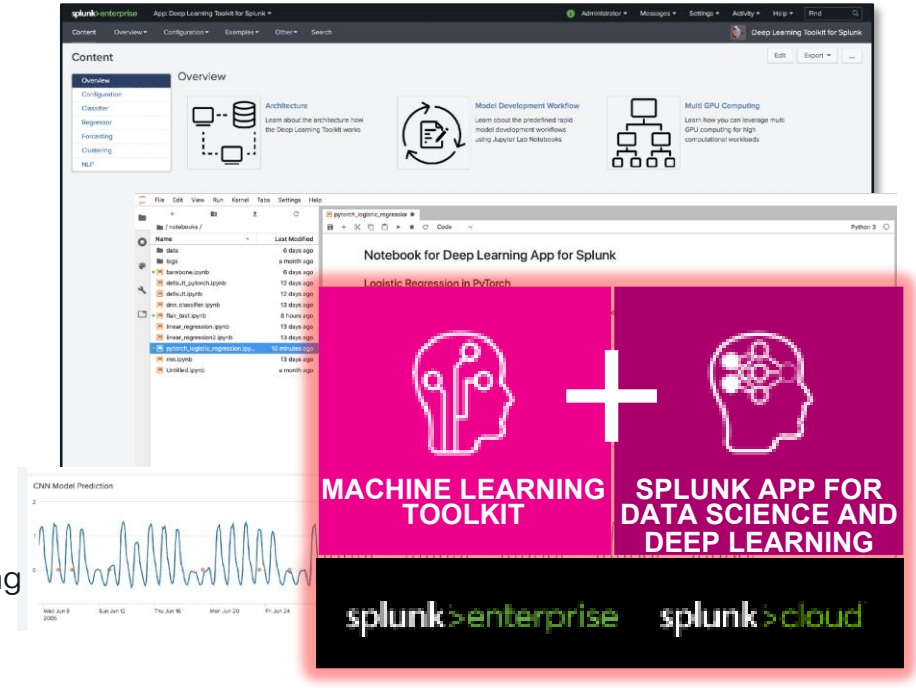


Splunk App for Data Science and Deep Learning (DSDL)

Extension for MLTK to operationalize advanced custom AI / ML use cases

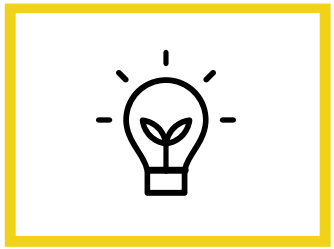
- **Built for Data Scientists**
- **Code Examples:** Guided model building, testing, and deployment of deep learning frameworks
- **Container Management:** Containers & Models can be productionized for scalability & optimization of resources, e.g. CPU & GPU
- **Frameworks:** PyTorch, Tensorflow, SpaCy, DASK, Rapids, Spark, Jupyter Notebooks & Tensorboard, ...

- **Extensible to operationalize any use case**
- **Open Source:**
  New frameworks & Python Libraries freely available for integration into DSDL via github
- **Acceleration:** Support for GPU accelerated machine learning and inference pipelines



Splunk – Investing in Foundational + Generative AI

Combining predictive analytics, accelerated investigation, and workflow enhancements



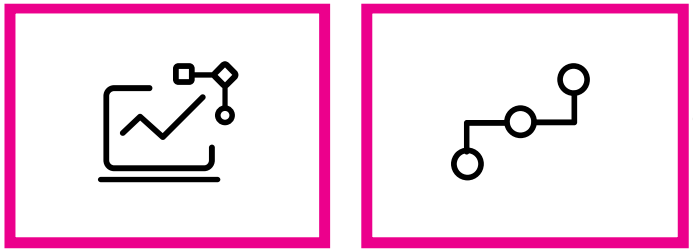
Generative AI Capabilities

Make Everyone an Expert

Reduce need for environment and tool expertise by simplifying content creation and investigation workflows

Detect and Predict

Real-time, streaming analysis to detect anomalies and forecast trends

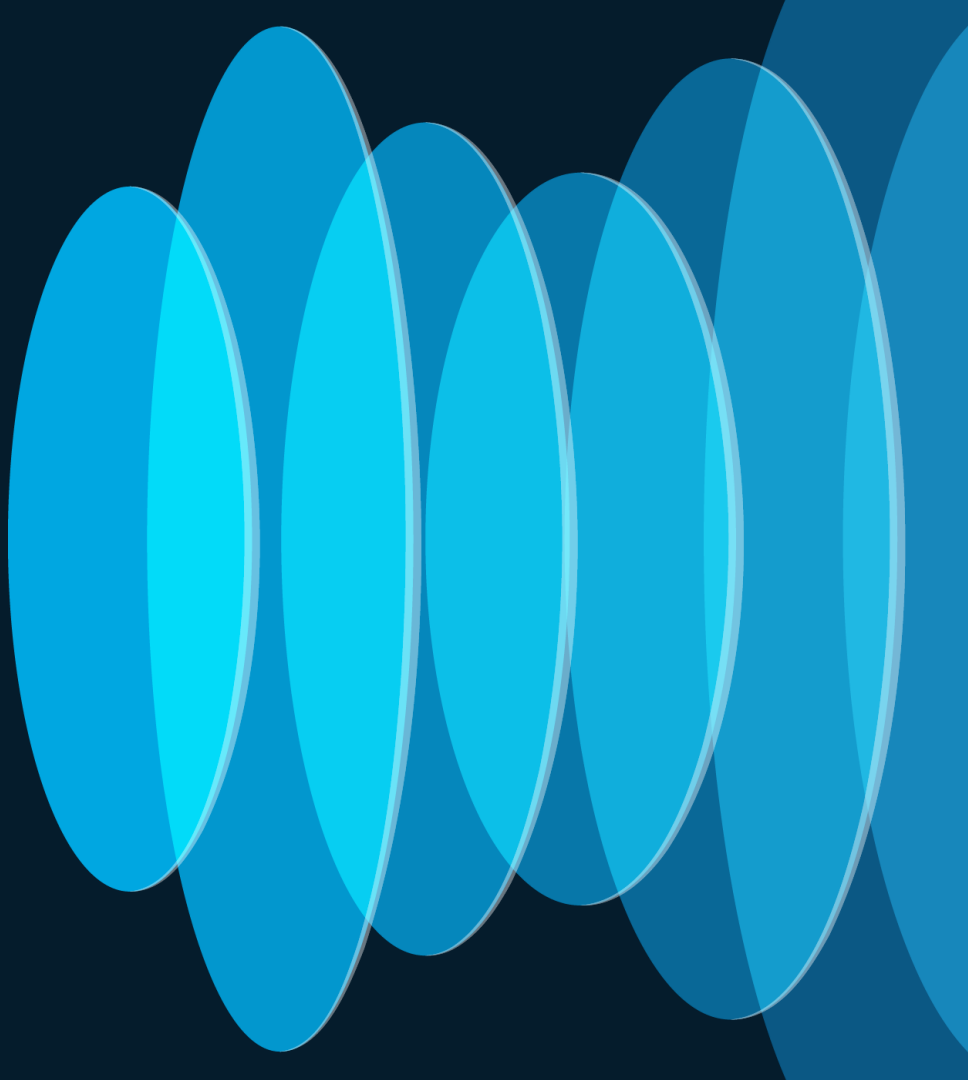


Foundational AI Capabilities

Correlate and Diagnose

Aggregate and analyze all data to investigate and identify root causes

Building Networks for AI/ML Workloads

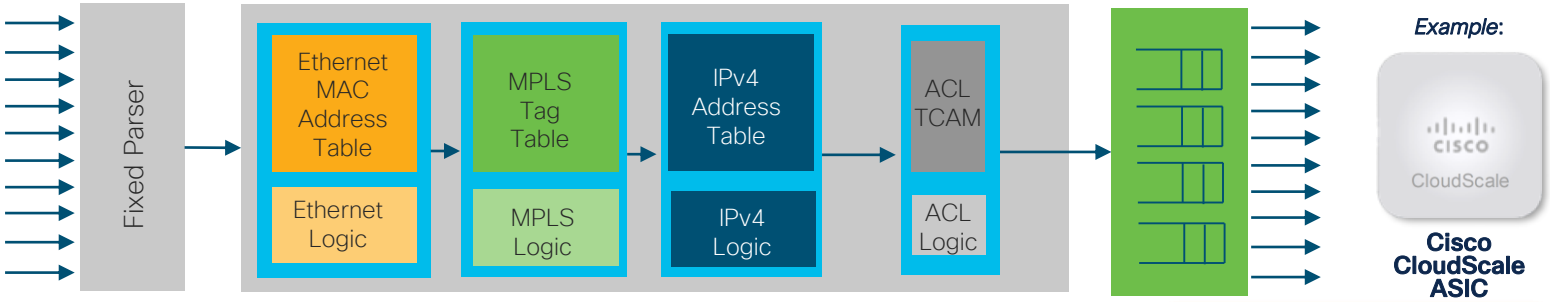


cisco Live!

Cisco Silicon

Foundational Elements

Fixed vs. Programmable Packet Processing



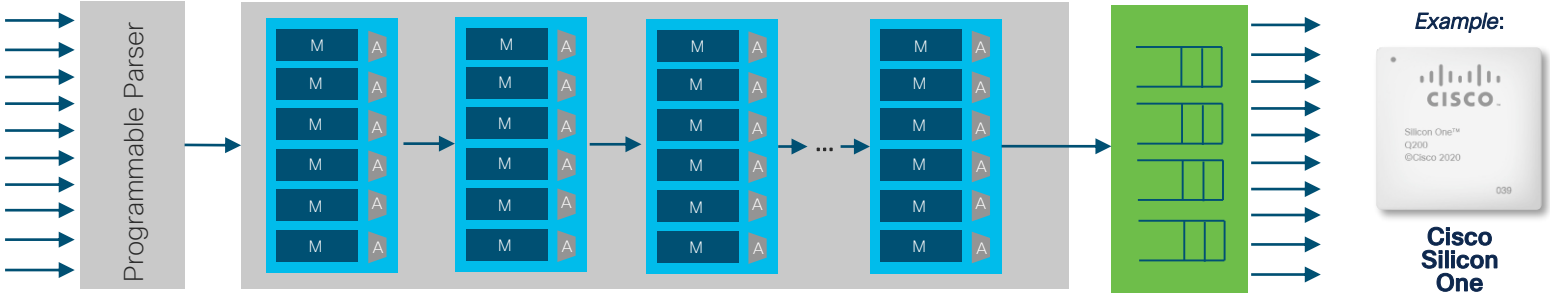
Fixed Pipeline: features and functionality are baked-in at design time

Want to know more about Cisco Silicon for AI?

BRKARC-2095
Tomorrow, 10:30am

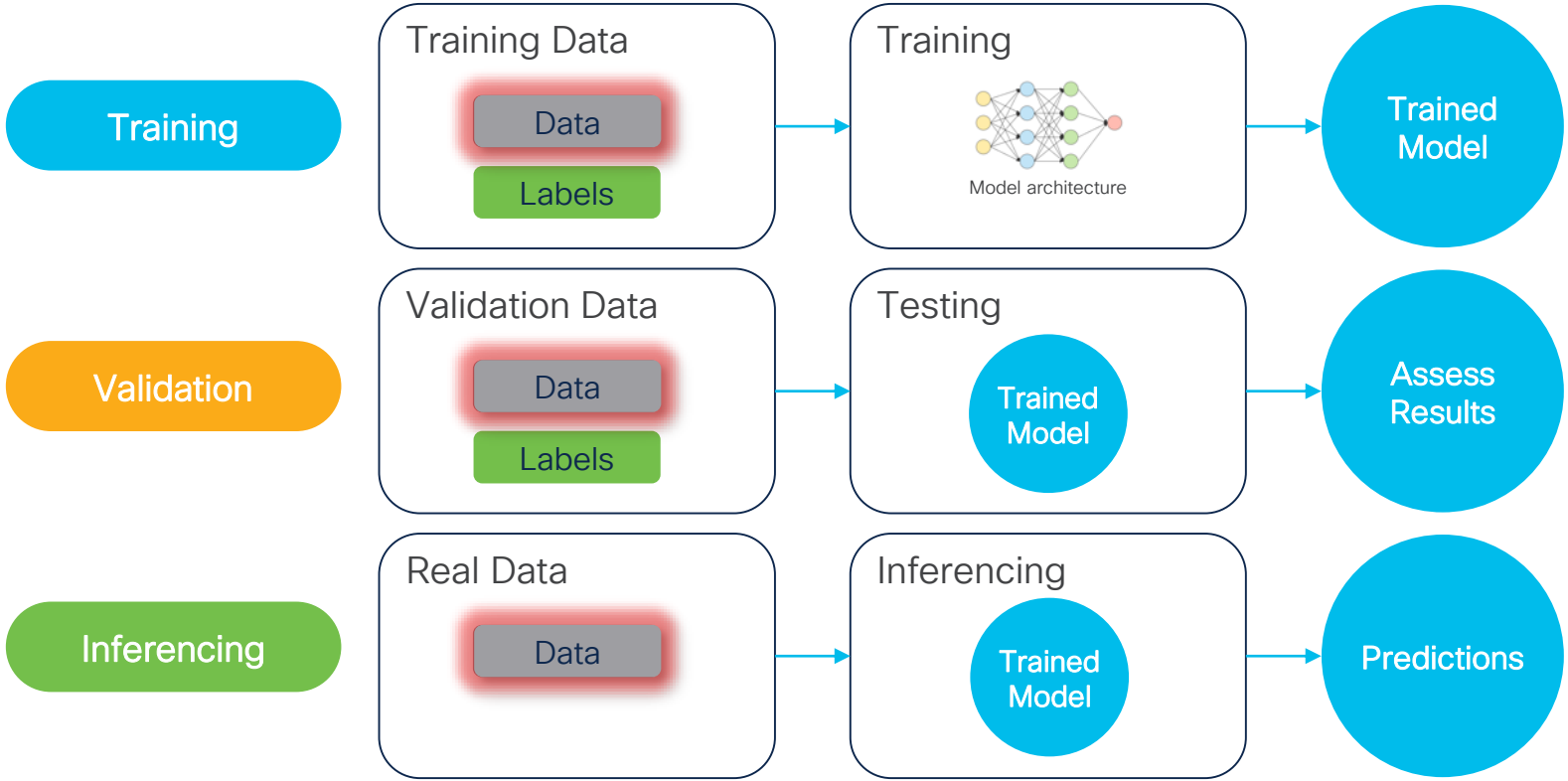
You declare which headers are recognized

You declare what tables are needed and how packets are processed



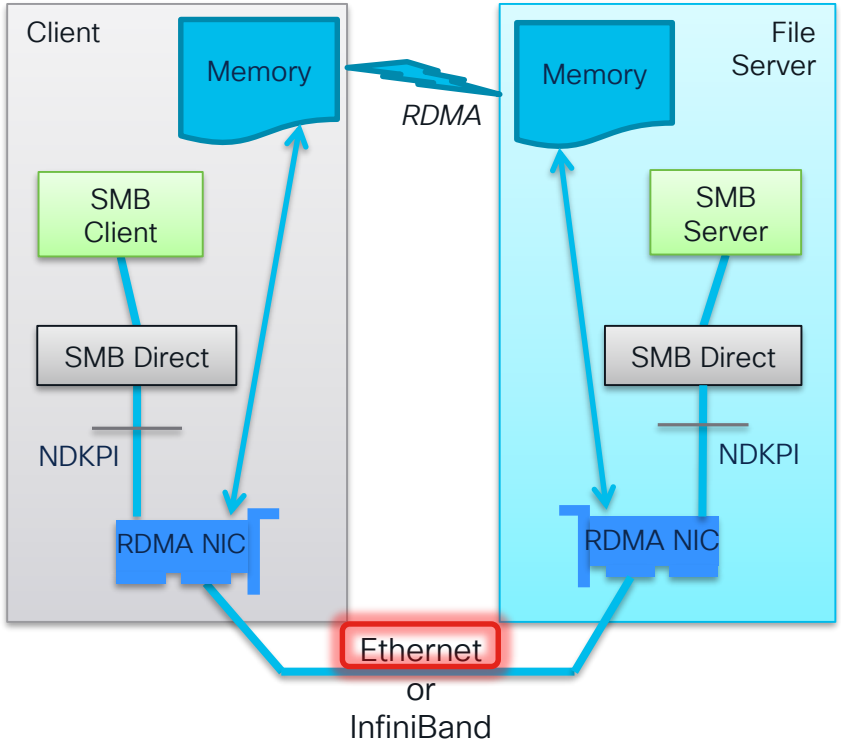
Programmable Pipeline: all stages identical, customer-defined match-action logic

AI - Stages



RDMA – Remote Direct Memory Access

- Allows application software to communicate directly with the hardware (RDMA NIC)
- Bypasses OS stack
- RDMA delivers, low latency, high throughput, zero copy capabilities
- RDMA Hardware Technologies
  - **RoCEv2: RDMA over Converged Ethernet**
  - iWARP: RDMA over TCP/IP
  - Infiniband



Why does the network matter for AI/ML?

Process

Execute instructions on GPU(s)

- **Compute the calculations** to inference / train a model

Synchronise

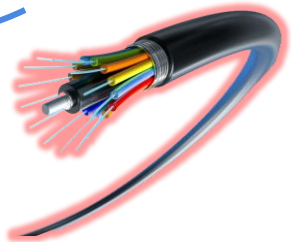
Wait for everyone to complete

- **Creates synchronization between GPUs**
- GPUs are engaged to a 'ready state' for the next stage of the job
- **Computation stalls waiting for the slowest path**
- **Job Completion Time (JCT) is based on the worst-case tail latency**

Notify

Send results of the computation

- **GPU-to-GPU Memory transfer**
  All-to-all Collective
  (Everyone sends to everyone, eg: All-Reduce)
- **High bandwidth compute can saturate network links**



Data Center Quantized Congestion Notification

- **IP ECN** (Explicit Congestion Notification) or **PFC** (Priority Flow Control) alone cannot provide a valid Congestion Management framework
- The **two of them together** provide the desired result of having lossless RDMA communications across Ethernet networks (this is called **DCQCN**)
- The requirements are:
  - Ethernet devices compatible with both techniques
  - Proper configurations applied

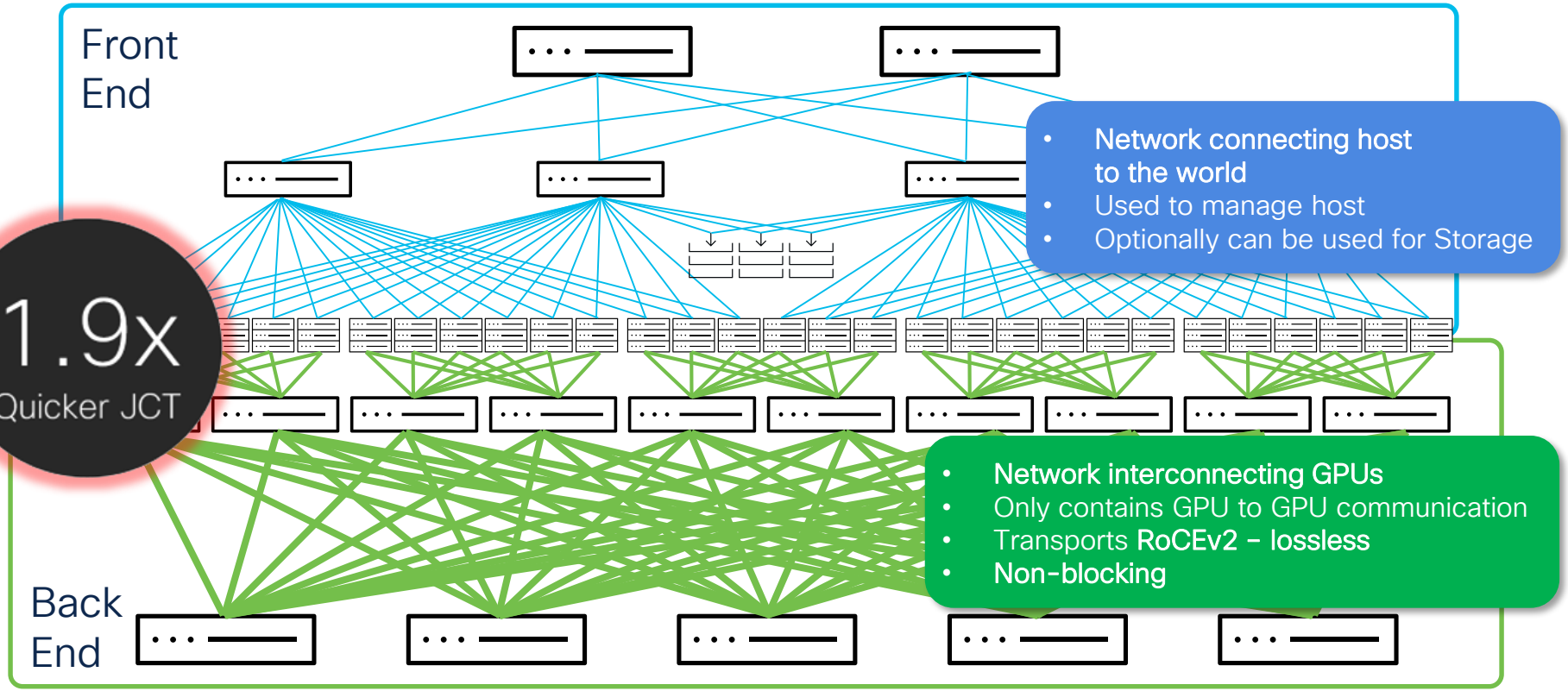
- **Elephant v. Mice flows**
- **AFD and Smart Buffers**

Non-blocking Network

TOMORROW!
3:00 – 4:30pm

MORE INFO AT …

Network Best Practices for Artificial Intelligence (AI/ML) Data Centre *BRKDCN-2921*



AI Data Center Blueprint for Networking

https://www.cisco.com/c/en/us/solutions/design-zone/ai-ready-infrastructure.html

Design guides for AI-ready infrastructure

Gen AI with Flexpod and Nvidia

Blueprint for deployment of Generative AI models for inferencing along with performance metrics.

MLOps with Flashstack and Red Hat

Architecture to operationalize end-to-end AI workflow using Red Hat OpenShift AI.

Gen AI with Intel

Operationalize Generative AI models with Intel AI accelerators.

AI/ML Networking

Build a modern, high-performance, lossless Ethernet fabric to address the stringent requirements of AI/ML workloads.

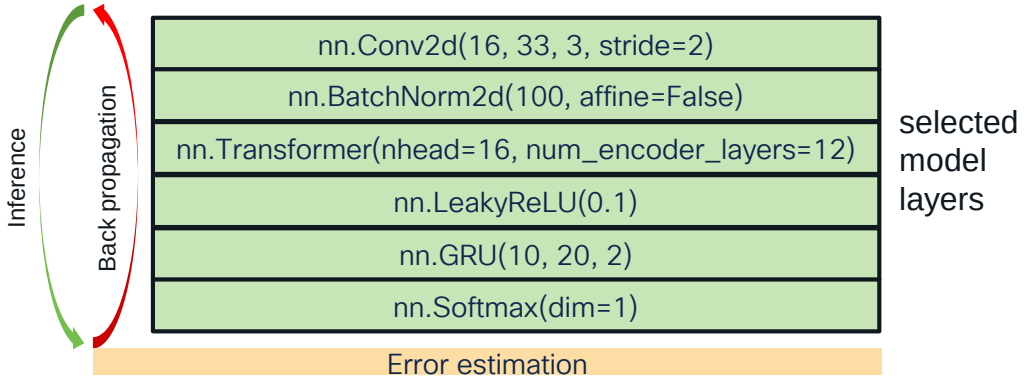
AI Compute

AI is a new programming paradigm – not just an optimality-vs-efficiency trade-off

Key characteristics:

- High parallelism
- Algorithm described by data-set, not procedure
- Coarse-grained selection of standard “layers” to compose network

labeled data set



selected model layers

nn.Conv2d(W,H,Kw,Kh):

```
for (int i = 0; i < outH; i++) {
   for (int j = 0; j < outW; j++) {
      int sum = 0;
      for (int m = 0; m < Kh; m++) {
          for (int n = 0; n < Kw; n++) {
              sum += input[i + m][j + n] * kernel[m][n];
          }
      }
    output[i][j] = sum;
    }
}
```

per ChatGPT 4/9/24

On-Prem AI Solutions – Open Source LLMs with Cisco UCS

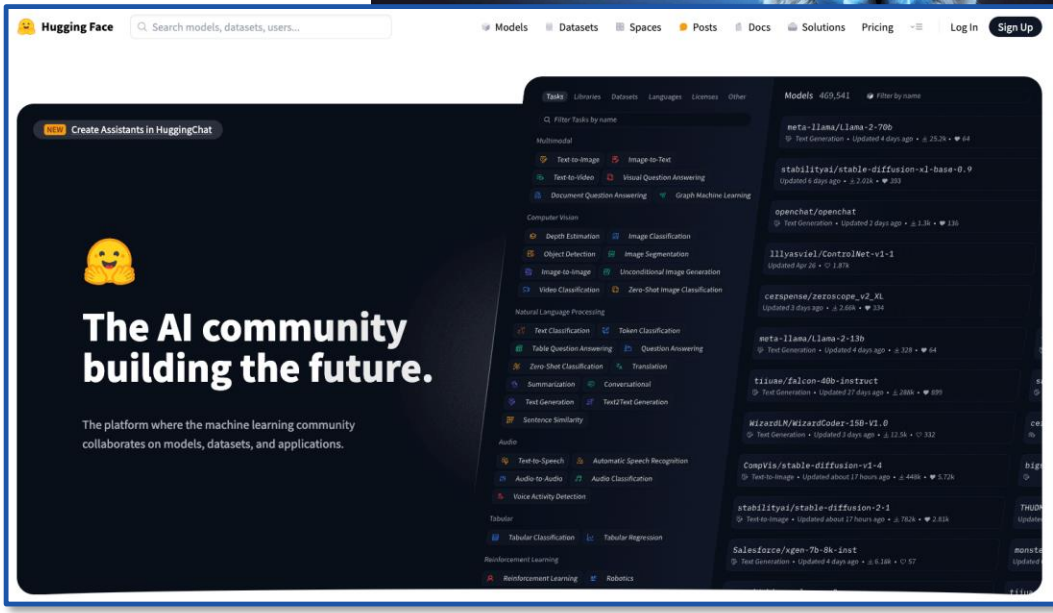
cisco Live!

On-Prem AI Solutions:
Business Challenges of Proprietary GenAI Systems

1. **Cost:** Cloud-based GenAI solutions are becoming a lucrative source of recurring revenue for OpenAI, Google, Anthropic, etc.
2. **Privacy and Security:** Commercially available LLMs are something of a “black box” for users. The users have no control over how they were trained, bias, etc.
3. **Training Gap:** Models are usually months or years out of date. Requires Retrieval Augmented Generation (RAG)
4. **Fine Tuning:** Cloud-based LLMs are “foundational models” and lack fine tuning for certain use cases

Meta has a different Approach – open source AI LLaMA (Large Language Model Meta AI)

- LLaMA (Large Language Model Meta AI) is an LLM released by Meta AI in February 2023
- Mar 3, 2023 (a week after it's release) LLaMA's model was leaked to the open source community
- Unlike other LLMs, Meta LLaMA is now available to the Open Source community



Open Source AI and Responsible AI Use

- Cisco has always have been **a champion of human rights and data privacy**
- Cisco has a **Responsible AI (RAI) framework** and **controls** around transparency, fairness, accountability, privacy, security and reliability

With the emergence of Generative AI, we are **evolving our RAI framework and controls** to surface new variants of issues, such as (but not limited to):

- **False content**
- **Hallucinations**
- **Bias**
- **Unanticipated output**
- **Harmful output**
- **Deep fakes**

Please keep in mind that **open source AI tools and capabilities may or may not have been checked against issues such as these (or others).**

As such, **please use any such tools at your own risk, and in accordance with any applicable AI policies from your organizations.**

Foundational vs. Fine Tuned Models

Confidential Documents

Private Images

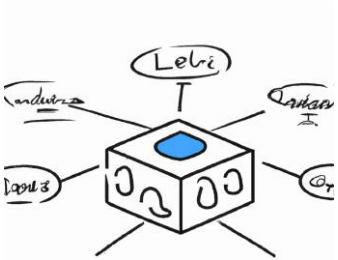
PII data

Financial data

Etc.

Training

Foundational Model



Fine Tuning

Industry Specializations

Healthcare

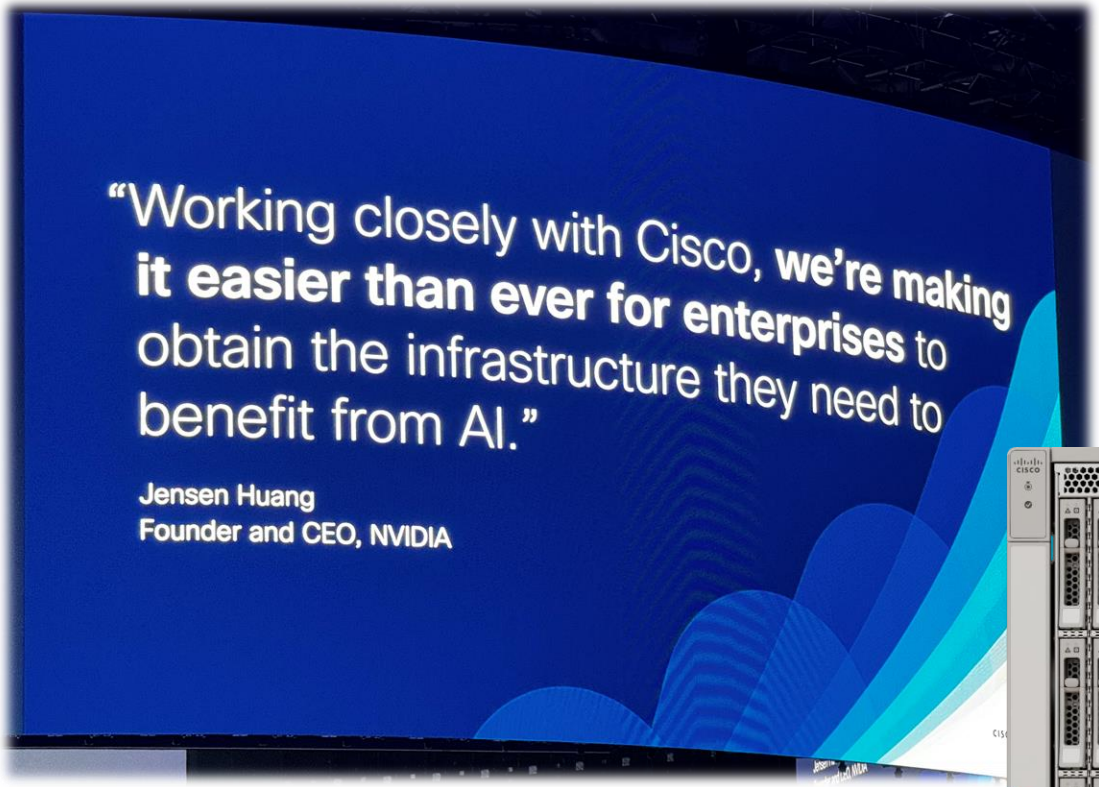
Legal Analysis

Writing Code

Power Utilities

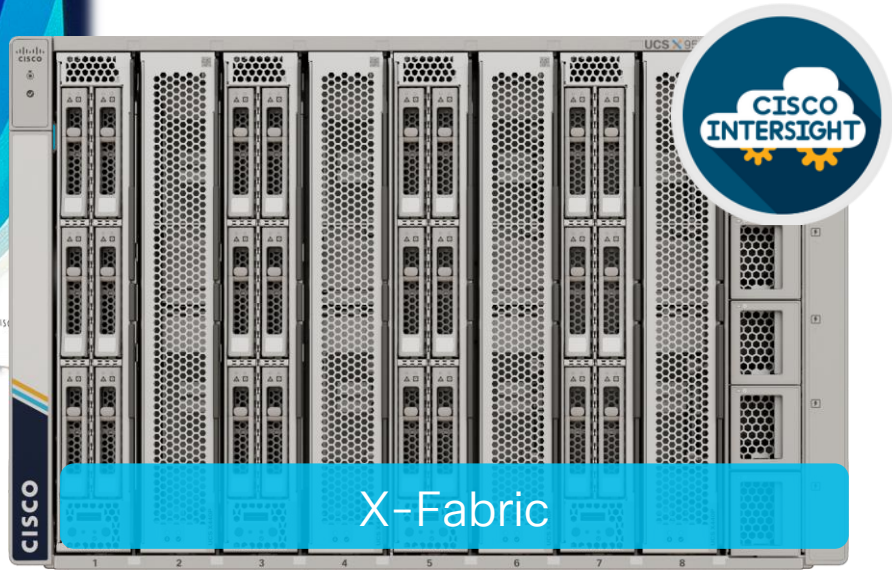
Fintech

Foundational models can be fine-tuned for specific tasks



Only chassis-based server system with expandable GPU modules

UCS-X Series + Nexus



Intersight provides single-click deployment of full-stack on-prem LLM

Demo Time!

www.github.com/pl247/ai-toolkit

Recovery

Your PC/Device needs to be repaired

The application or operating system couldn't be loaded because a required file is missing or contains errors.

File: \WINDOWS\system32\Xxxxxx.xx
Error code: 0xc0000185

You'll need to use recovery tools. If you don't have any installation media (like a disc or USB device), contact your PC administrator or PC/Device manufacturer.

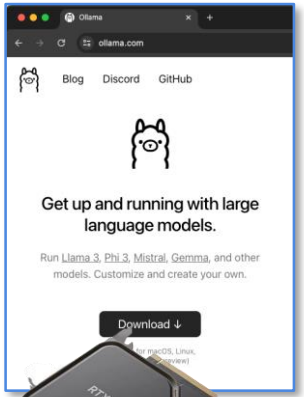
Press Enter to try again
Press F8 for Start-up Settings
Press Esc for UEFI Firmware settings

And Keep in Mind ...

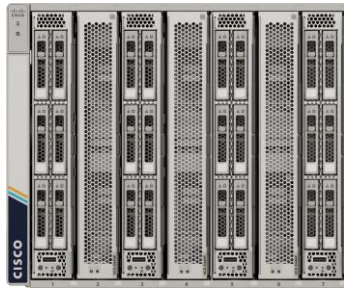
You can **experiment with all of the AI capabilities we are showing, and get your hands dirty**, almost no matter what hardware you can get your hands on ...

Model quantization is the process of reducing the precision of a model's parameters from floating-point to lower bit-width representations (such as 8-bit integers) to decrease its memory footprint and computational requirements while aiming to maintain accuracy.

https://ollama.com/library



Cisco UCS with Nvidia H100 GPU
up to 70B parameter models, 8-bit quantized



Gaming PC / Mac, larger / discrete GPU
up to 30-40B parameter models, 4-bit quantized

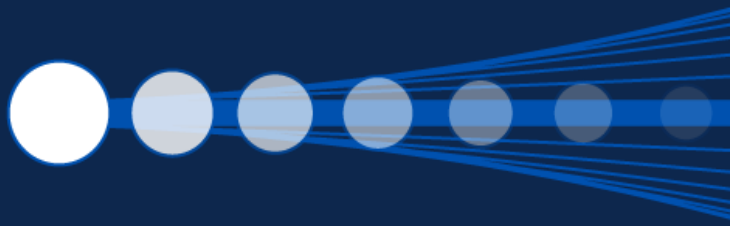
Generic PC / Mac, no GPU or small GPU
up to 7-15B parameter models, 4-bit quantized

Raspberry Pi 5
up to 7B parameter models, 4-bit quantized

Large Language Models (LLMs), GPTs, and more ... and their use in Networking

John

Langchain



An open-source framework for AI development

- Python or JavaScript
- Highly abstracted
- Easy
- Accessible

Integrated openAI

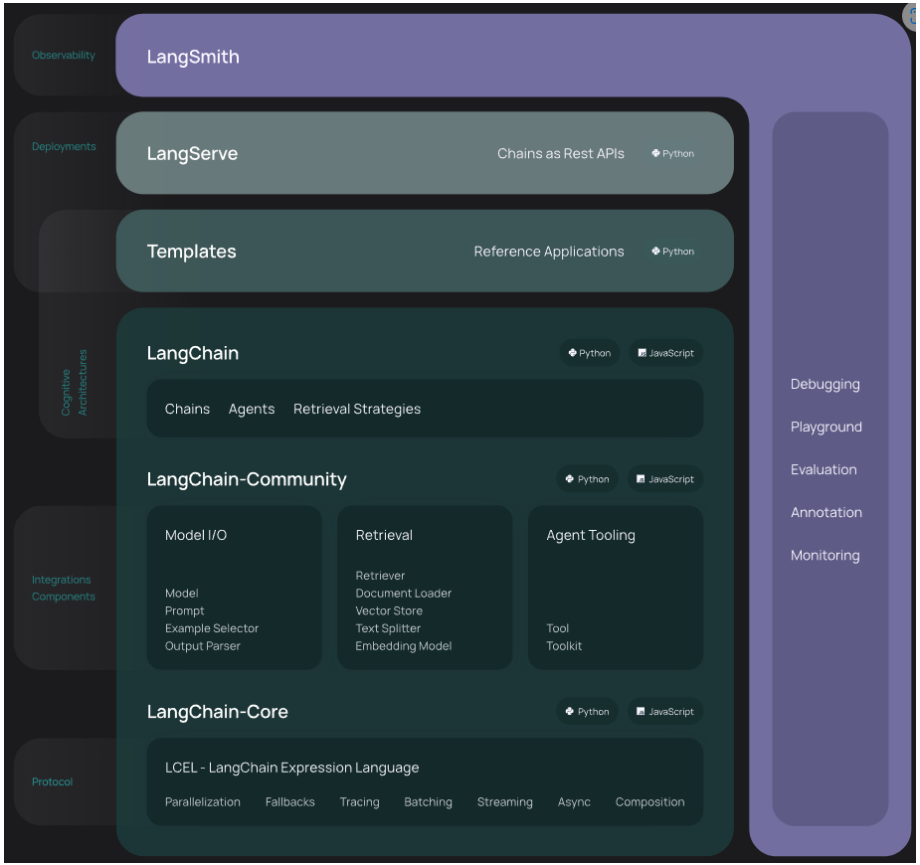
Options for ***private LLMs***

Evolving quickly and daily

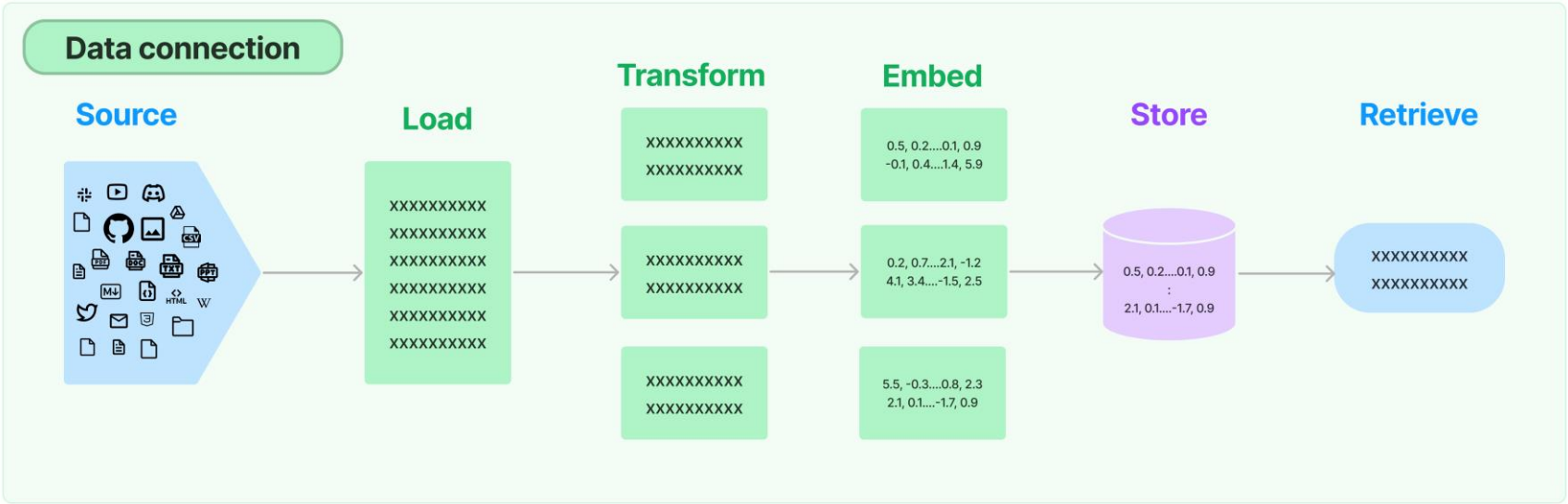
“Like developing on wet cement”

LCEL (Langchain Expression Language)

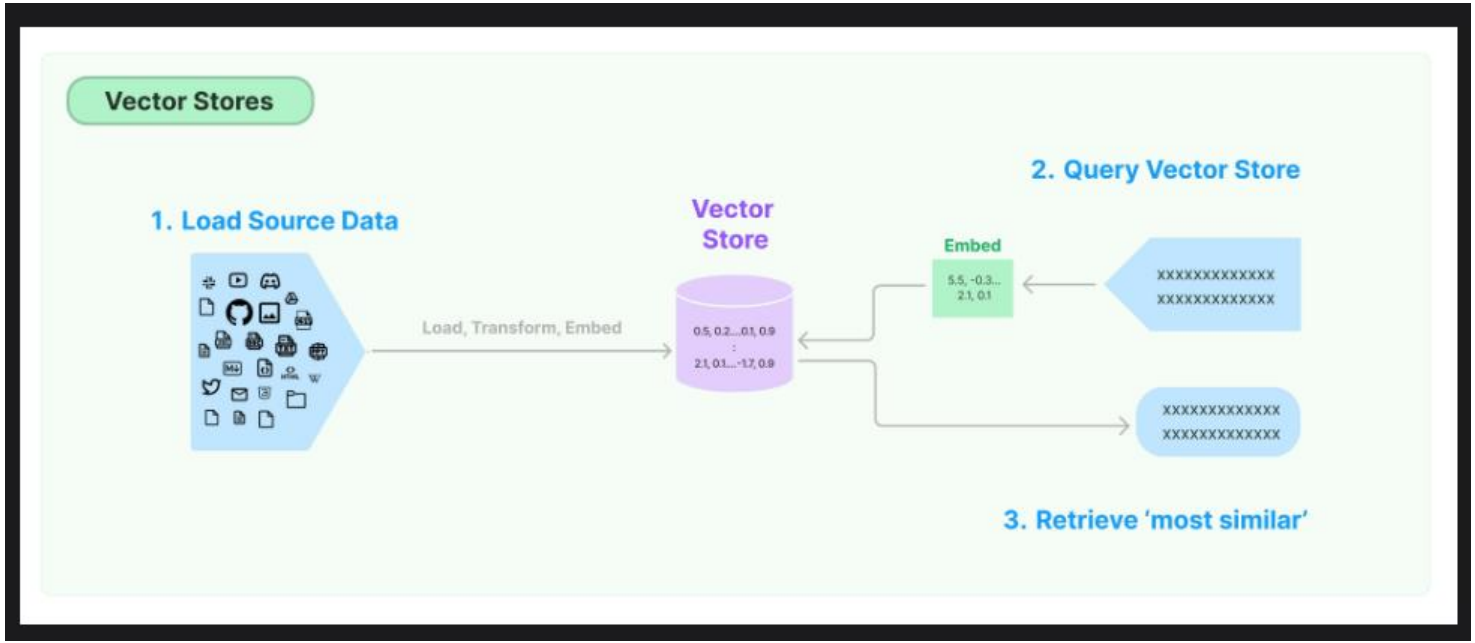
The LangChain ecosystem



A high-level langchain abstraction



Vector Store



```python
#Load OpenAI API Key
os.environ["OPENAI_API_KEY"] = openai.api_key

#Setup Loader - in this case a PDF Loader
loader = PyMuPDFLoader("cisco-annual-report-2022.pdf")

# Load and split the pdf into pages
pages = loader.load_and_split()

# setup a text splitter
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=250,
    chunk_overlap=20,
    length_function=len,
)
```

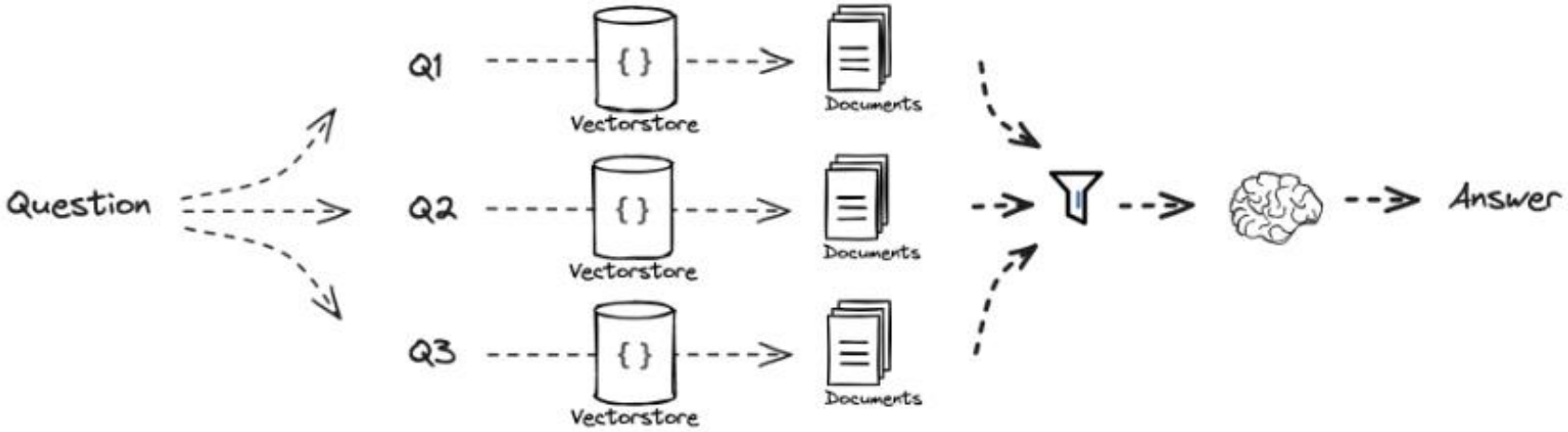
```python
# split the pages into smaller chunks called docs
docs = text_splitter.split_documents(pages)

# transform to embeddings
embeddings = OpenAIEmbeddings()

# setup and store docs and embeddings into ChromaDB
vectordb = Chroma.from_documents(docs, embedding=embeddings,
                                 persist_directory=".")

#Make the database persisten
vectordb.persist()

# setup memory so it remembers previous questions and answers
memory = ConversationBufferMemory(memory_key="chat_history", return_messages=True)
```

```python
# perform the Conversational Retreival Chain
pdf_qa = ConversationalRetrievalChain.from_llm(OpenAI(temperature=0.5) , vectordb.as_retriever(), memory=memory)

#Run the question
question = "Who is the senior leadership team at Cisco?"
result = qa.run(question)

#Print the results to the screen
print(result)
```

Advanced RAG – RAG-Fusion

Part 6: RAG-Fusion

Flow:



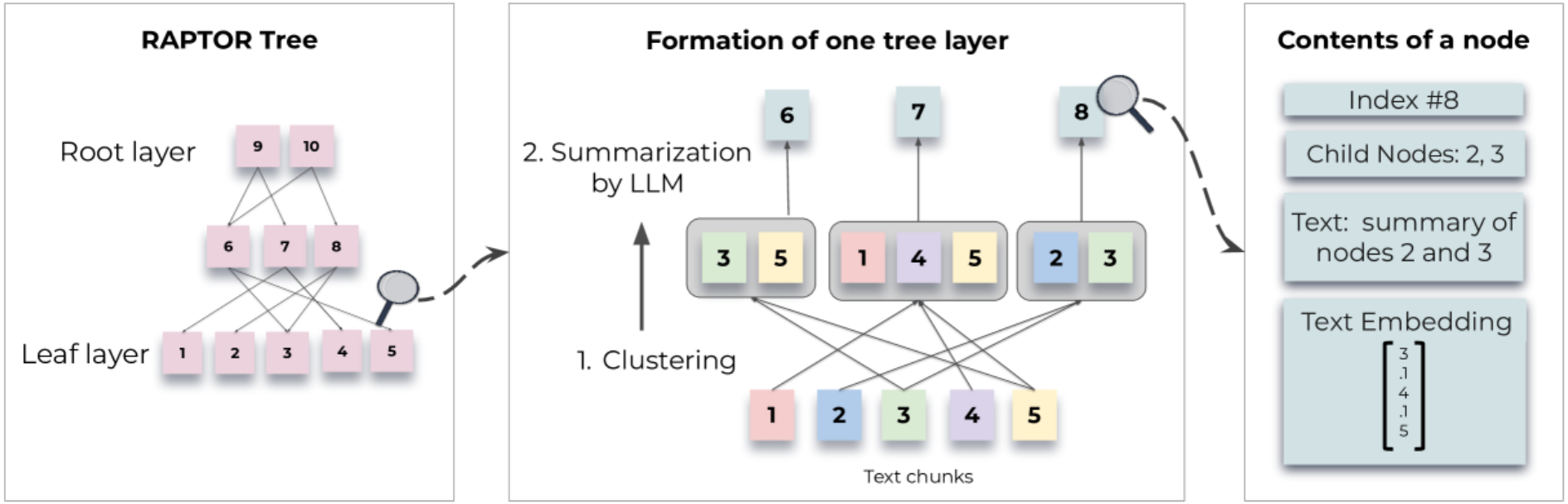
[rag-from-scratch/rag_from_scratch_5_to_9.ipynb at main · langchain-ai/rag-from-scratch · GitHub]

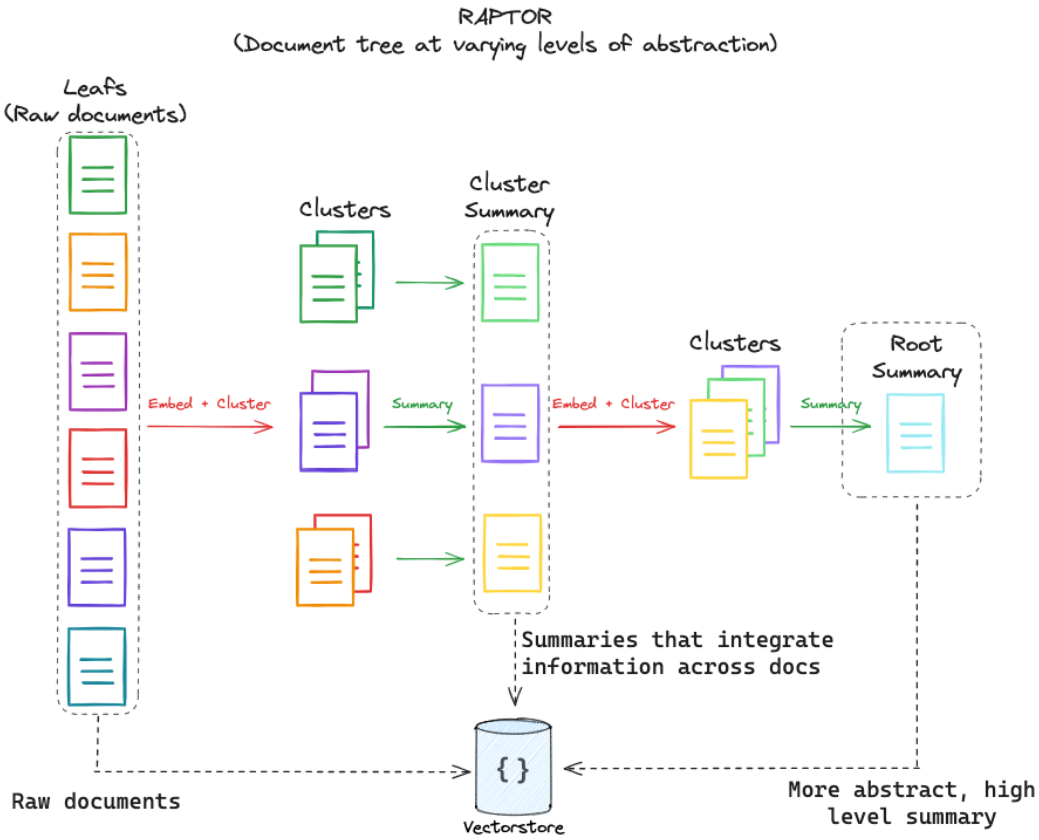
The drawbacks and flaws of RAG approaches

RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval (arxiv.org)

Nevertheless, existing retrieval-augmented approaches also have flaws. The one we tackle is that most existing methods retrieve only a few short, contiguous text chunks, which limits their ability to represent and leverage large-scale discourse structure. This is particularly relevant for thematic questions that require integrating knowledge from multiple parts of a text, such as understanding an entire book, as in the NarrativeQA dataset (Kočiskỳ et al., 2018). Consider the fairy tale of Cinderella, and the question "How did Cinderella reach her happy ending?". The top-k retrieved short contiguous texts will not contain enough context to answer the question.

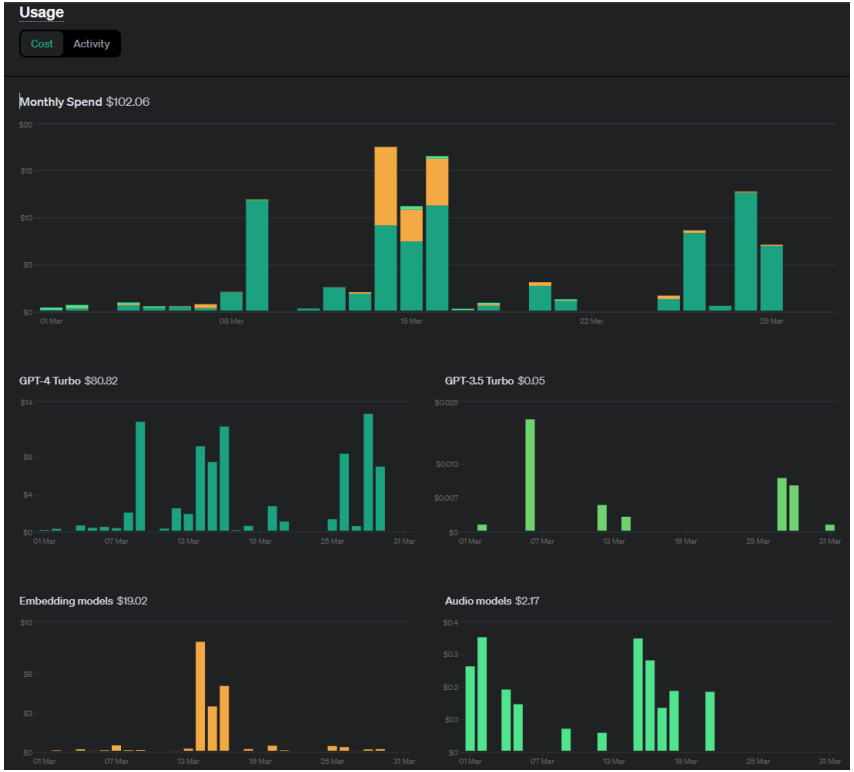
Look familiar? Like a well designed data center!





[langchain/cookbook/RAPTOR.ipynb at master · langchain-ai/langchain · GitHub]

Personal Monthly Spend on openAI Services



Open Source: Free; Private; Local

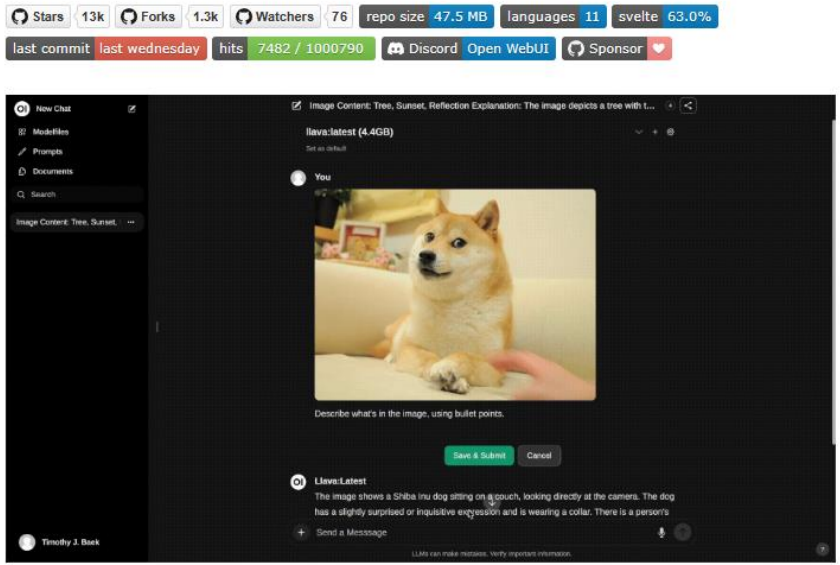
Get up and running with large language models.

Run Llama 2, Code Llama, and other models.
Customize and create your own.

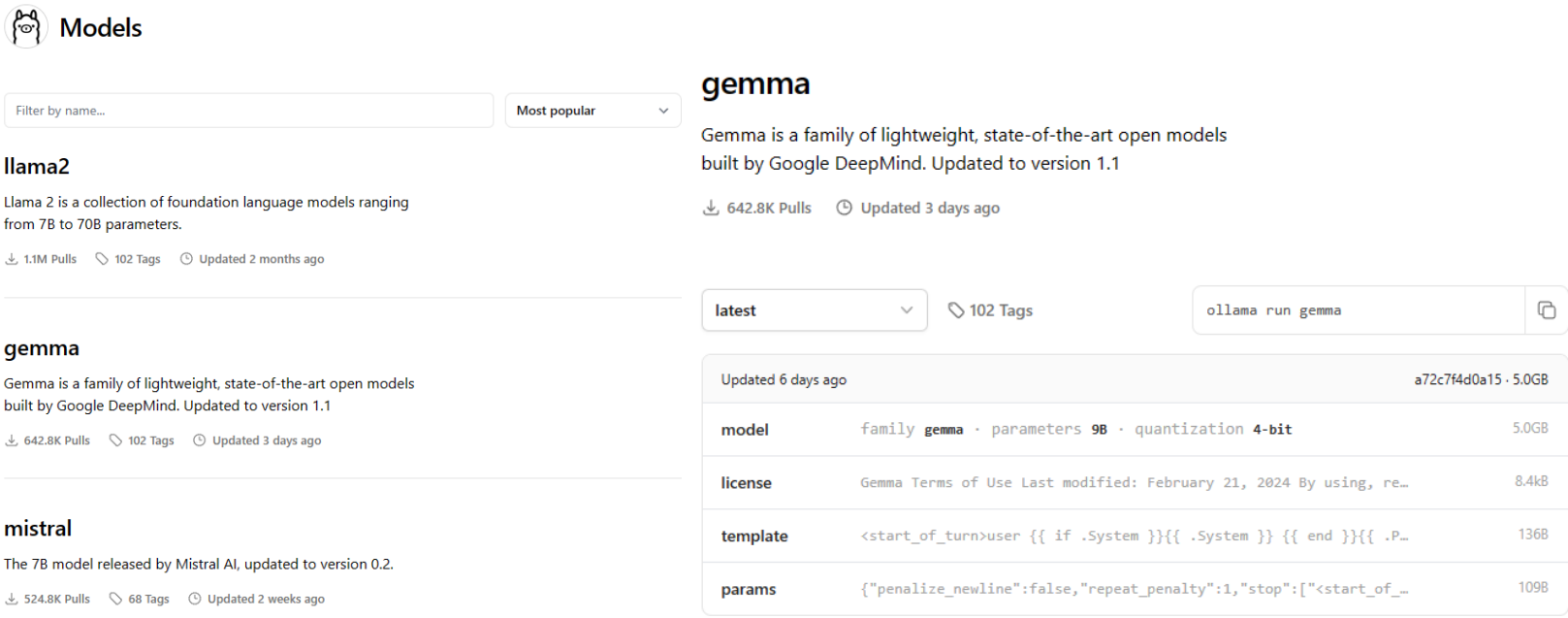
Open WebUI

Open WebUI

Open WebUI is an extensible, feature-rich, and user-friendly self-hosted WebUI designed to operate entirely offline. It supports various LLM runners, including Ollama and OpenAI-compatible APIs.



Pick your model(s)



Local AI development with Ollama

```
ollama_1  | llm_load_print_meta: model ftype      = Q4_0
ollama_1  | llm_load_print_meta: model params     = 7.24 B
ollama_1  | llm_load_print_meta: model size       = 3.83 GiB (4.54 BPW)
ollama_1  | llm_load_print_meta: general.name     = mistralai
ollama_1  | llm_load_print_meta: BOS token        = 1 '<s>'
ollama_1  | llm_load_print_meta: EOS token        = 2 '</s>'
ollama_1  | llm_load_print_meta: UNK token        = 0 '<unk>'
ollama_1  | llm_load_print_meta: LF token         = 13 '<0x0A>'
ollama_1  | ggml_cuda_init: GGML_CUDA_FORCE_MMQ:   yes
ollama_1  | ggml_cuda_init: CUDA_USE_TENSOR_CORES: no
ollama_1  | ggml_cuda_init: found 1 CUDA devices:
ollama_1  |   Device 0: NVIDIA GeForce RTX 2060 SUPER, compute capability 7.5, VMM: yes
ollama_1  | llm_load_tensors: ggml ctx size =    0.22 MiB
ollama_1  | llm_load_tensors: offloading 32 repeating layers to GPU
ollama_1  | llm_load_tensors: offloading non-repeating layers to GPU
ollama_1  | llm_load_tensors: offloaded 33/33 layers to GPU
ollama_1  | llm_load_tensors:        CPU buffer size =    70.31 MiB
ollama_1  | llm_load_tensors:      CUDA0 buffer size =  3847.55 MiB
```

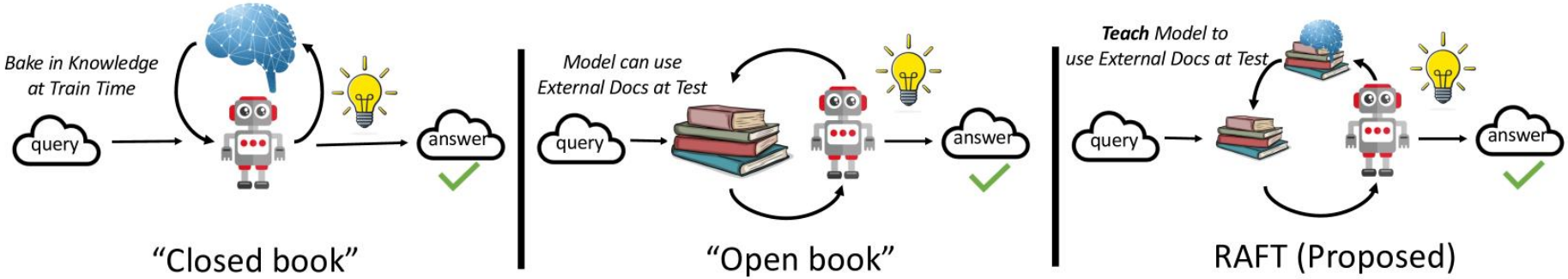
Fine Tuning

Fine-tuning is the process of taking a pretrained machine learning model and further training it on a smaller, targeted, domain-specific, data set. The aim of fine-tuning is to maintain the original capabilities of a pretrained model

1. **Setup:** Loading our dataset and filtering down to one domain to fine-tune on.
2. **Data preparation:** Preparing your data for fine-tuning by creating training and validation examples, and uploading them to the `Files` endpoint.
3. **Fine-tuning:** Creating your fine-tuned model.
4. **Inference:** Using your fine-tuned model for inference on new inputs.

https://github.com/openai/openai-cookbook/blob/main/examples/How_to_finetune_chat_models.ipynb

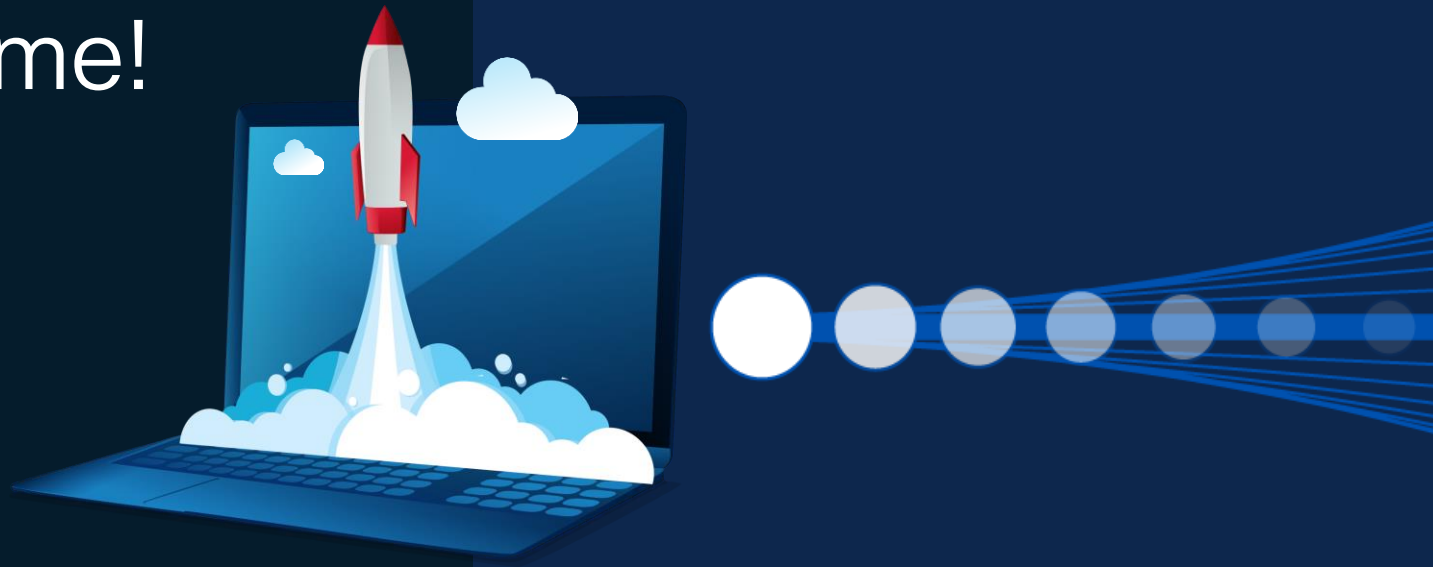
RAFT (Retrieval Augmented Fine Tuning)



[RAFT: Adapting Language Model to Domain Specific RAG (arxiv.org)]

[RAFT (Retrieval Augmented Fine-tuning): A new way to teach LLMs (Large Language Models) to be better at RAG (Retrieval Augmented Generation) (microsoft.com)]

Demo Time!



10.85.134.92 - PuTTY

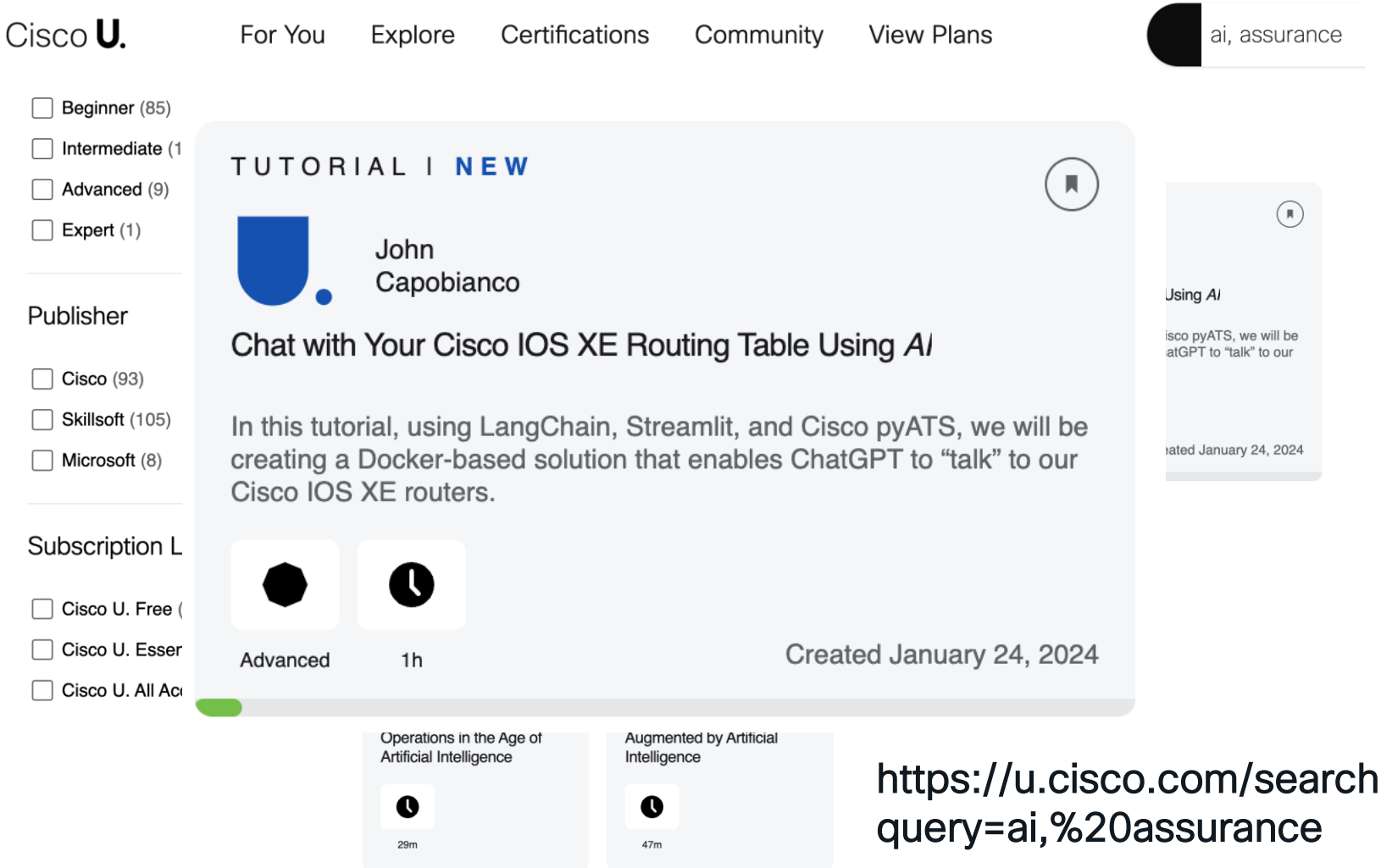
```
| Password:
End of keyboard-interactive prompts from server


VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
```

10.85.134.92 - PuTTY

```
| Password:
End of keyboard-interactive prompts from server


VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
```

10.85.134.92 - PuTTY

```
| Password:
End of keyboard-interactive prompts from server


VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
VNC2-BORDER1-X#
```



https://u.cisco.com/search?query=ai,%20assurance

Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to **win 1 of 5 full conference passes** to Cisco Live 2025.

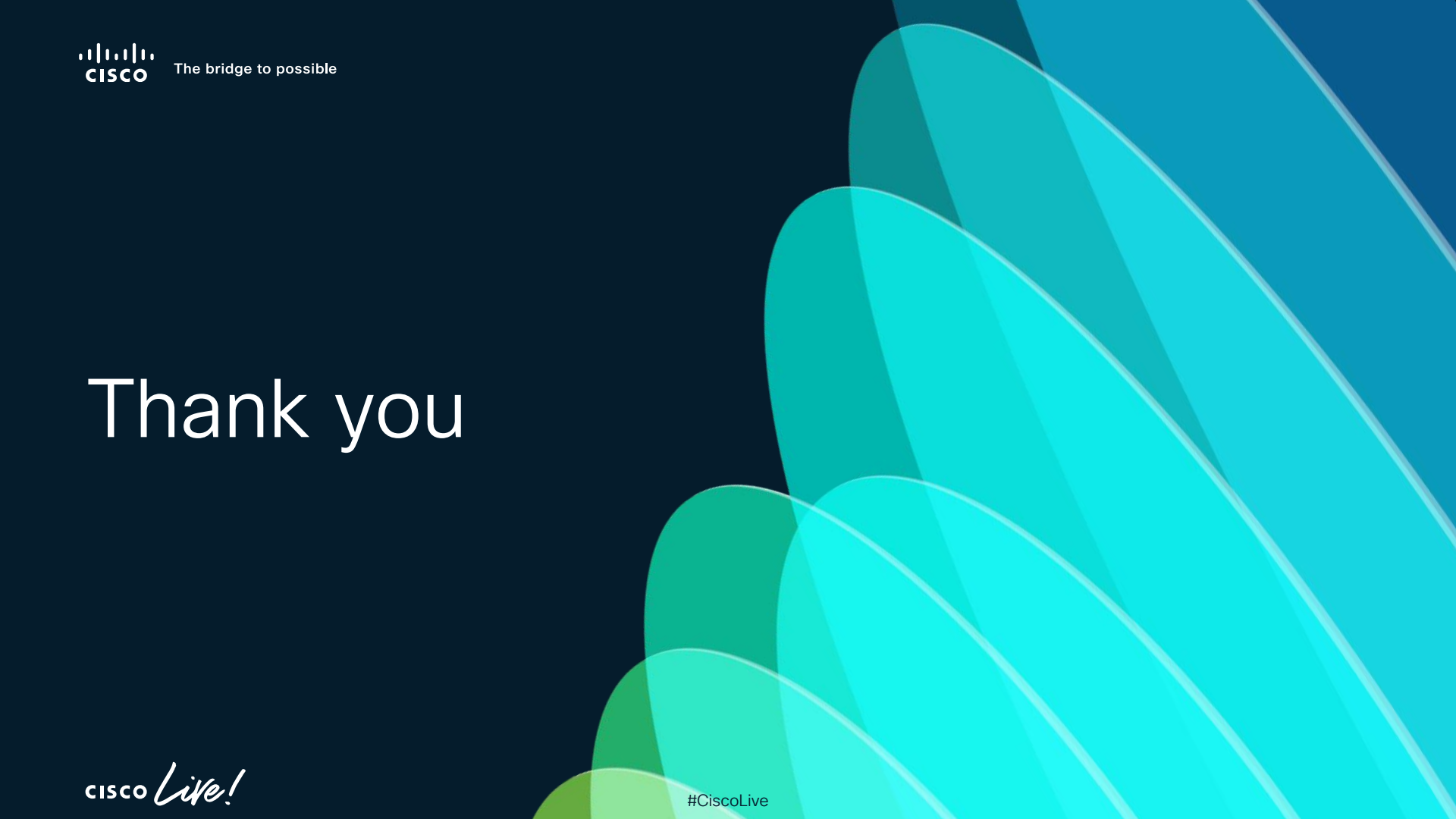
Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn **exclusive prizes!**

Complete your surveys in the **Cisco Live mobile app.**

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand


The bridge to possible
Thank you
cisco Live!
#CiscoLive