EXPLORE

CMTY2254LV

Unlocking the Magic of Gen-AI in VMware VCF

Where dreams meet reality

Tony Foster – Dell Technologies

Gina Rosenthal – Digital Sunshine Solutions

#VMwareExplore #CMTY2254LV

EXPLORE

Please take your survey.

Speakers

Gina Rosenthal

- CEO – Digital Sunshine Solutions
- Virtualized since ~2008 (ESX 3.x)
- Find out more at: DigitalSunshineSolutions.com, TechAunties.com
- **Mastodon:** @gminks@mas.to **LinkedIn**: linkedin.com/in/gminks

Tony Foster (WonderNerd)

- Sr. Principal Engineering Technologist – Dell Technologies
- vExpert, NVIDIA vGPU Community Advisor, Omnissa Tech Insider
- Virtualized since 2005 (ESX 2.0)
- Find out more at: wondernerd.net
- **X:** @wonder_nerd **LinkedIn:** linkedin.com/in/wondernerd

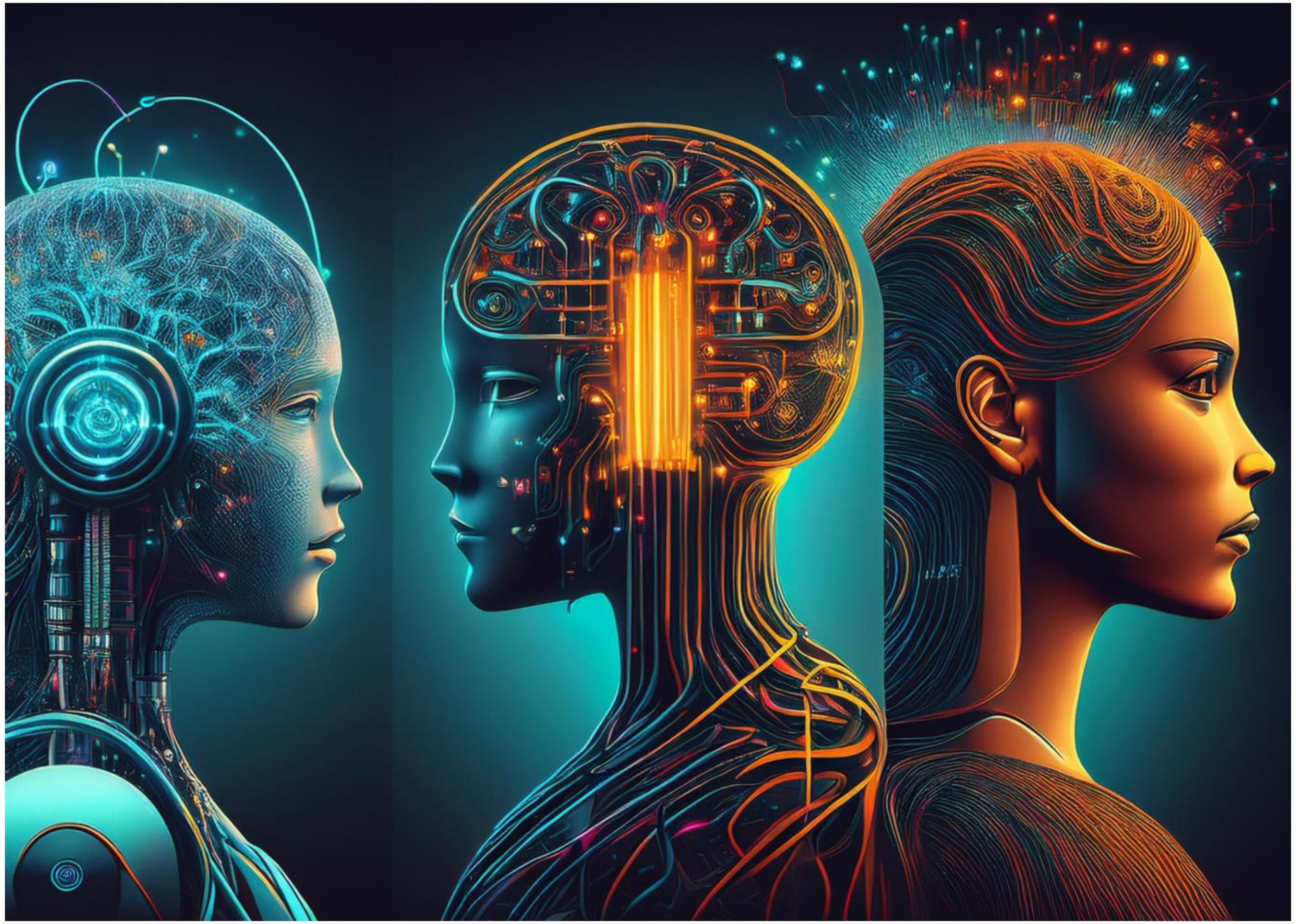
Agenda

- AI Level Set
- What Can the Enterprise Do With Gen-AI?
- How Do LLMs Work
- Putting it All Together

EXPLORE

AI Level Set

Three Types of Artificial Intelligence



Generated in Adobe Photoshop: "three different types of artificial inteligence"

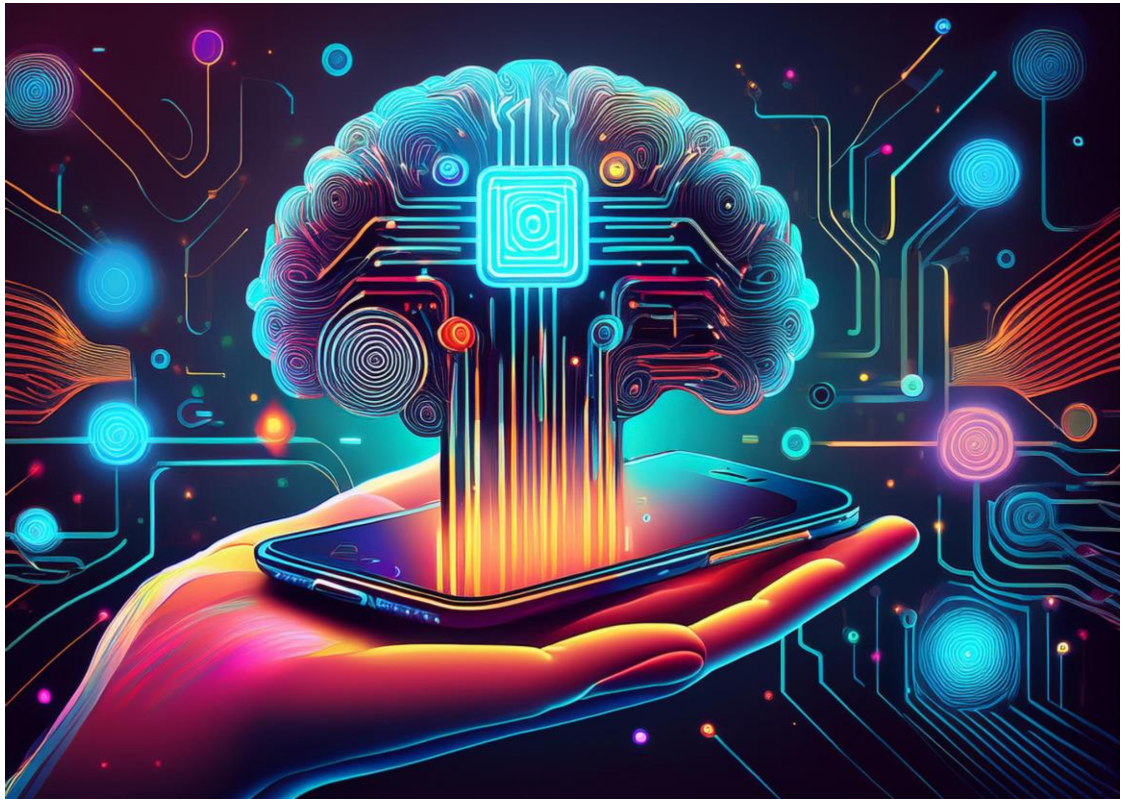
The Three Types of Artificial Intelligence

Narrow (Weak) AI

Artificial General (Strong) Intelligence (AGI)

Artificial Super Intelligence (ASI)

What is Narrow AI?



Generated in Adobe Photoshop: “limited artificial intelligence with a mobile device”

Definition of Narrow AI

- Designed and trained to perform a single, specific task
- Also known as Artificial Narrow Intelligence (ANI) or Weak AI
- The only AI in use today

Characteristics

- Has a narrow focus – built to do a single thing
- Operates under a pre-defined set of rules
- Cannot apply knowledge beyond specific programming

Types of Narrow AI

Reactive

(Chess, recommendation engines, basic instructions to voice assistants)

Generative

(Creates text, images, music, from existing data)

Predictive

(Stock market trends, buying behaviors)

Descriptive

(Business intelligence tools)

Diagnostic

(Analyze health data, root causes of equipment failure)

Limited Memory

(Self driving cars, VAs, recommendation system)

Generative AI is a Narrow AI

It can create new content from existing data

BUT

It still operates within a specific domain or task. It can’t “think” or create anything original.

Generated in Adobe Photoshop: “robot painting a self portrait”

AI for what it is

Via artist Rob Sacchetto's Zombie Art Facebook page

We should stop treating AI like science fiction.
AI is computer science.

We are not animating computers.

We are building infrastructures to run AI workloads.

EXPLORE

Headline

What Can The Enterprise Do With Gen-AI?

Subtitle

vmware EXPLORE

Las Vegas | August 26 – 29, 2024

Barcelona | November 4 – 7, 2023

Enterprise Use Cases

Customer Support & Chatbots

Coding Assistants

Spear Phishing Detection

Marketing Support

EXPLORE

Generative AI Recipe

What you need to enable generative AI in your enterprise

Ingredients

Data & Storage

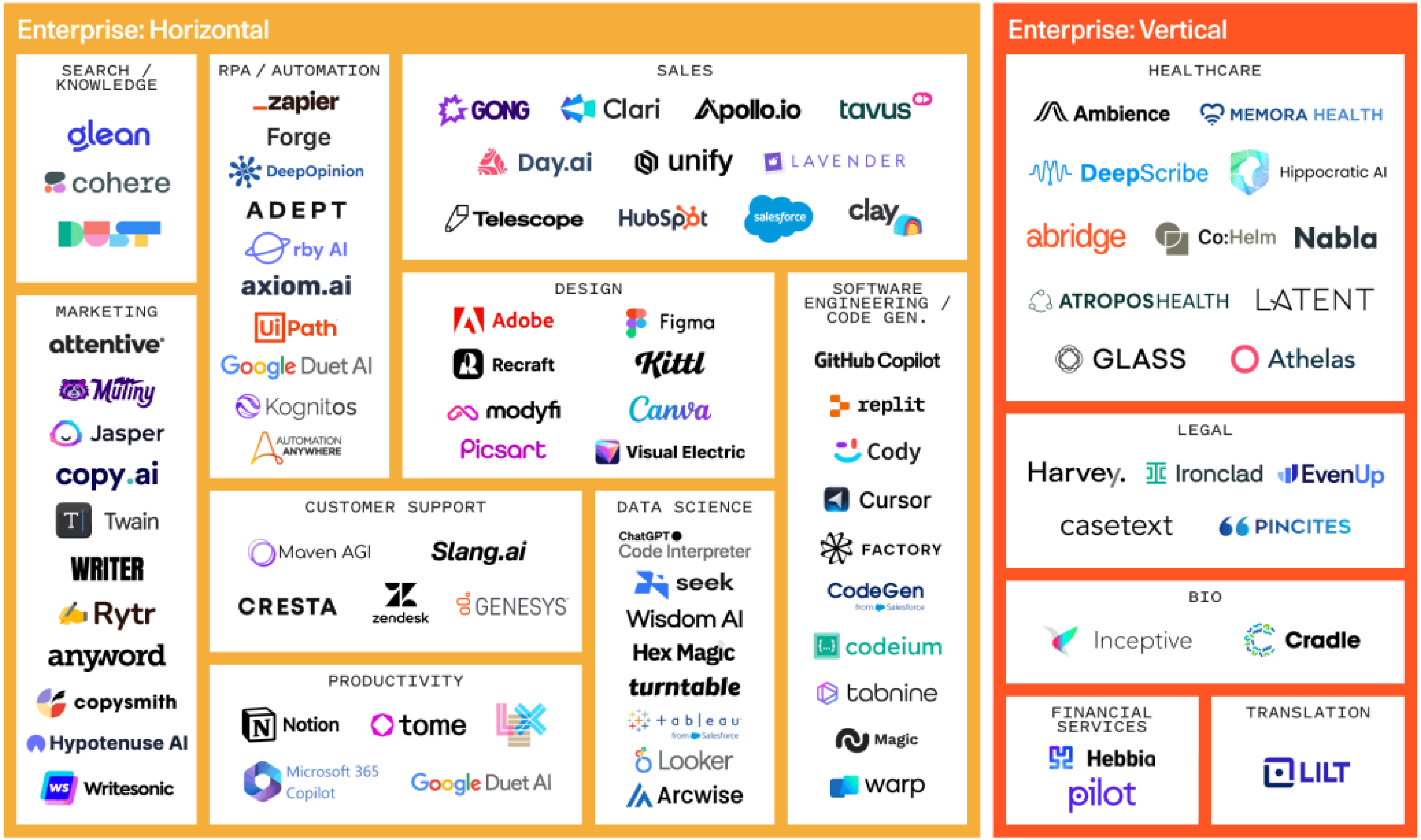
Compute & Accelerators

Algorithms (NLG, NLP,…)

Frameworks (SLM, LLM,…) & Models (RAG)

Pre-trained Models (GPT, BERT, BART, Etc.)

Or Buy Cookie Dough…



Enterprise: Horizontal

SEARCH / KNOWLEDGE
glean
cohere
DUST

MARKETING
attentive
Mutiny
Jasper
copy.ai
Twain
WRITER
Rytr
anyword
copysmith
Hypotenuse AI
Writesonic

RPA / AUTOMATION
zapier
Forge
DeepOpinion
ADEPT
rby AI
axiom.ai
UiPath
Google Duet AI
Kognitos
AUTOMATION ANYWHERE

SALES
GONG
Clari
Apollo.io
tavus
Day.ai
unify
LAVENDER
Telescope
HubSpot
salesforce
clay

DESIGN
Adobe
Figma
Recraft
Kittl
modyfi
Canva
Picsart
Visual Electric

SOFTWARE ENGINEERING / CODE GEN.
GitHub Copilot
replit
Cody
Cursor
FACTORY
CodeGen from Salesforce
codeium
tabnine
Magic
warp

CUSTOMER SUPPORT
Maven AGI
Slang.ai
CRESTA
zendesk
GENESYS

PRODUCTIVITY
Notion
tome
Microsoft 365 Copilot
Google Duet AI

DATA SCIENCE
ChatGPT Code Interpreter
seek
Wisdom AI
Hex Magic
turntable
tableau from Salesforce
Looker
Arcwise

Enterprise: Vertical

HEALTHCARE
Ambience
MEMORA HEALTH
DeepScribe
Hippocratic AI
abridge
Co:Helm
Nabla
ATROPOSHEALTH
LATENT
GLASS
Athelas

LEGAL
Harvey.
Ironclad
EvenUp
casetext
PINCITES

BIO
Inceptive
Cradle

FINANCIAL SERVICES
Hebbia
pilot

TRANSLATION
LILT

Generative AI Market Map 2024

Enterprise section

Via Sequoia Capital

Foundational Components

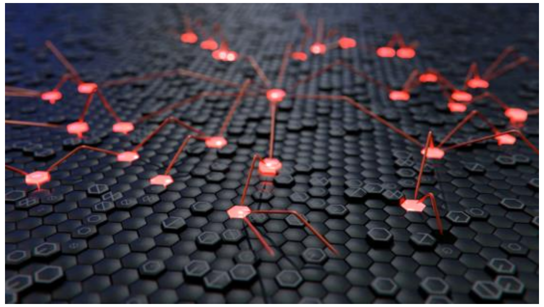
NVIDIA Riva

https://www.nvidia.com/en-us/glossary/text-to-speech/
https://www.nvidia.com/gtc/posters/?search=dell#/

NVIDIA Triton

https://developer.nvidia.com/blog/deploy-an-ai-coding-assistant-with-nvidia-tensorrt-llm-and-nvidia-triton/

NVIDIA Morpheus



https://catalog.ngc.nvidia.com/orgs/nvaie/collections/spear_phishing_detection

EXPLORE

How Do LLMs Work

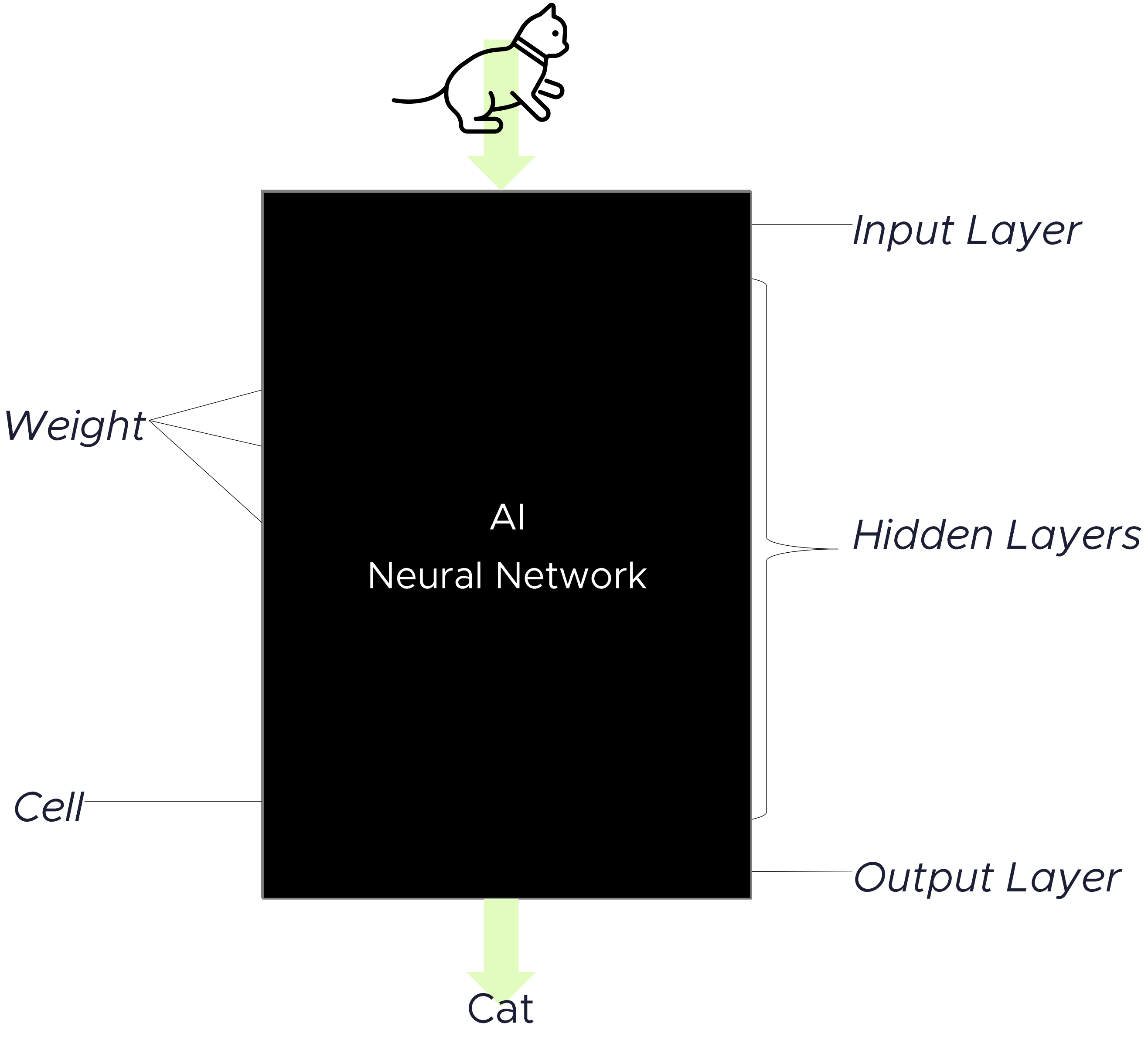
Inside the Black Box of AI Models

Lots of Code

Large Statistics Problem

Solving the Probability of X

Each node is a liner regression model



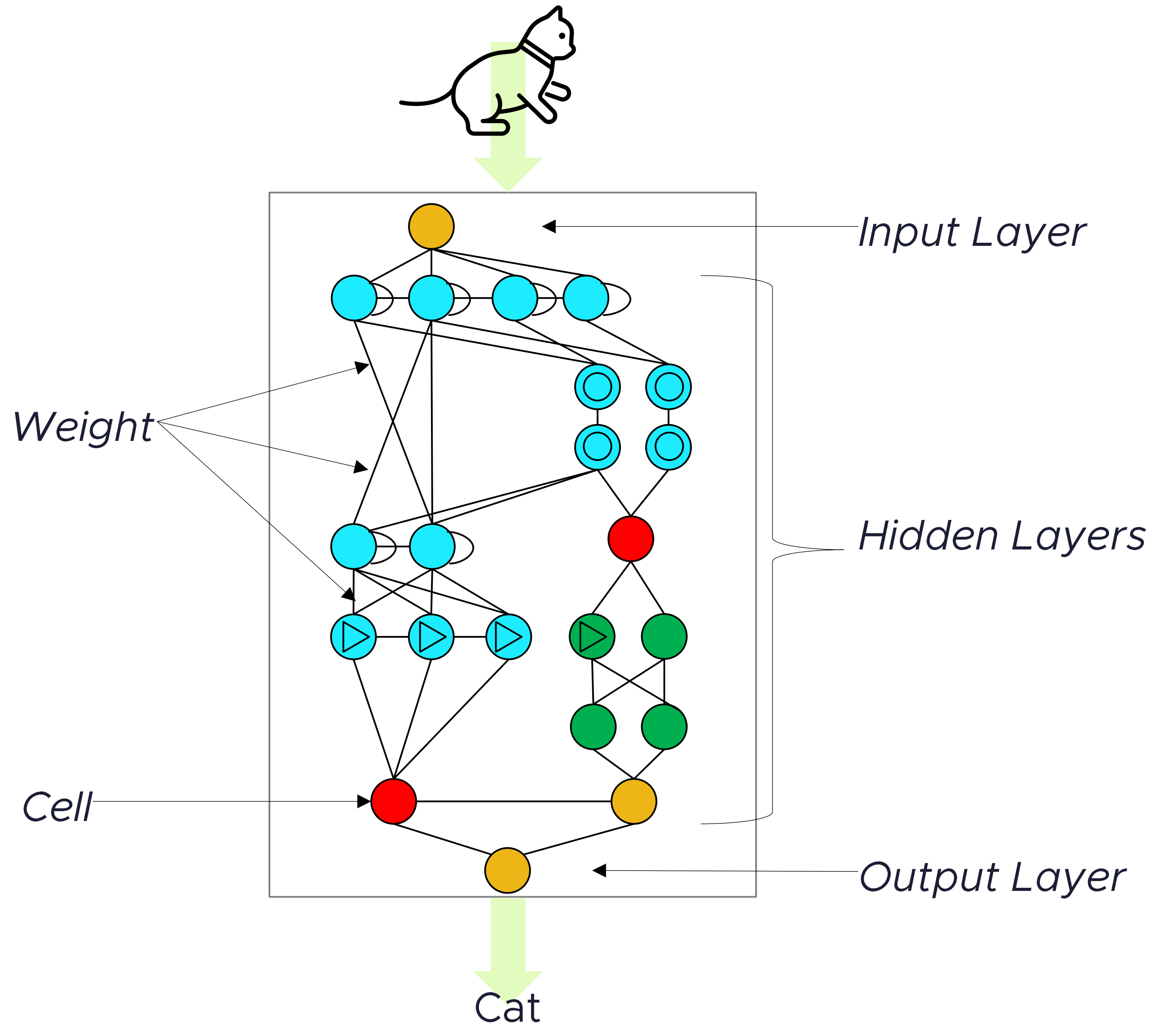
Inside the Black Box of AI Models

Lots of Code

Large Statistics Problem

Solving the Probability of X

Each node is a liner regression model



Navigating AI model black box “gotchas”

Know what the model was designed to do

LLMs handle word content and images

But what if you want analytics?

Was the model trained responsibly?

No standard LLM responsibility benchmarks.

How can you evaluate models?

Can you trust the training data?

Many were trained on internet data.

Can you trust internet data?

If it was collected without consent, does that open up risk?

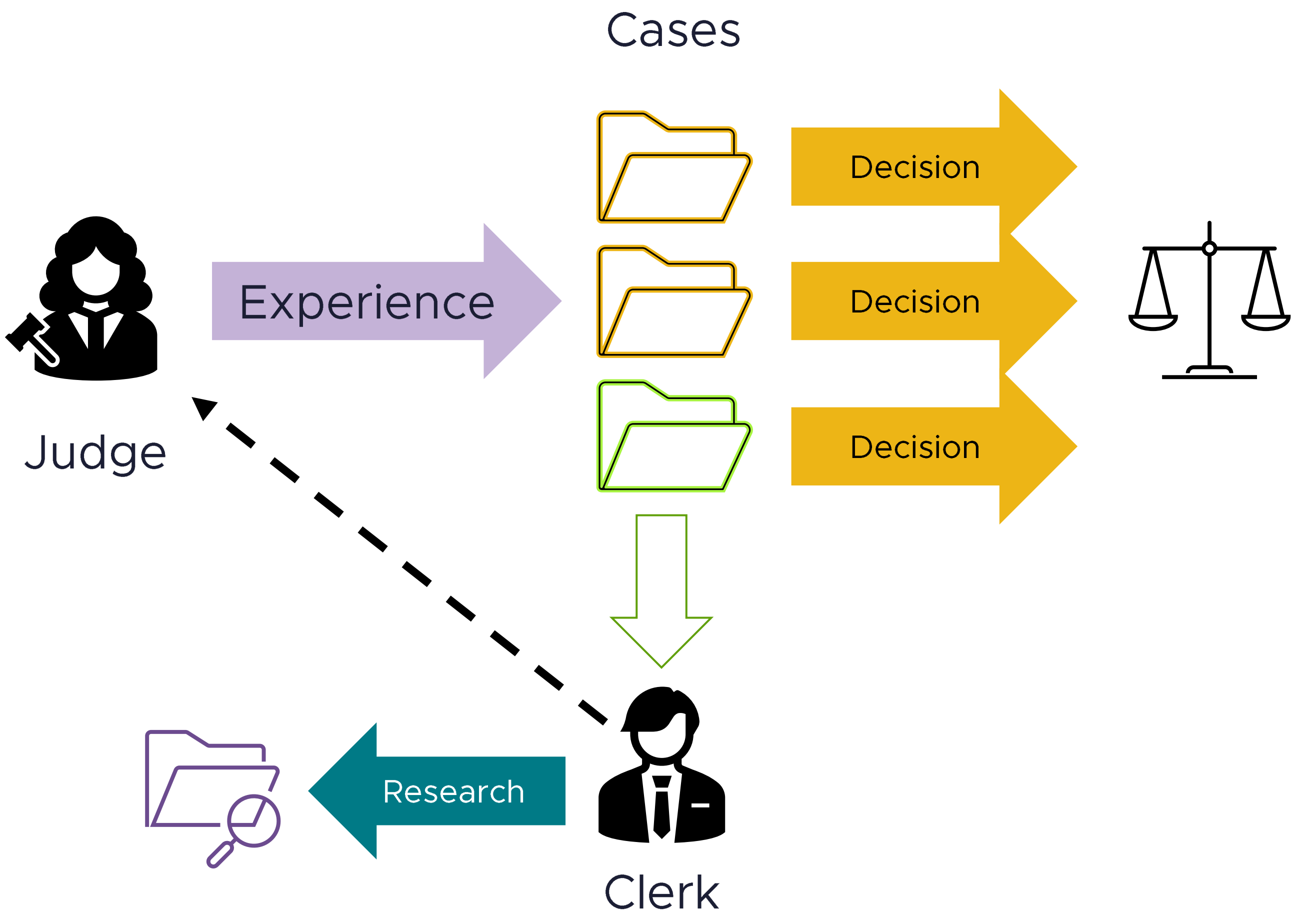
Can an LLM negatively impact your company’s ESG score?

ESG = Environmental, social, and governance frameworks.

It takes an incredible amount of energy to train LLMs.

Bias in training data can perpetuate and amplify existing biases.

Retrieval-Augmented Generation (RAG)



"Retrieval-augmented generation (RAG) is a technique for enhancing the accuracy and reliability of generative AI models with facts fetched from external sources." – Rick Merritt *https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/*

EXPLORE

Headline

Subtitle

Putting it all together

vmware EXPLORE

Las Vegas | August 26 – 29, 2024

Barcelona | November 4 – 7, 2023

Training and Inferencing

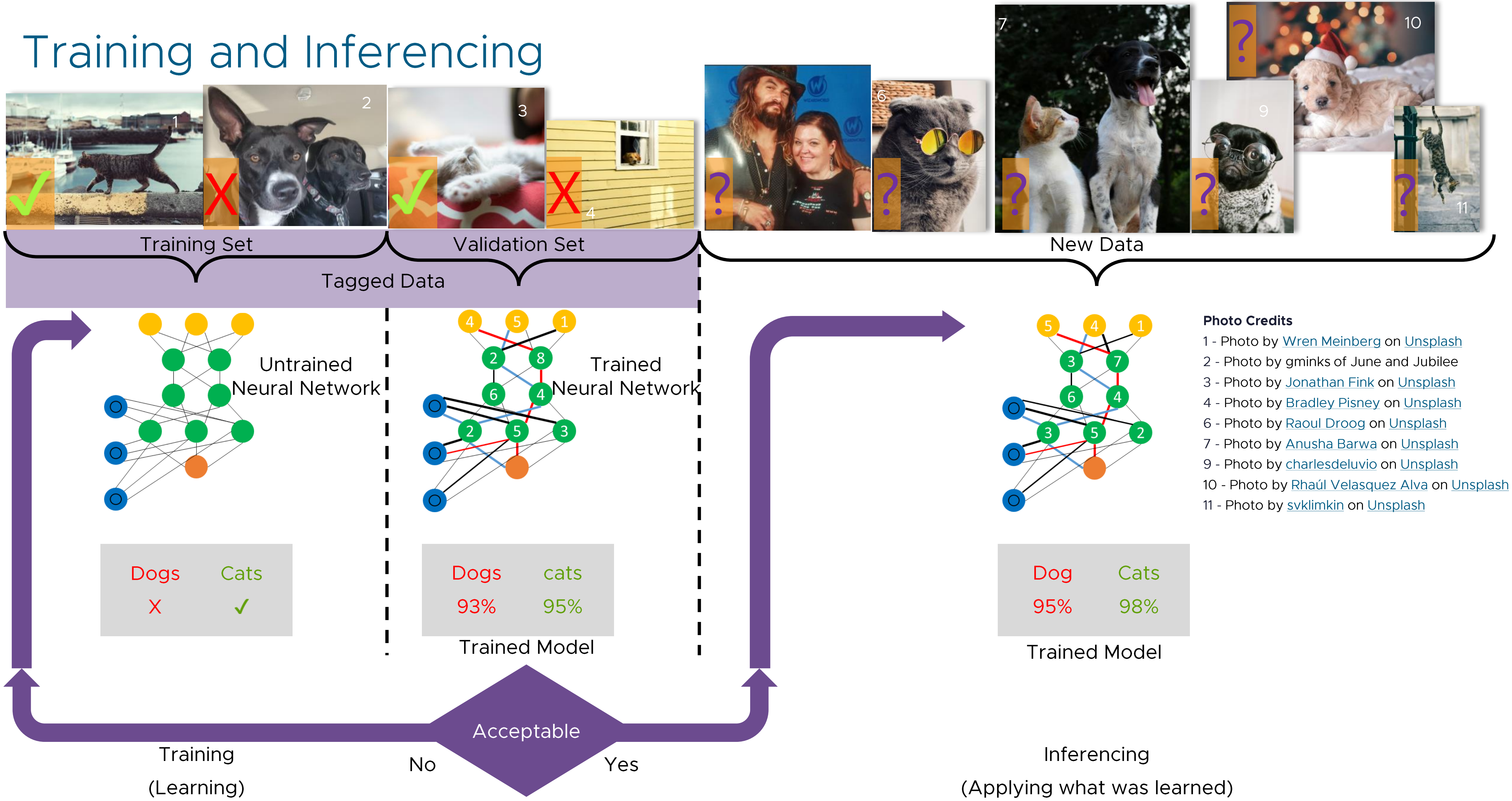


Photo Credits

1 - Photo by Wren Meinberg on Unsplash
2 - Photo by gminks of June and Jubilee
3 - Photo by Jonathan Fink on Unsplash
4 - Photo by Bradley Pisney on Unsplash
6 - Photo by Raoul Droog on Unsplash
7 - Photo by Anusha Barwa on Unsplash
9 - Photo by charlesdeluvio on Unsplash
10 - Photo by Rhaúl Velasquez Alva on Unsplash
11 - Photo by svklimkin on Unsplash

Virtualizing AI With VCF

Start Small and Target Low Hanging Fruit

Virtualize AI Workloads

Expand to The Edge

Monitor and Adjust

EXPLORE

Thank You

Please complete your surveys

Slides are available at wondernerd.net