Istio Day
NORTH AMERICA

Unlock the Full Potential of Generative AI via Microservices and Istio

Lin Sun, Solo.io
Iris Ding, Intel

History of AI

Neural Networks

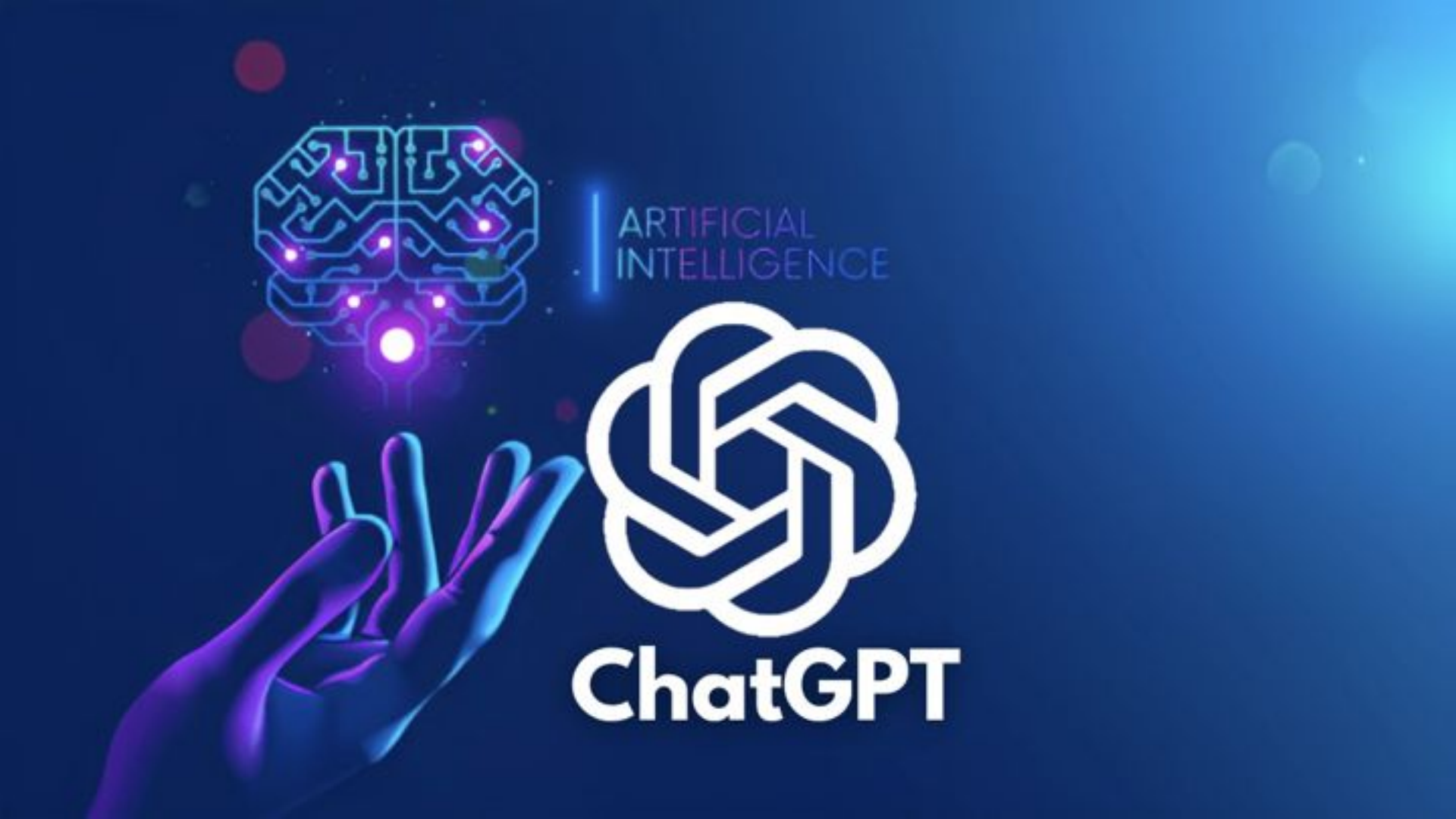
1950-1970

Machine Learning

1980-2010

Deep Learning

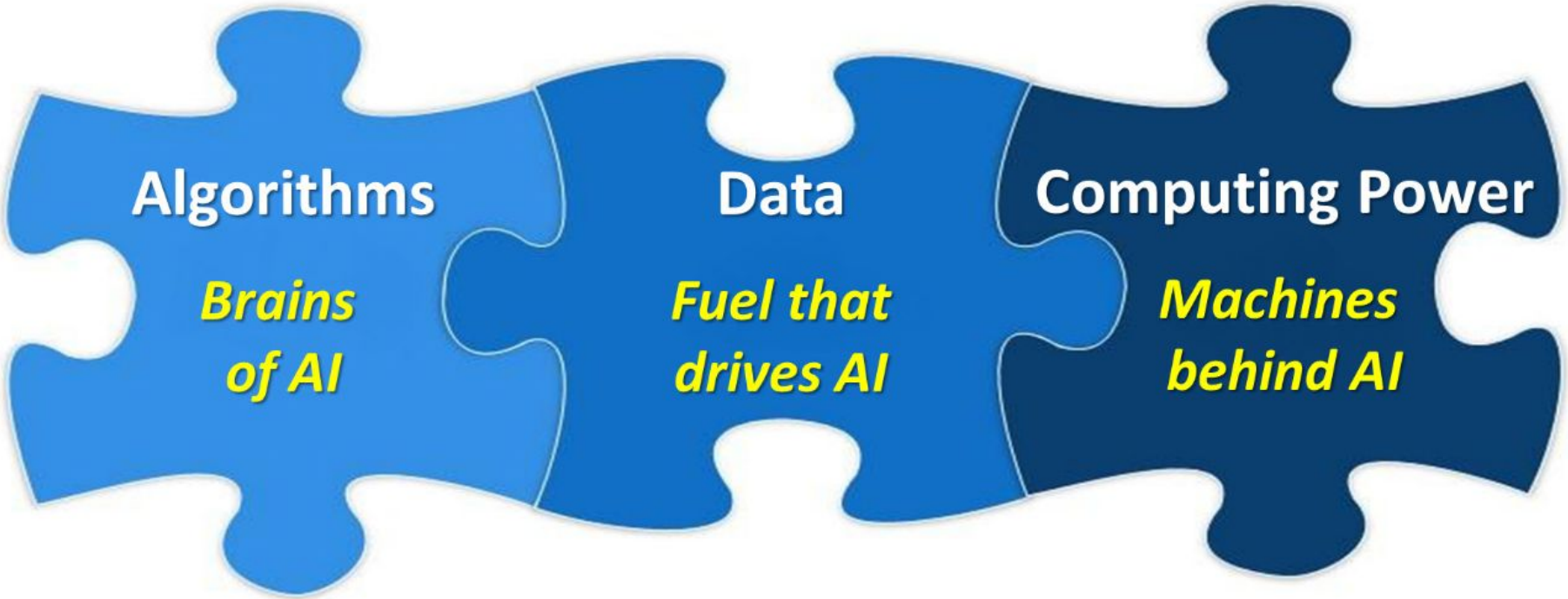
Today


ARTIFICIAL
INTELLIGENCE
ChatGPT

AI Painting (2018)
Edmond de Belamy

AI Painting (2023)
Girl with Glowing Earrings

AI Photo (2023)
Pseudomnesia : The Electrician



Algorithms tell computers what to do. Data tells computers what to learn. Computing power gives machines the power to learn and make decisions

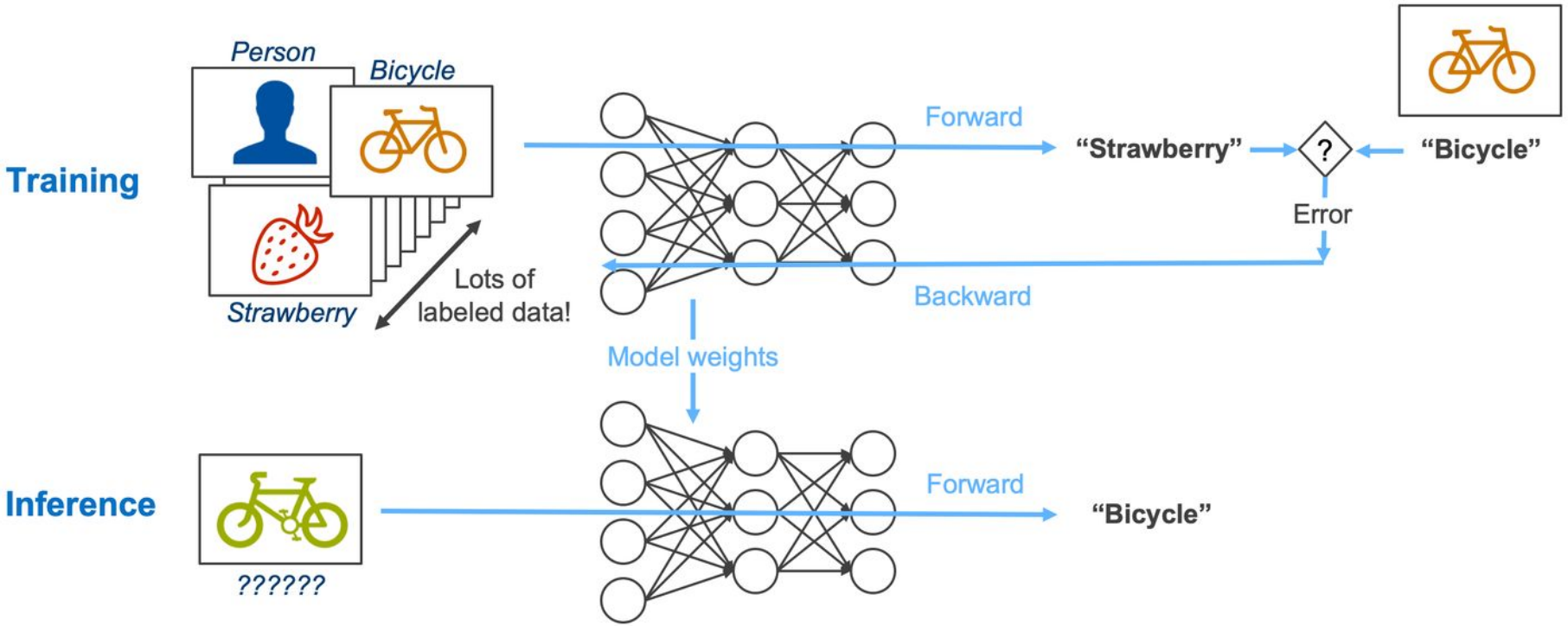
AI Training Techniques

Supervised learning
- Comparing AI-generated outcomes or predictions to the correct answer

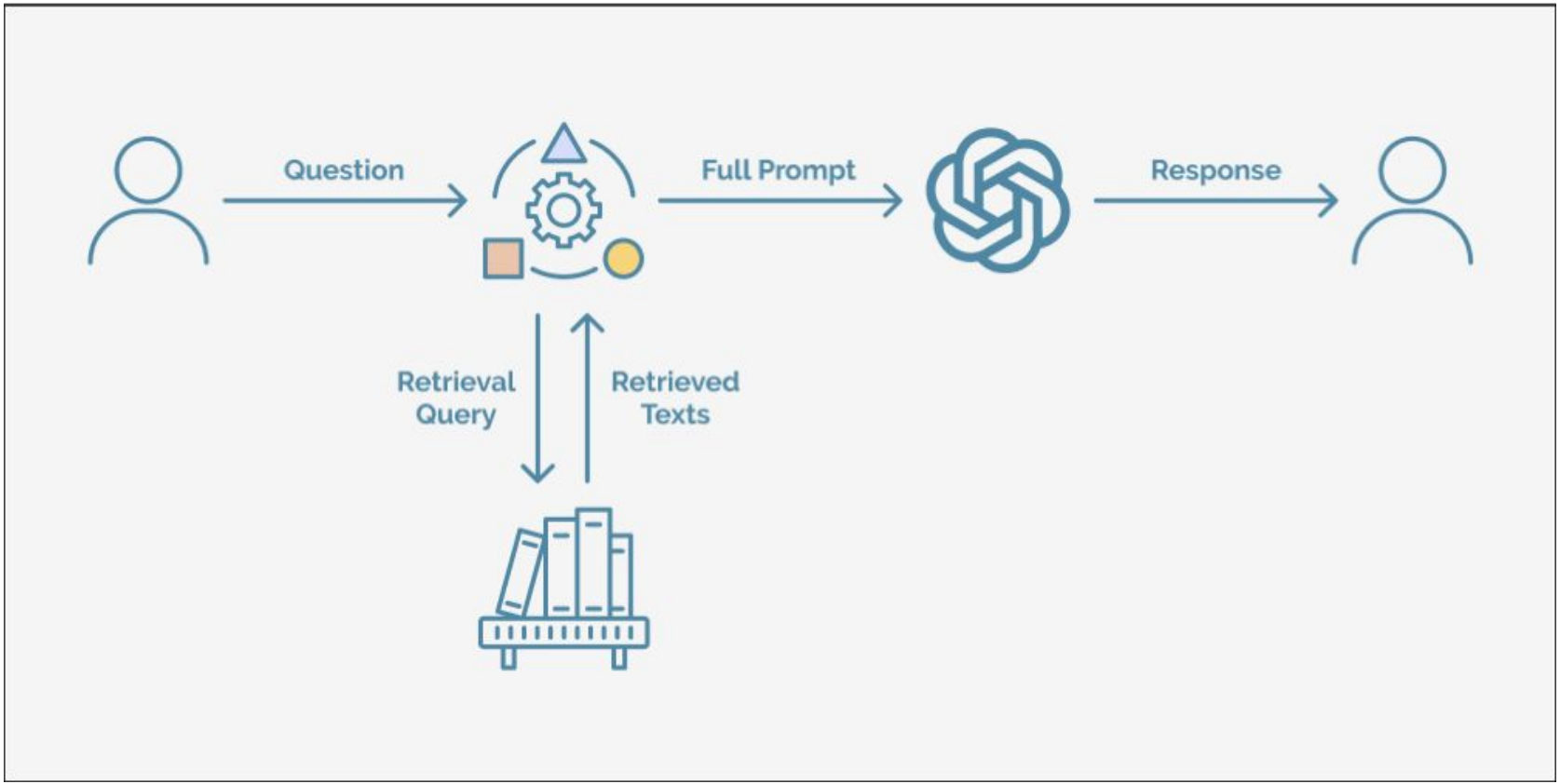
Unsupervised learning
- Finding patterns in data

Reinforced learning
- Trial & Error

Inference


Training
Person
Bicycle
Strawberry
Lots of
labeled data!
Forward
"Strawberry"
?
"Bicycle"
Error
Backward
Model weights
Inference
??????
Forward
"Bicycle"

Retrieval-Augmented Generation (RAG)



Taming Ecosystem Complexity

Ease development and production deployment

- Start with existing/deployed infrastructure
- Scale with new hardware
- Procure new hardware quickly (short lead time)
- Integrate within existing enterprise workflow
- Integrate within ecosystem stack
- Enable fast prototyping and agile deployment of AI apps on existing systems

Speed return on investment

- Fast time-to-result (latency, throughput)
- Result quality (accuracy, effectiveness)
- Low hardware procurement cost
- Low operational cost

Secure

- On-prem data center support
- Virtual private cloud support
- End-to-end data security

End-to-end supported use cases

- Augmented services (e.g., QnA, generation, summarization)
- Defect/Anomaly detection, Theft prevention
- Fraud detection, Pattern anomaly
- Research & discovery

https://github.com/opea-project

OPEA core intent

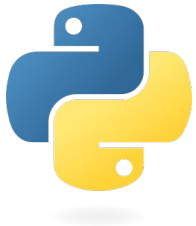
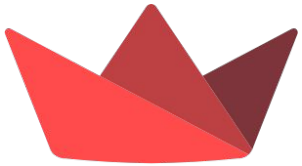
Construction of GenAI solutions, including retrieval augmentation:

- Set of building blocks to be used in the composition of GenAI solutions.
  - GenAI models – Large Language Models (LLMs), Large Vision Models (LVMs), etc.
  - System components – e.g., Embedding Models; Vector DB; Ranking, Prompt processing, and more
- Set of compositional capabilities for building AI agents & creating full end-to-end GenAI flows
- Tools for fine-tuning, customizing and optimizing, including for datacenter/on-prem settings
- A variety of validated, ready for deployment end-to-end reference flows

Evaluation of GenAI solutions, including retrieval augmentation:

- Means and services to fully evaluate and grade components and end-to-end GenAI solutions
  - **Assessment** – Detailed tests done for particular modules or attributes of the end-to-end flow.
  - **Grading** - Aggregation of the individual assessments to a grade per each of the four domains -
    - ✓ Performance
    - ✓ Features
    - ✓ Trustworthiness
    - ✓ Enterprise-readiness
  - **Certification** (if offered) - meeting a minimum level of grading on all four domains.

Demo



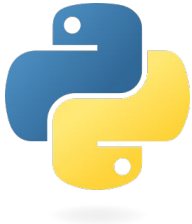
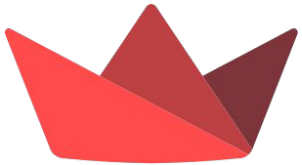





spiffe

gateway api

Demo Recap






spiffe



gateway api


Questions

Istio Day
NORTH AMERICA