Are You Ready for AI?:

A Guide to Running AI Workloads Smoothly and Securely

Anjelica Ambrosio, You.com
Christian Hernandez, Cisco

Presenters

Anjelica Ambrosio
DevRel Engineer, You.com

Christian Hernandez
Technical Marketing Engineer, Cisco

You.com

AI Pioneers

Our founders have pushed the frontier of AI, NLP, and LLMs for over a decade.

salesforce NVIDIA NORWEST VENTURE PARTNERS

Richard Socher
CEO & Co-Founder
Former Chief Scientist & EVP at Salesforce & Stanford Professor

Bryan McCann
CTO & Co-Founder
Former Lead AI Researcher at Salesforce

2014
Stanford research pioneered deep learning + NLP with Nvidia GPUs

2016
Socher appointed Chief Scientist to lead AI at Salesforce

2018
Revolutionized NLP and search with contextualized word vectors

2021
First to incorporate GenAI in search - patent pending

2022
First LLM fully connected to the web for real-time knowledge

2023
LM Firsts: Personalization, with a full Web API

2024
Launches LLM selection and multi-modal input

You.com supports the world's leading AI teams

Our APIs operate at **true internet scale** for customers **globally**.

Scale

1B+

Queries served monthly

Engagement

5K+

API customers to date

Enterprise Grade

57%

Fortune 500 penetration

OpenAI

Alibaba

DuckDuckGo

The Telegraph

amazon

databricks

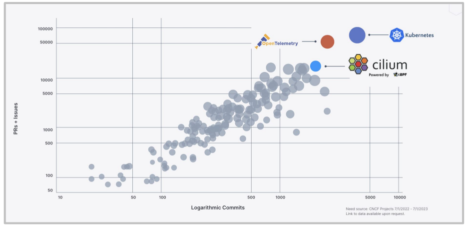
Our Story:

We Are Isovalent, now Cisco – Creators of eBPF & Cilium

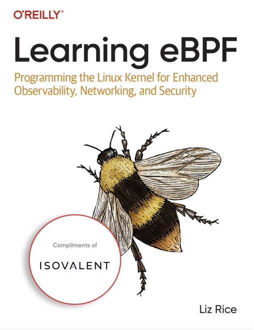
- Founded in 2017, Isovalent is the enterprise company behind Cilium.
- We are passionate about reinventing **Networking & Security for the Cloud Native Era with eBPF**.

THOMAS GRAF
Co-founder & CTO

DAN WENDLANDT
Co-founder & CEO



- Linux Networking & Security – eBPF maintainer
- Kubernetes & Distributed Systems
- Open-source community



WILD LIFE · PostFinance · Trip.com · IKEA · aws
Alibaba.com · radiofrance · Capital One · Google · sky
FORM3 FINANCIAL CLOUD · overstock · Adobe · telenor · Bell
DigitalOcean · SAP Concur · COGNITE · DATADOG

Agenda

1. Introduction
2. The Rise of AI
3. Why AI Breaks Traditional Infrastructure
4. 3 Common Mistakes
5. Building the Right Foundation
6. Enforcing Security and Compliance
7. How to Add Live Data to Your AI
8. High Level Overview
9. Who Benefits, and How?
10. Closing

The Rise of AI

AI adoption is accelerating, but infrastructure readiness is lagging.

The Rise of AI

Core Problem:

Enterprises are rapidly adopting AI without proper infrastructure foundations

Challenges

- Overprovisioned GPUs drive high costs
- Legacy VM mindset limits scale
- Weak inter-node traffic security
- Compute-heavy distributed AI workloads
- Sensitive data requires compliance controls

Why AI Breaks Traditional Infrastructure

- Extreme GPU and compute demand
- High performance networking required
- Strict security and compliance needs
- AI workloads exceed VM design limits

Architecture Layers of Generative AI system

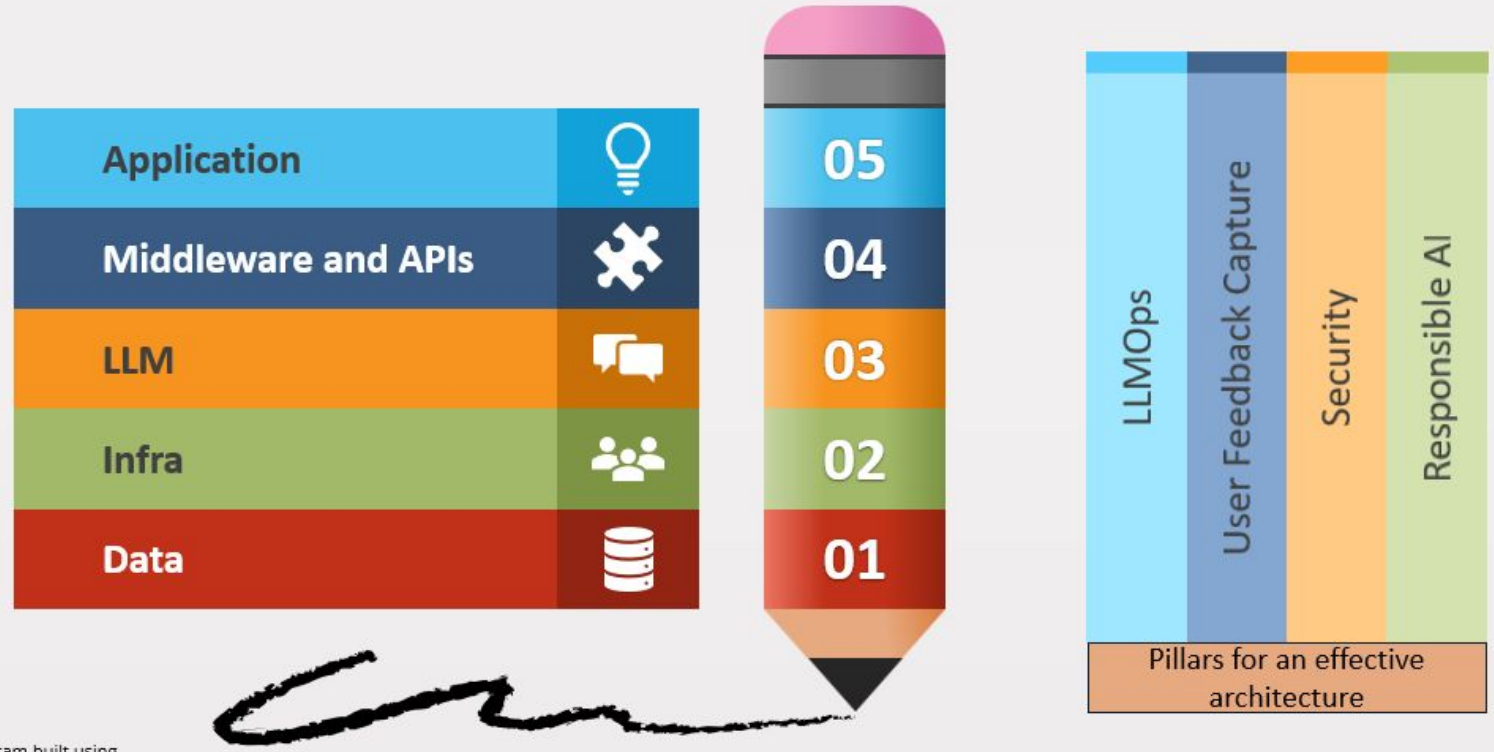


Diagram built using
PRESENTATIONGO
.com

3 Common Mistakes

The three common mistakes enterprises make when building AI infrastructure.

3 Most Common Mistakes

1. Overprovisioning Resources
   - Resource overprovisioning
   - Idle GPUs
   - Fear-driven overplanning

2. Legacy Infrastructure Mindset
   - Poor alignment with K8s
   - Limited flexibility for AI workloads

3. Insufficient Security
   - Increased risk of data leaks
   - Network security limited to isolation

Building the Right Foundation

Build the right foundation for your AI workloads with Kubernetes + k0rdent.

Why Use Kubernetes for AI?

- Scalable distributed workload orchestration
- Efficient GPU resource scheduling
- Automated deployment and scaling
- Cloud-native infrastructure flexibility

You Can’t Just Run AI on Kubernetes As Is!

Bottlenecks 😕:

01 | GPU Scheduling

02 | Long-Running & Stateful

03 | Complex Workflows

04 | Data Movement

What is k0RDENT

Open-source Kubernetes-native platform by Mirantis

- Unified Kubernetes management (cloud, on-prem, edge)
- GitOps automation
- Built-in governance & policies
- Built for distributed + AI workloads
- Observability & FinOps visibility (via KOF)

k0rdent Observability & FinOps (KOF)

Clusters: Select clusters... Nodes: Select nodes... Search by endpoints, labels or scrape pool

0 Unknown 0 Down 6 Up

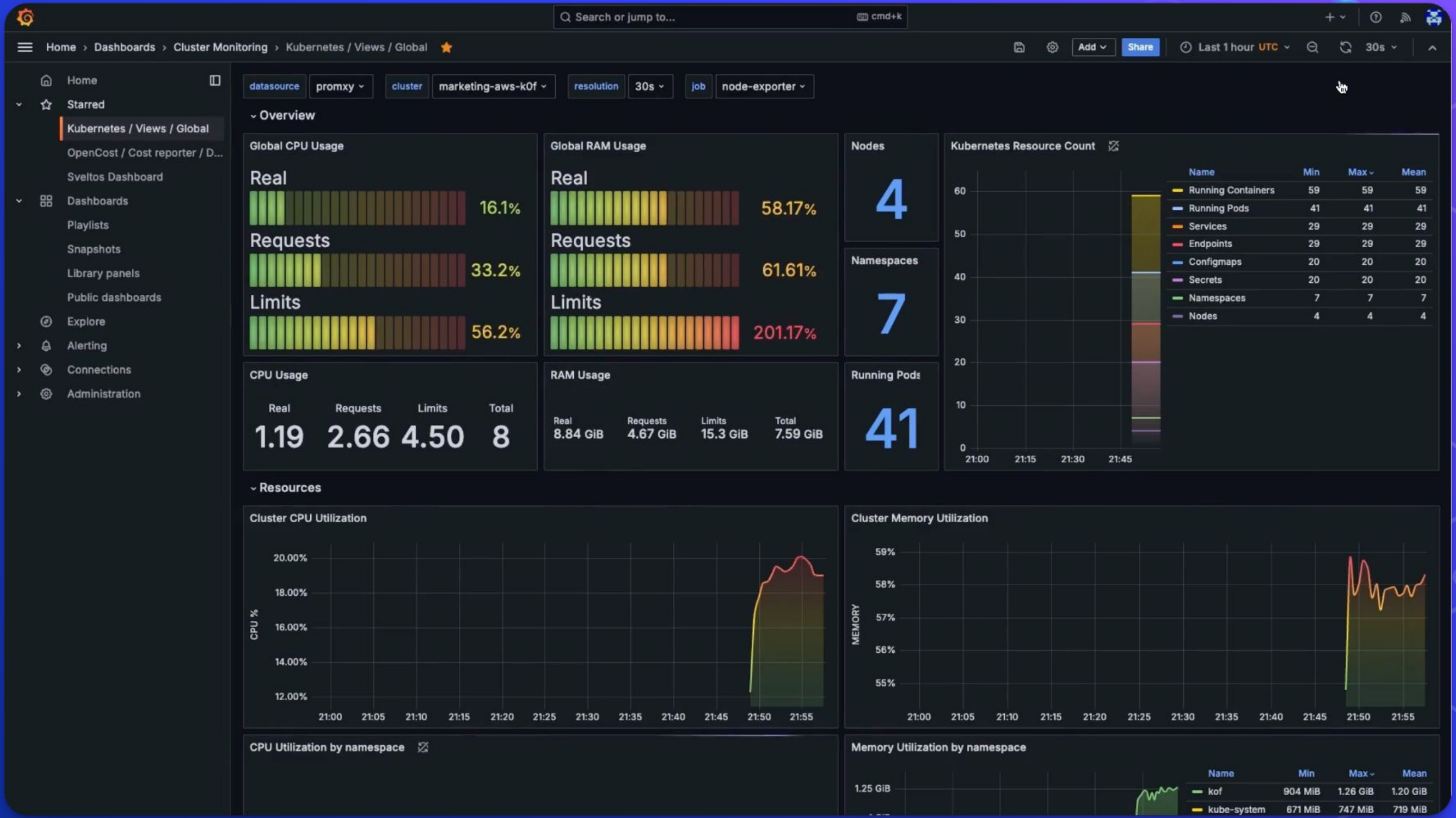
regional-adopted

0 Unknown • 0 Down • 6 Up

Endpoint	State	Labels	Last Scrape	Scrape Duration	Error
https://kubernetes.default.svc.cluster.local/api/v1/nodes/regional-adopted-control-plane/proxy/metrics/cadvisor	Up	2 Labels	a few seconds ago	81.281ms	
https://kubernetes.default.svc.cluster.local/api/v1/nodes/regional-adopted-control-plane/proxy/metrics	Up	2 Labels	a few seconds ago	32.223ms	
https://172.18.0.5:6443/metrics	Up	5 Labels	a few seconds ago	165.283ms	
http://kof-collectors-kube-state-metrics-86d4cdf499-jg266:8080/metrics	Up	7 Labels	a few seconds ago	18.618ms	
http://kof-collectors-opencost-68b6dc8ccb-fs22g:9003/metrics	Up	7 Labels	a few seconds ago	4.239ms	
http://kof-collectors-prometheus-node-exporter-xc2qr:9100/metrics	Up	8 Labels	a few seconds ago	23.250ms	

KOF UI via OSS k0rdent

k0rdent Observability & FinOps (KOF)

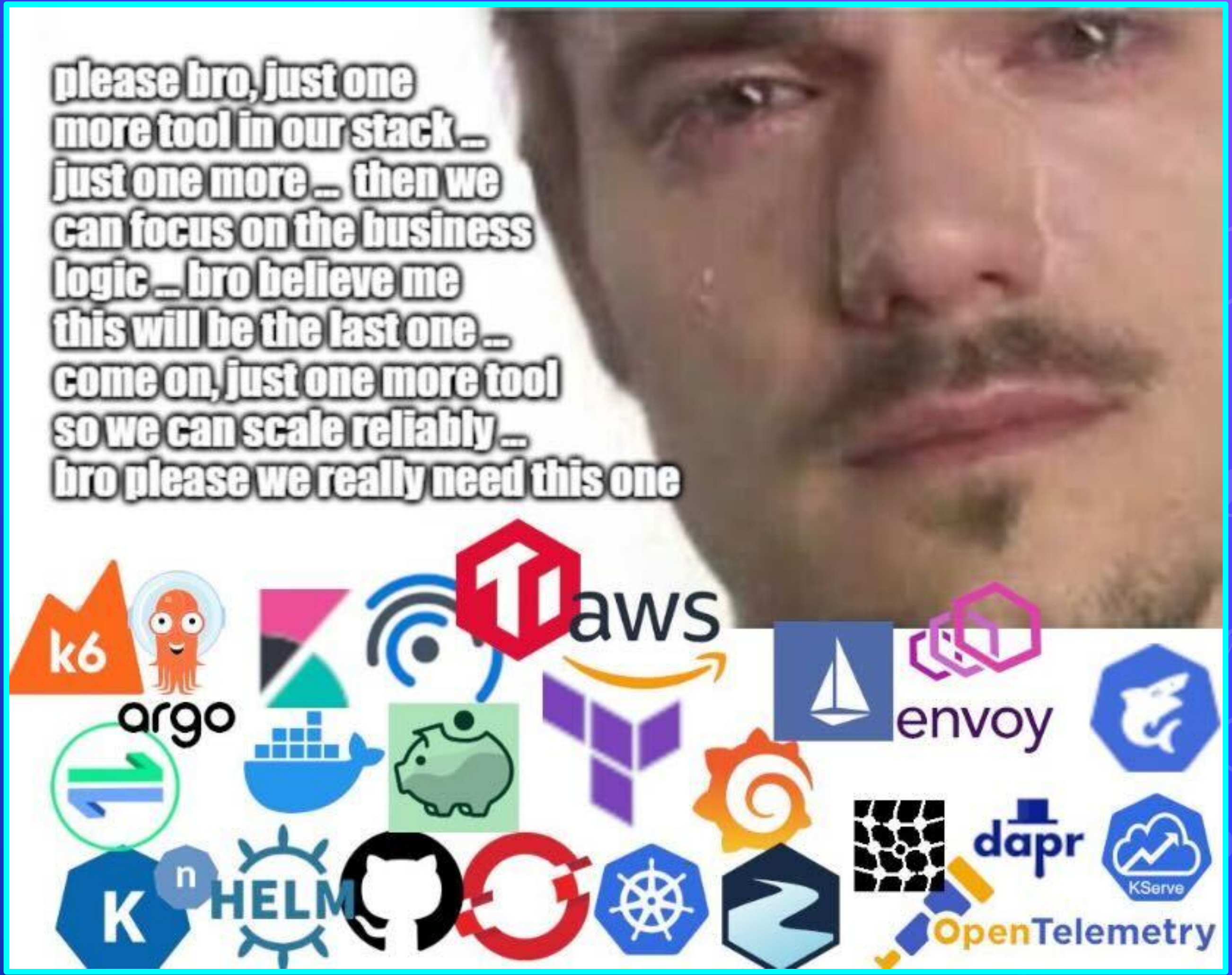


Screenshot by Xnapper.com

KOF with Grafana

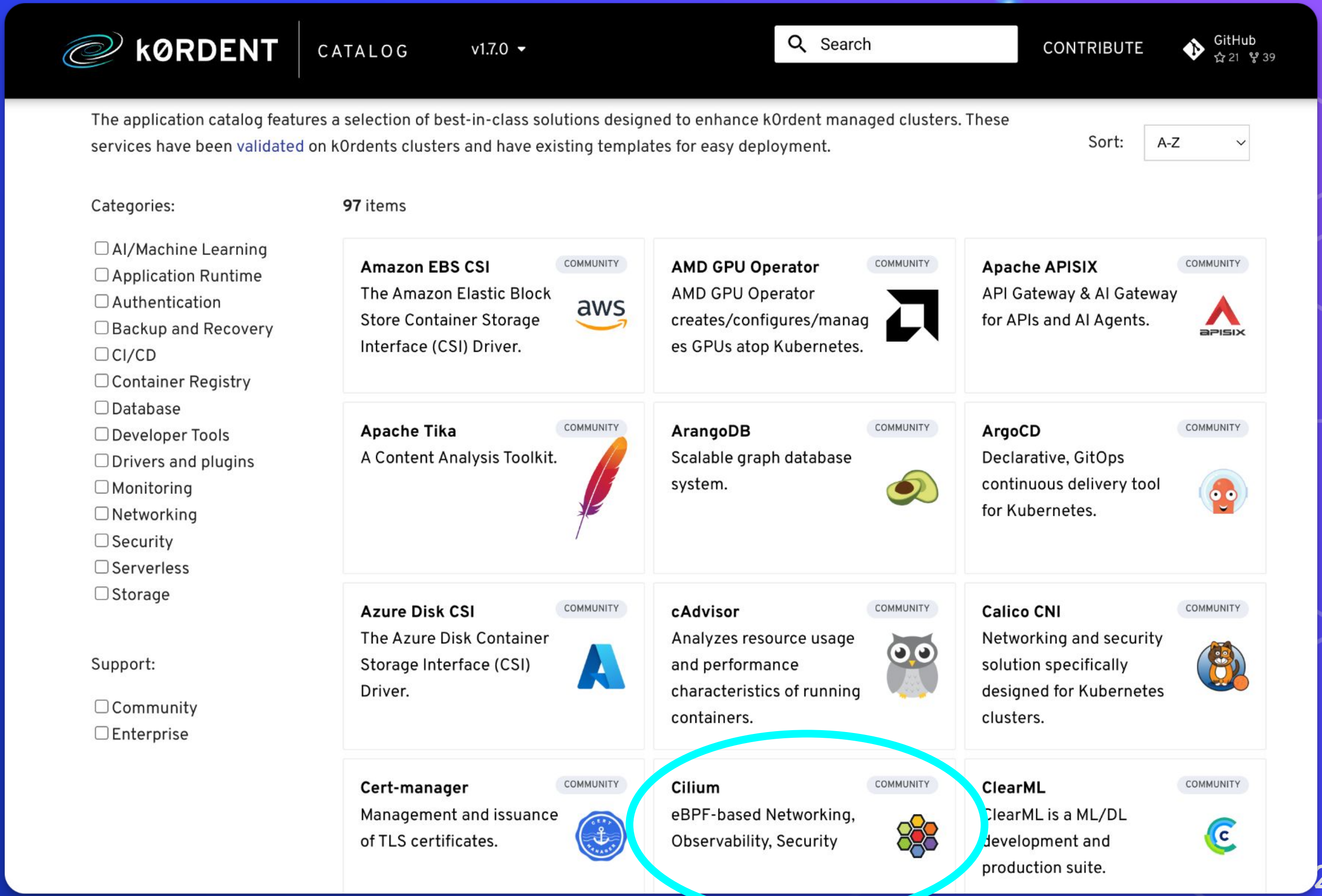
Outcomes: k0rdent + Kubernetes

- Scalable, multi-cloud AI infrastructure
- Centralized cluster management
- Faster model deployment (GitOps)
- Stronger security & policy control
- Lower operational complexity
- Visibility into resource usage and costs via KOF


please bro, just one
more tool in our stack...
just one more ... then we
can focus on the business
logic ... bro believe me
this will be the last one ...
come on, just one more tool
so we can scale reliably...
bro please we really need this one
k6
argo
aws
envoy
dapr
HELM
KServe
OpenTelemetry

Use Your Favorite Services with Ease

Check out k0rdent!



Enforcing Security and Compliance with Cilium

Keep your data safe and compliant with Cilium in the age of AI

What is Cilium?

Service Mesh

cilium Service Mesh

RESILIENCE	TRAFFIC MANAGEMENT	AUTHENTICATION
Ingress	Gateway API	spiffe

Observability

cilium Hubble

METRICS	TRACING	SERVICE MAP	LOGS

SIEM

JSON

fluentd

Grafana

OpenTelemetry

Networking

cilium CNI

ENCRYPTION	LOAD-BALANCING	NETWORK POLICY	NETWORKING
IPsec **Wireguard**	**Kubernetes** **DSR • Maglev**	**DNS • L7** **L3/L4**	**IPv4 • IPv6 • Overlay • BGP** **Egress Gateway • Multi-Cluster** **Bandwidth Management**

Runtime Security

OBSERVABILITY

ENFORCEMENT

tetragon

SIEM

JSON

fluentd

Kubernetes | **Container** | **VM** | **Metal**

aws

Google Cloud

Azure

Alibaba Cloud

Red Hat OpenShift

vmware

How Can We Secure AI Systems Against Evolving Threats?

- MITRE ATLAS Framework
  - T0018: Backdoor ML Model Tampering
  - T0010: ML Supply Chain Compromises
  - T0015: Data and Model Poisoning

MITRE ATLAS https://atlas.mitre.org/

How Can We Secure AI Systems Against Evolving Threats?

- Top 10 Risk & Mitigations for LLMs and Gen AI Apps
  - LLM01: Prompt Injection (PoisonGPT Attack)
  - LLM03: Supply Chain (PyTorch Supply Chain Attack)
  - LLM04: Data & Model Poisoning (Probllama CVE-2024-370032)

- Paired it with MITRE ATLAS

ML Supply Chain Compromise

- Attack Vector: Compromise of third-party ML dependencies and data sources
- Real Example: PyTorch supply chain attacks targeting ML developers
- Defense: FQDN network policies control external data sources w/ Cilium

Mitigation Strategy:

```yaml
# Cilium Network Policy for Data Source Control
apiVersion: cilium.io/v2
kind: CiliumNetworkPolicy
spec:
  egress:
  - toFQDNs:
    - matchName: "azureopendatastorage.blob.core.windows.net"
```

Backdoor ML Model

- Attack Vector: Manipulation of training data to insert hidden triggers
- Real Example: PoisonGPT attack on Hugging Face models
- Defense: eBPF-based file monitoring detects training data tampering

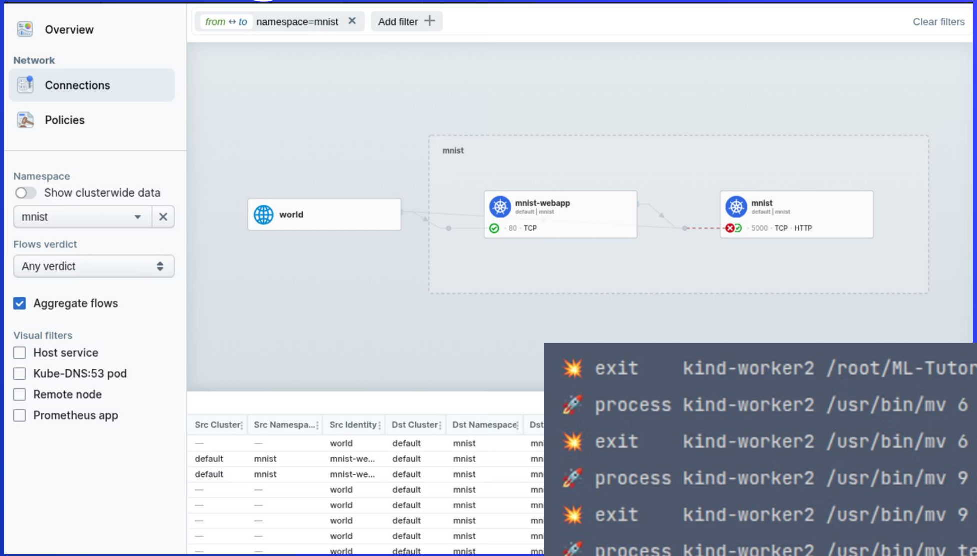
```yaml
# Tetragon TracingPolicy for File Integrity Monitoring
apiVersion: cilium.io/v1alpha1
kind: TracingPolicy
spec:
  kprobes:
  - call: "security_file_open"
    syscall: false
    args:
    - index: 0
      type: "file"
```

LLM Prompt Injection

- Attack Vector: Adversarial prompts manipulating model behavior
- Real Example: CVE-2024-37032 (Probllama) path traversal via model uploads
- Defense: Runtime monitoring detects malicious LLM behavior patterns

```yaml
# Tetragon TracingPolicy for Path Traversal Detection
apiVersion: cilium.io/v1alpha1
kind: TracingPolicy
spec:
  kprobes:
  - call: "security_path_truncate"
    args:
    - index: 0
      type: "path"
```

Using Cilium and Tetragon Together




```
💥 exit    kind-worker2 /root/ML-Tutorial/poison_attack.sh /root/ML-Tutorial/poison_attack.sh 0
🚀 process kind-worker2 /usr/bin/mv 6 temp_backup_6
💥 exit    kind-worker2 /usr/bin/mv 6 temp_backup_6 0
🚀 process kind-worker2 /usr/bin/mv 9 6
💥 exit    kind-worker2 /usr/bin/mv 9 6 0
🚀 process kind-worker2 /usr/bin/mv temp_backup_6 9
💥 exit    kind-worker2 /usr/bin/mv temp_backup_6 9 0
🚀 process kind-worker2 /usr/bin/head -1 6/sample_1.jpg
💥 exit    kind-worker2 /usr/bin/head -1 6/sample_1.jpg 0
🚀 process kind-worker2 /usr/bin/head -1 9/sample_1.jpg
💥 exit    kind-worker2 /usr/bin/head -1 9/sample_1.jpg 0
💥 exit    kind-worker2 /root/ML-Tutorial/poison_attack.sh /root/ML-Tutorial/poison_attack.sh 0
```

Make Your AI Smarter with Real-Time Search with you.com

Presented by: Anjelica Ambrosio

Search API for the LLM Age

You.com's APIs empower developers to build more accurate, faster, and fresher AI experiences—at a lower cost.

01 One API, All of the Data

Seamlessly blend the latest web and news search from leading sources

02 Long, Rich Snippets

Get rich context for every result; perfect for LLMs & agentic workflows

03 Real-Time Freshness

Answer questions with leading recency and semantic accuracy

04 Enterprise Reliability

Built for scale, with robust uptime and unified billing

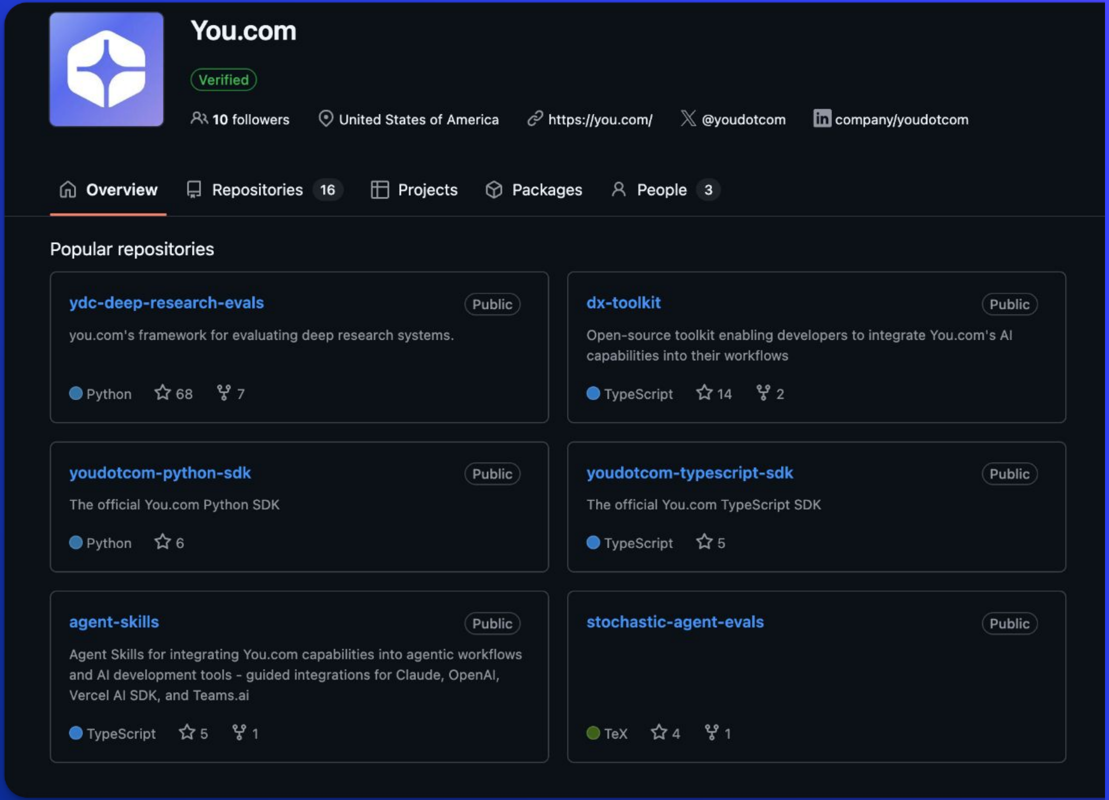
You.com Search API: What do we solve for?

Use cases generalize across numerous agentic workflows.

Guide to Our APIs

OpenAI	De facto Web Search API provider for OpenAI OSS model.
DuckDuckGo	News API to serve the latest breaking news for all their US visitors.
Alibaba.com	Web Search API with Accio, Alibaba's AI-powered B2B sourcing agent, to acquire the latest product information enhancing sourcing accuracy.
databricks	Web Search API to allow any enterprise agentic workflow to have access to the web index in real-time.
amazon	Product recommendations, spotting user trends early on, dynamic pricing, and more.

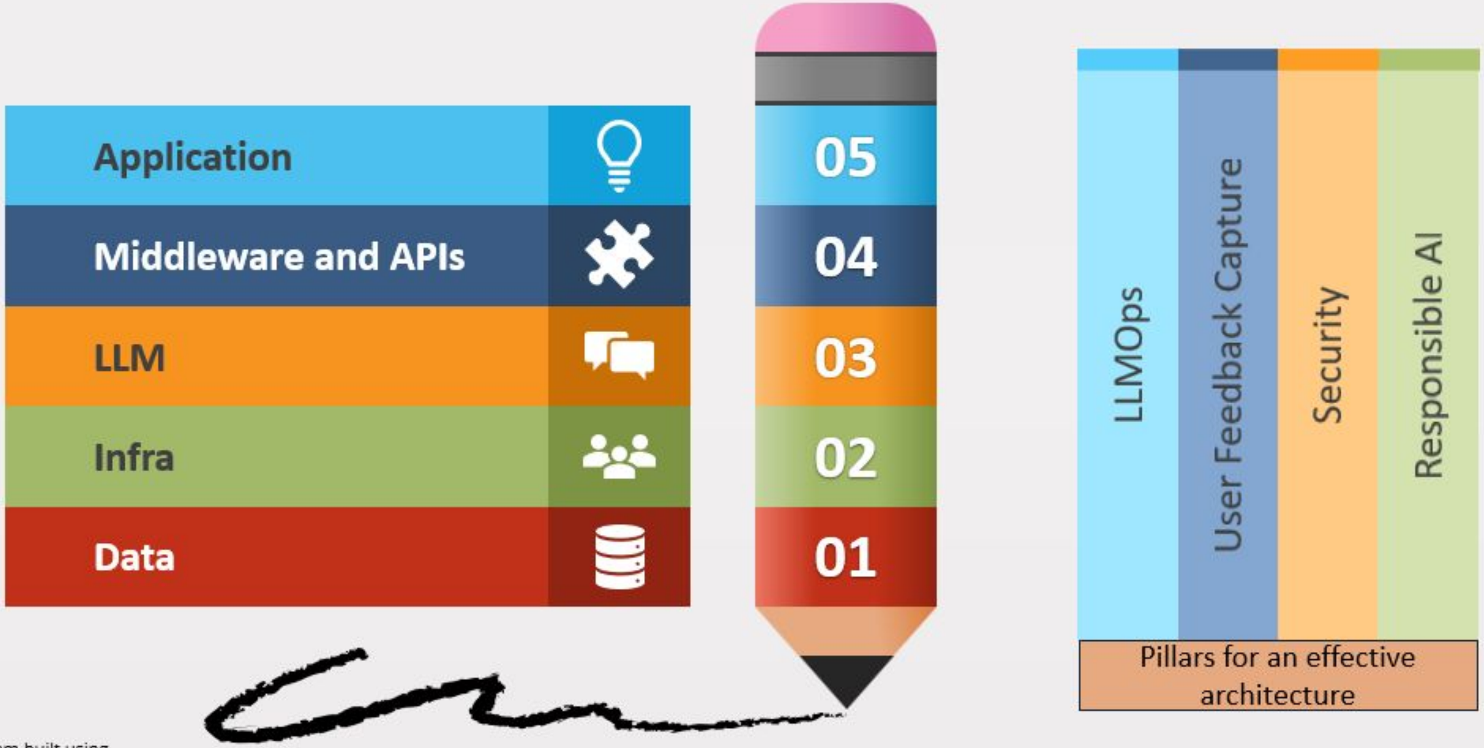
Getting Started with You.com



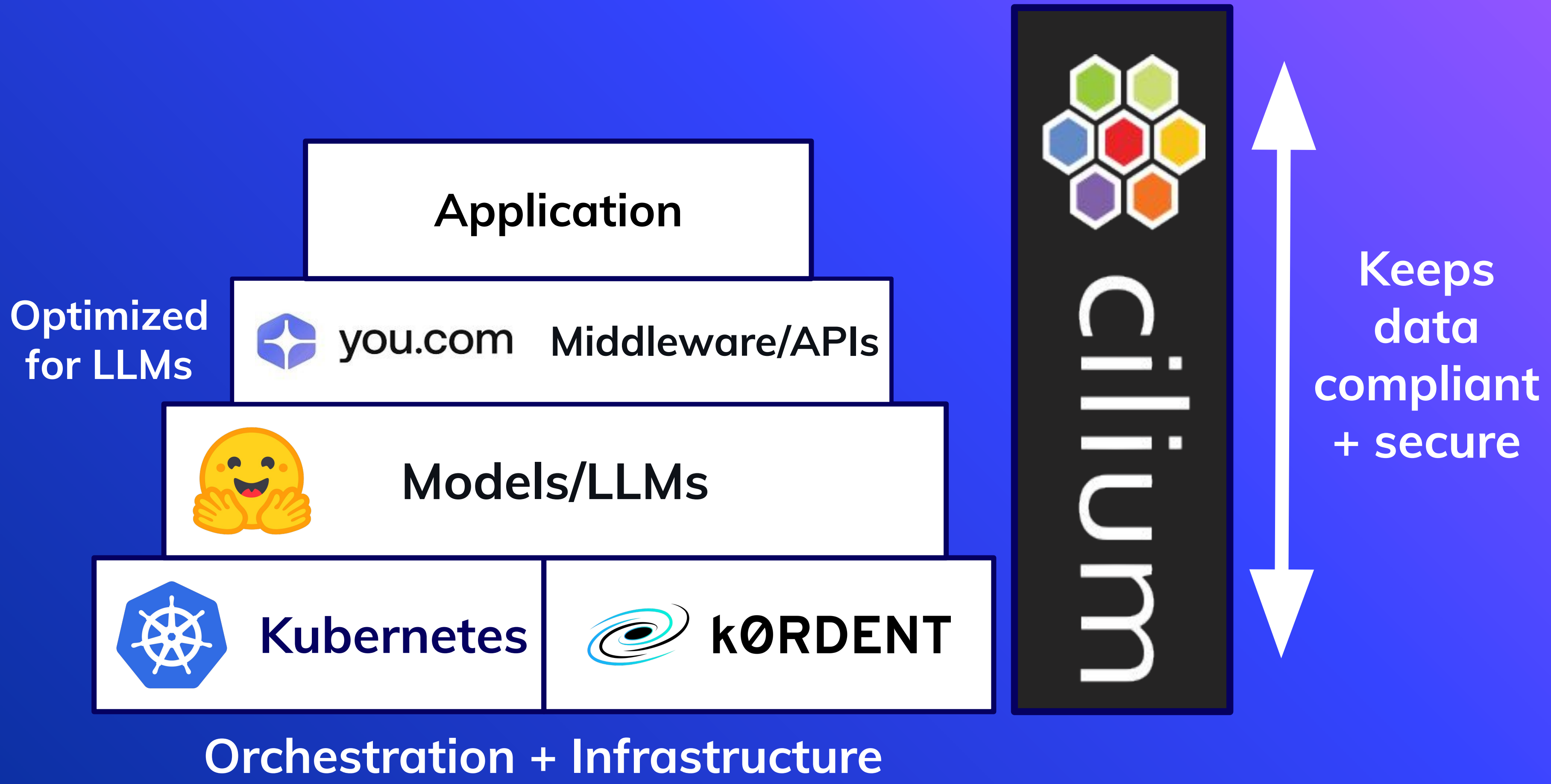
Review

By Anjelica & Christian

Architecture Layers of Generative AI system



High-Level Architecture Overview



Who Benefits? For Teams Scaling AI

Platform Engineers

AI Developers

Thank You!

You.com: Try it for Free!

AI/ML Free Lab

Scan me