EE 392b
Industrial AI
Spring Quarter 2024

Daniel C O’Neill

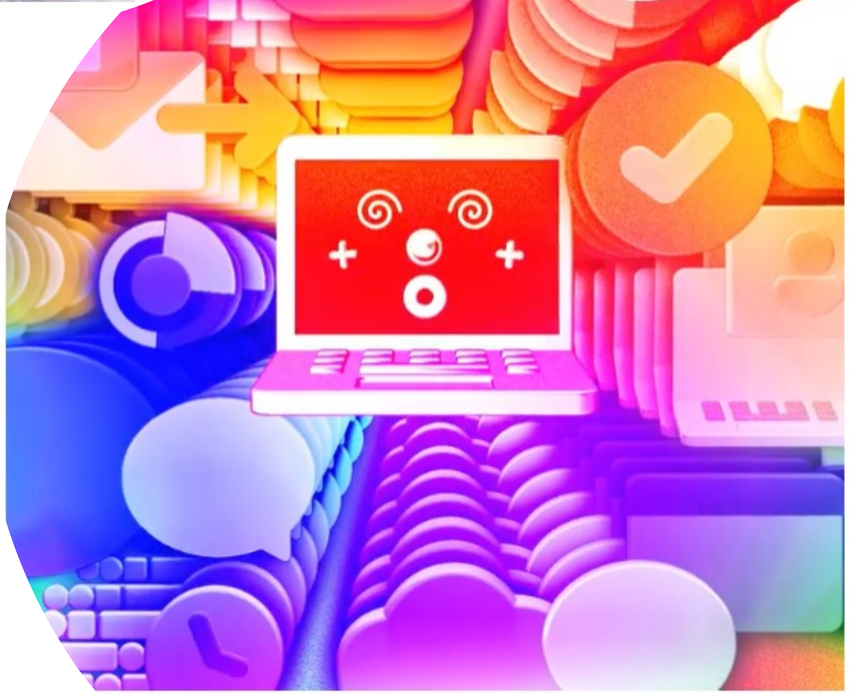
www.stanford.edu/~dconeill

https://web.stanford.edu/class/ee392b/

Agenda

- Class information
- I-AI Impact
- The I-AI Industry
  - A bit of tech
  - A bit about players
- Speakers
- Example applications

Economist



NYT

Instructors

- Daniel C O'Neill, Adjunct Professor in EE
  - Generative Models (AIOps, AI-genomics)
  - CEO & Founder @2 startups
    Partner AIOps@ Microsoft, Senior Director @TI, Senior Director SUN, VC General Partner
  - PhD EE Stanford (AFOSR, Google, Microsoft), MBA UC Berkeley
  - http://www.stanford.edu/~dconeill

- Dimitry Gorinevsky, Adjunct Professor in EE
  - Industrial AI and analytical applications in robotics, automotive, process control, energy, defense and aerospace
  - CEO of startup in AI for Supply Chain
  - IEEE Fellow
  - www.stanford.edu/~gorin

Class Mechanics

- Sequence of ten industry talks
  - Overview
  - Concepts
- Weekly on Tuesday’s
  - Some speakers on ZOOM
  - Check out class website at ee392b.stanford.edu
- 1 unit graded CR/NC
  - No formal pre-requisites
  - Attendance and questions
  - Term paper: one page report/summary
    - Will post formal requirements

Impact: I-AI is Everywhere

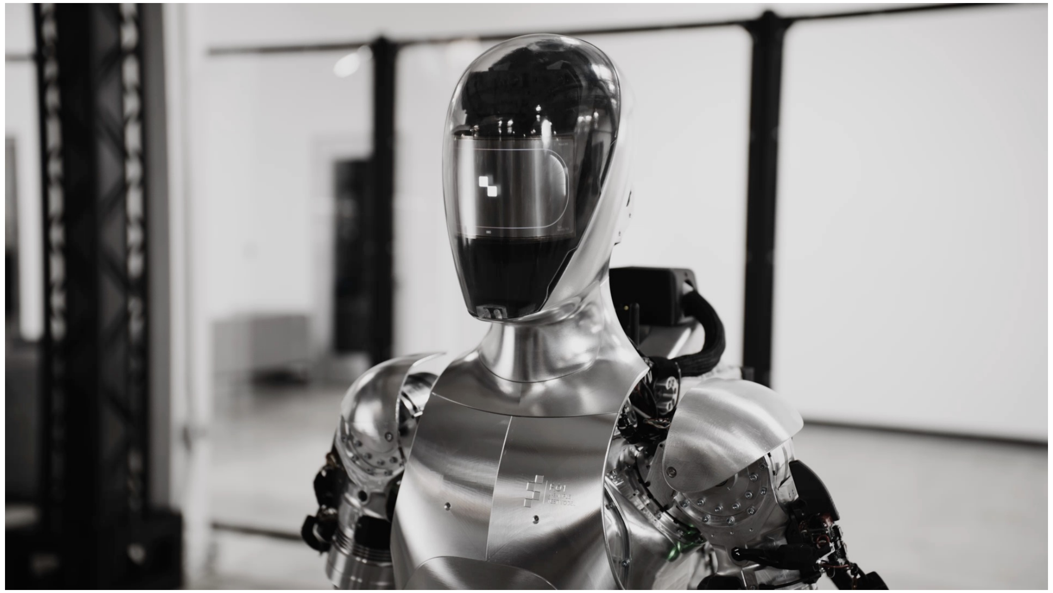
Video generation models as world simulators

OpenAI

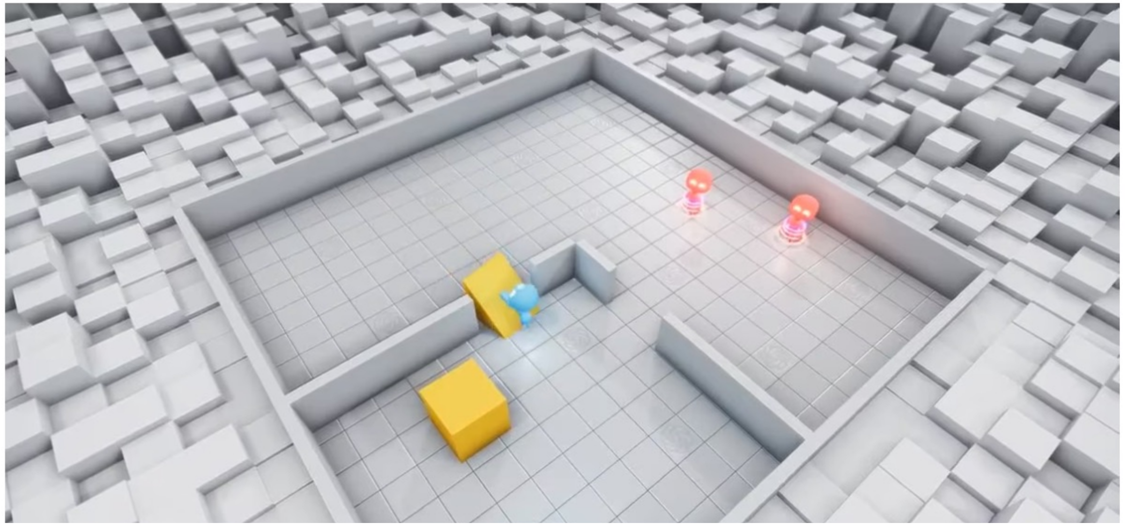
Microsoft and Siemens revolutionise industry with AI-powered Copilot

Change: Too Fast To Track

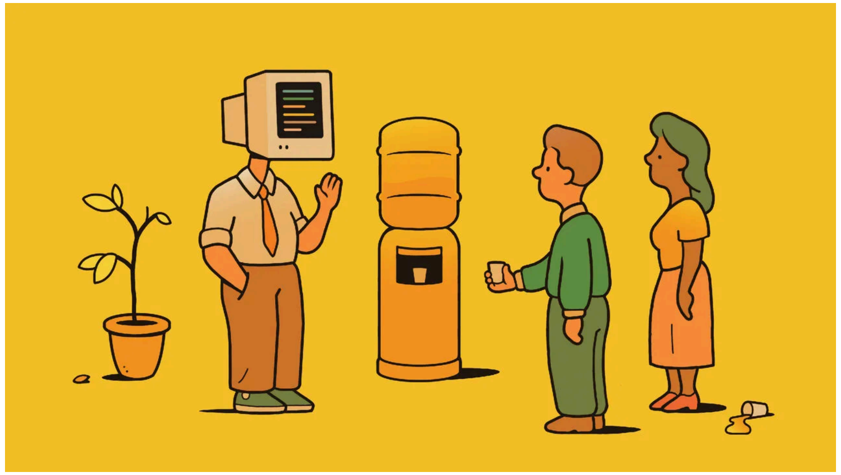
2024



2019



Context Is Everything:
Things To Know About The I-AI Industry

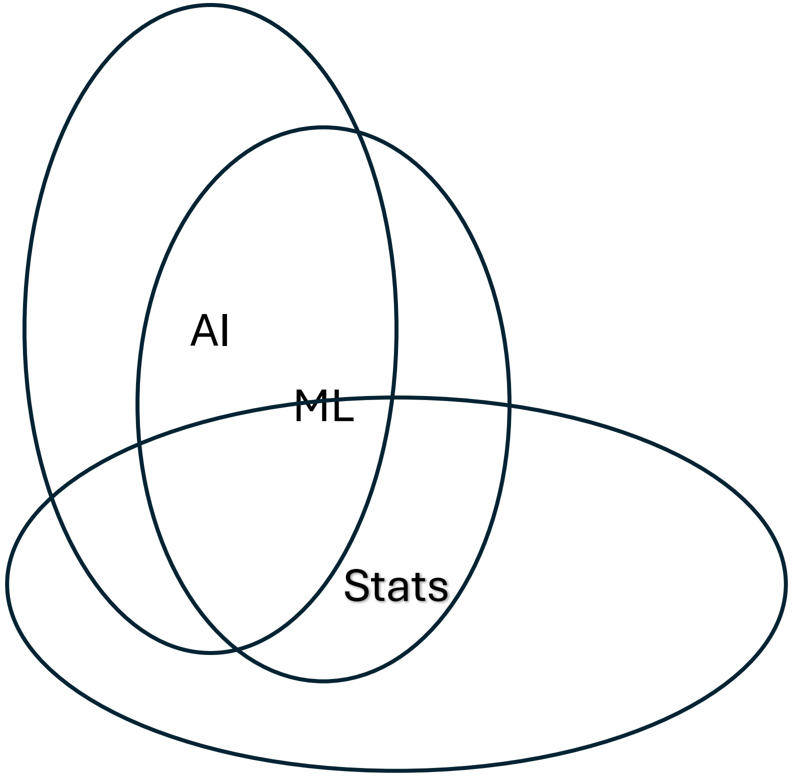


The Economist

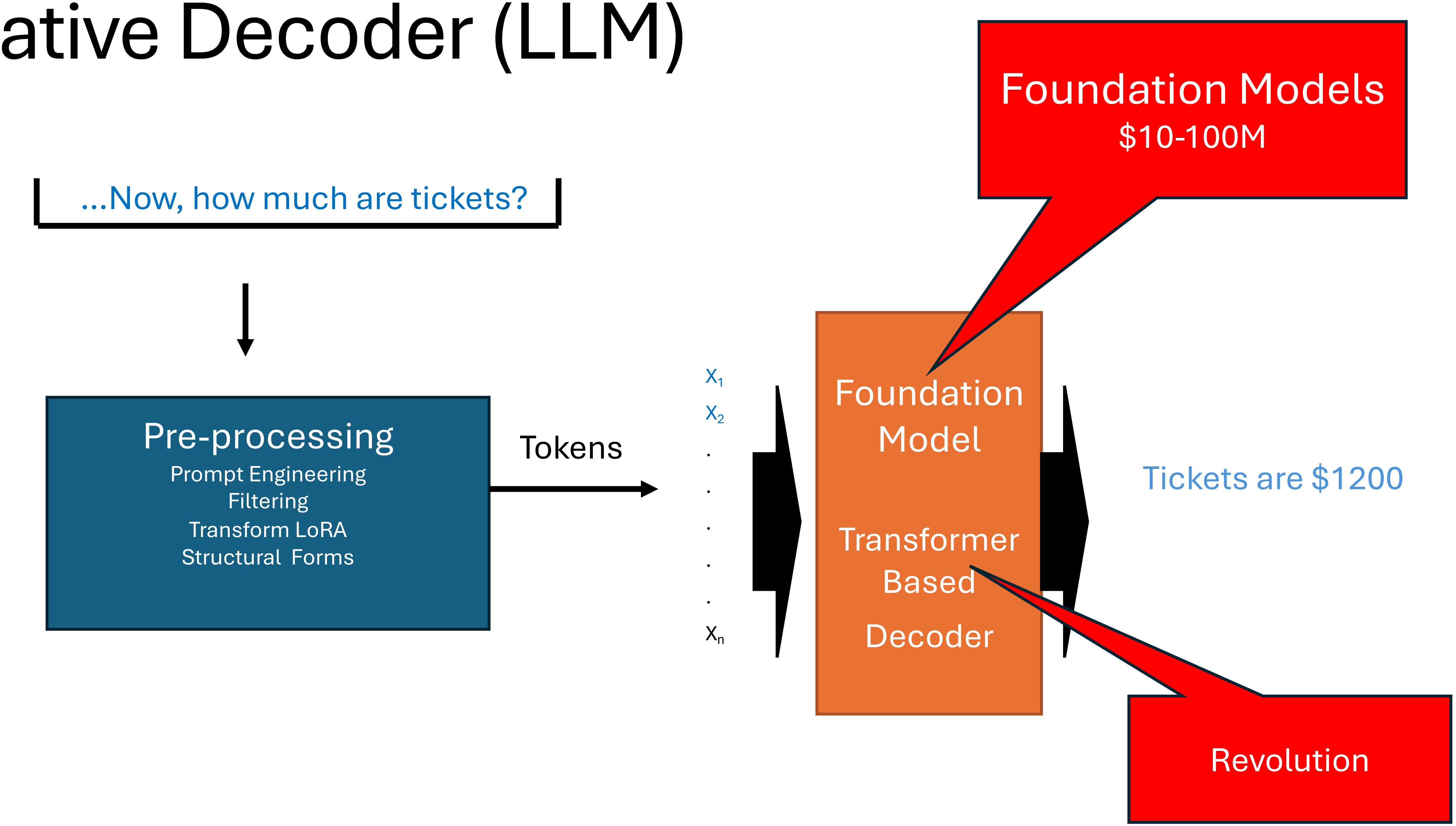
- What makes I-AI different?
    - Downside risk
        - Physical loss
        - Business impact
        - Legal problems
    - Often Mission Critical
    - Cost is an issue
- A bit of background
    - AI tech
    - I-AI industry

AI Basics

- **ML and AI overlap, a lot**
  - High dimensional data
  - Large parameter spaces (up to 10^12)
  - Nonlinear
  - Supervised or Unsupervised
- **In common everyday usage**
  - **AI =>** ***Generative Decoder (GPT)***
    - Generative model is a model of a distribution $p(x_{i+1} \mid x_1,\ldots x_i)$. P(cat | sequence of images)
    - Decoder ”reads L to R”, autoregressive
- **Industrial applications require a bit more**
  - Encoders, VAE, RL
  - Physical models
  - Models of ‘reality’

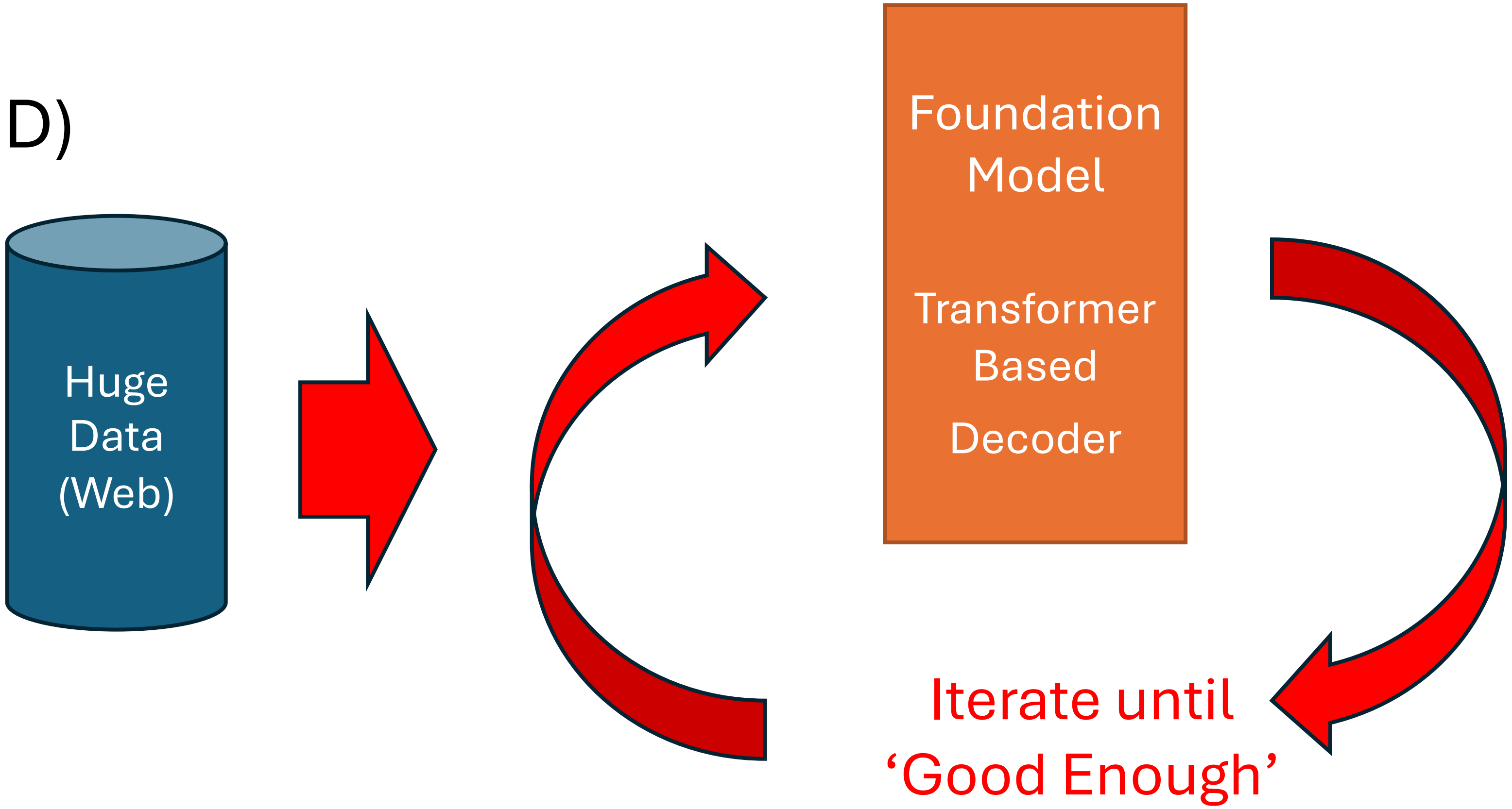


The Common Everyday Experience: Generative Decoder (LLM)



Backup: Training Foundation Models

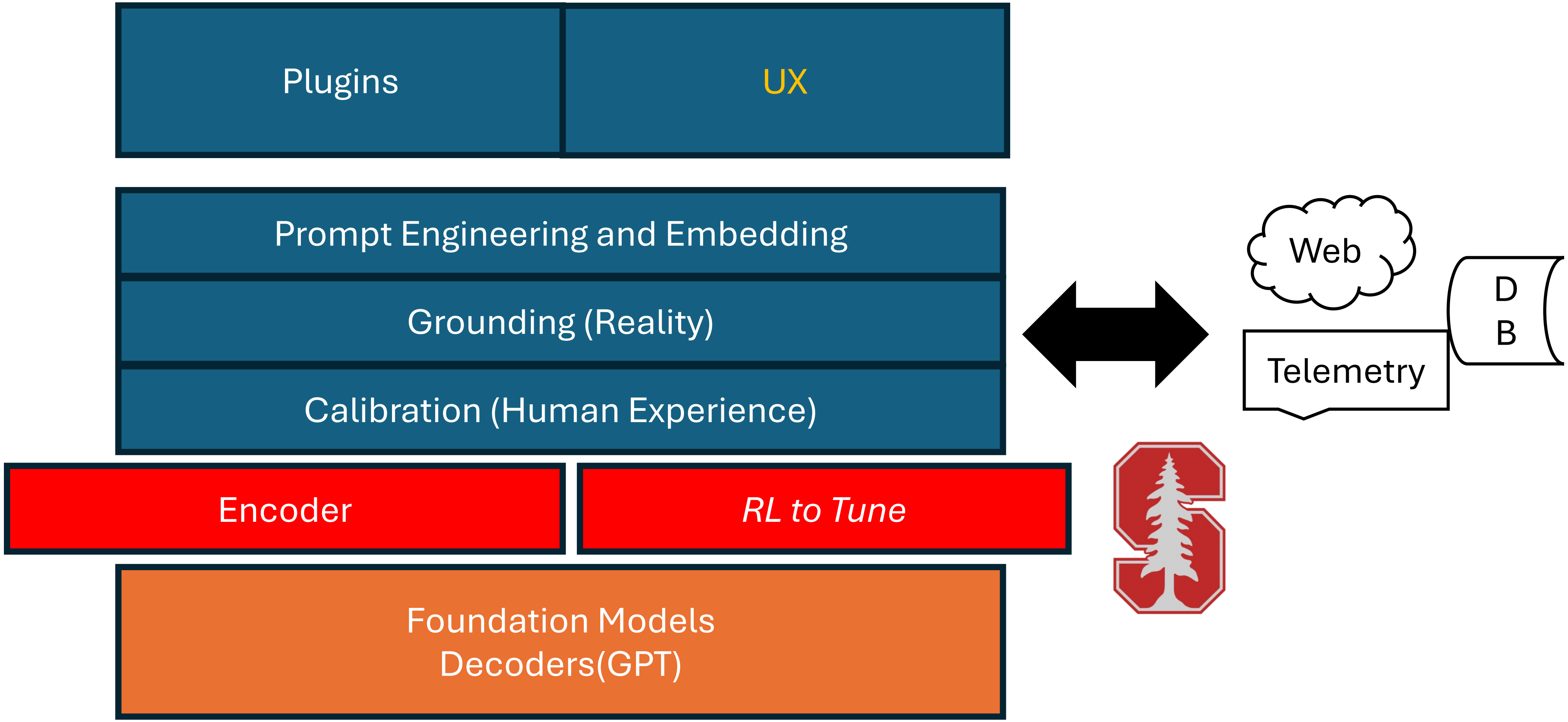
- Huge Data 100-1000’s GB
- Cost $10-100M
- Massive iteration (SGD)



But, I-AI More Complex

Multi-modal

- **Telemetry**
- Text
- Images
- Sounds



Plugins

UX

Prompt Engineering and Embedding

Grounding (Reality)

Calibration (Human Experience)

Encoder

RL to Tune

Foundation Models
Decoders(GPT)

Web

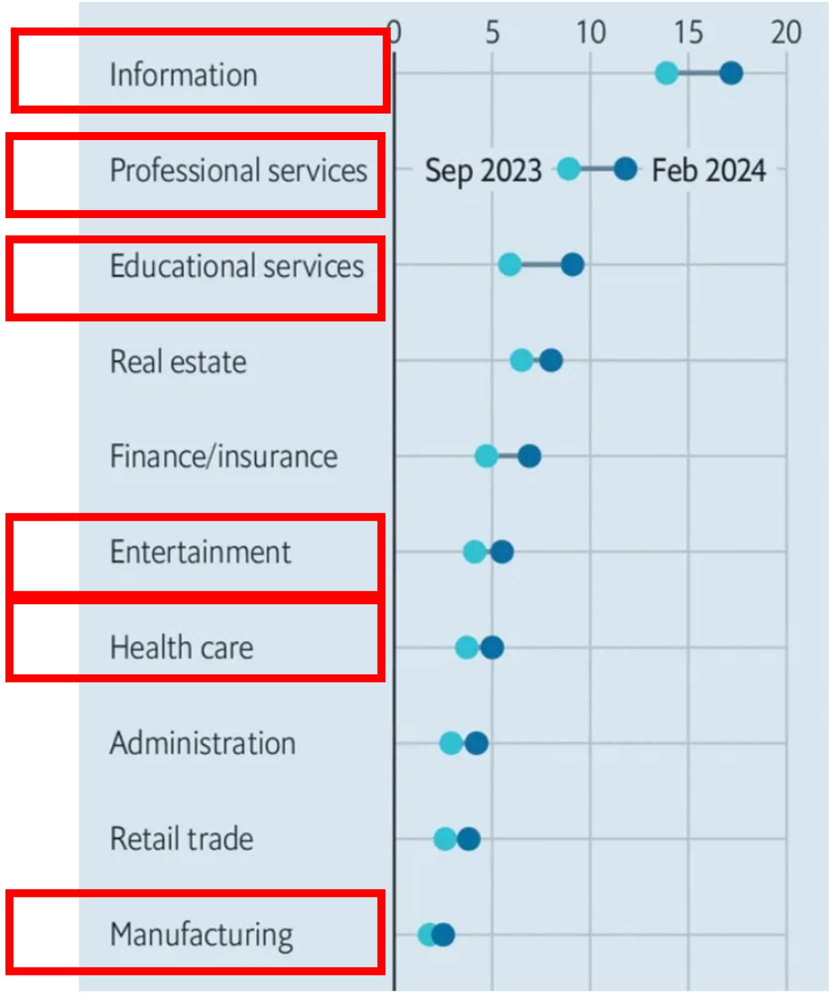
D
B

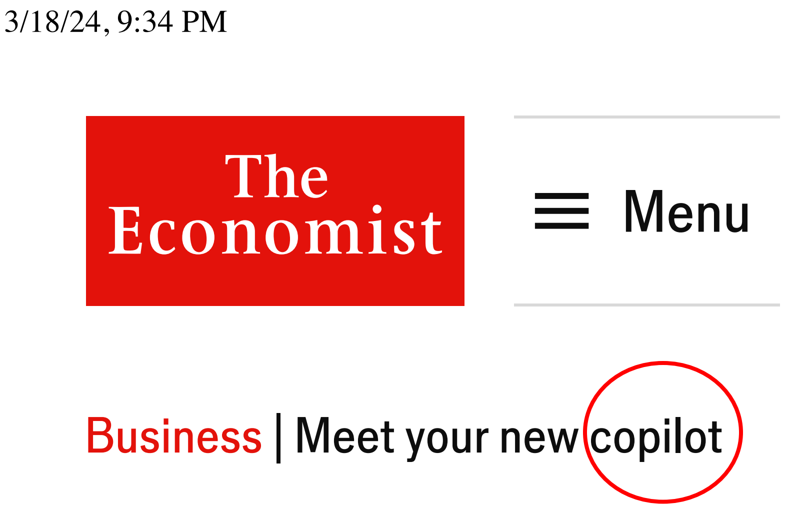
Telemetry

I-AI Technical Challenges

- Managing downsides
  - Mission critical
  - Often Real Time
  - Few to no errors
- AI safety
  - Hallucinations…
  - False Alarms…
- IP
  - What data can I use
  - Right to forget
- Cost

Definition: I-AI Industries



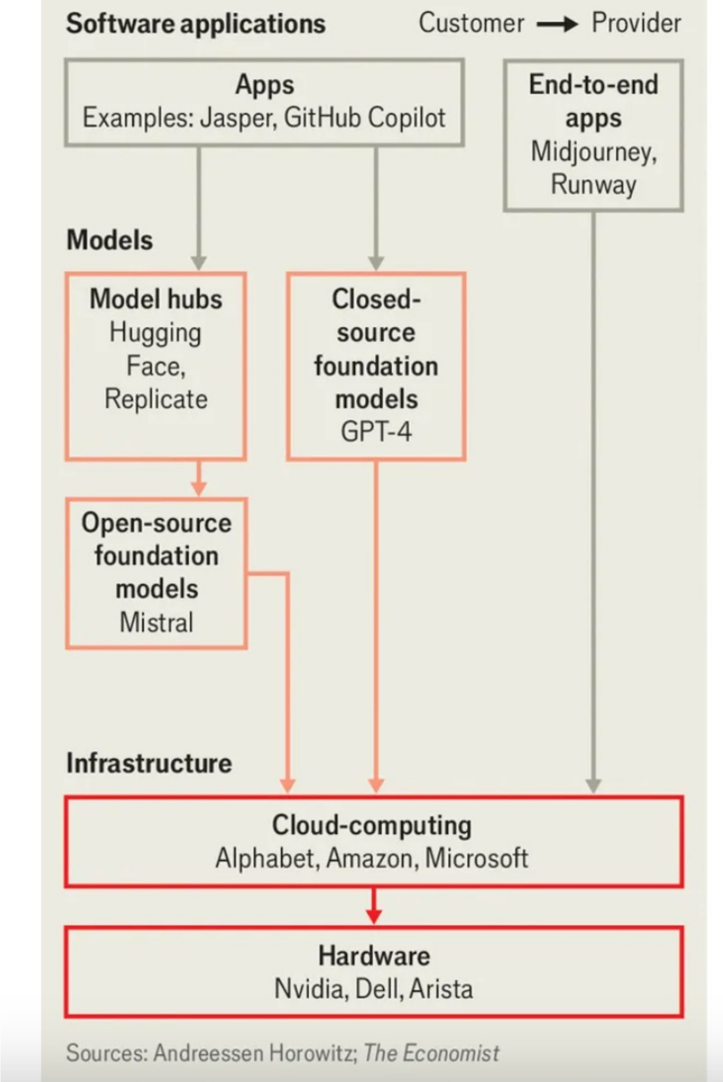

3/18/24, 9:34 PM

Business | Meet your new copilot

Convergence

- Professional Services
- Entertainment
- Education

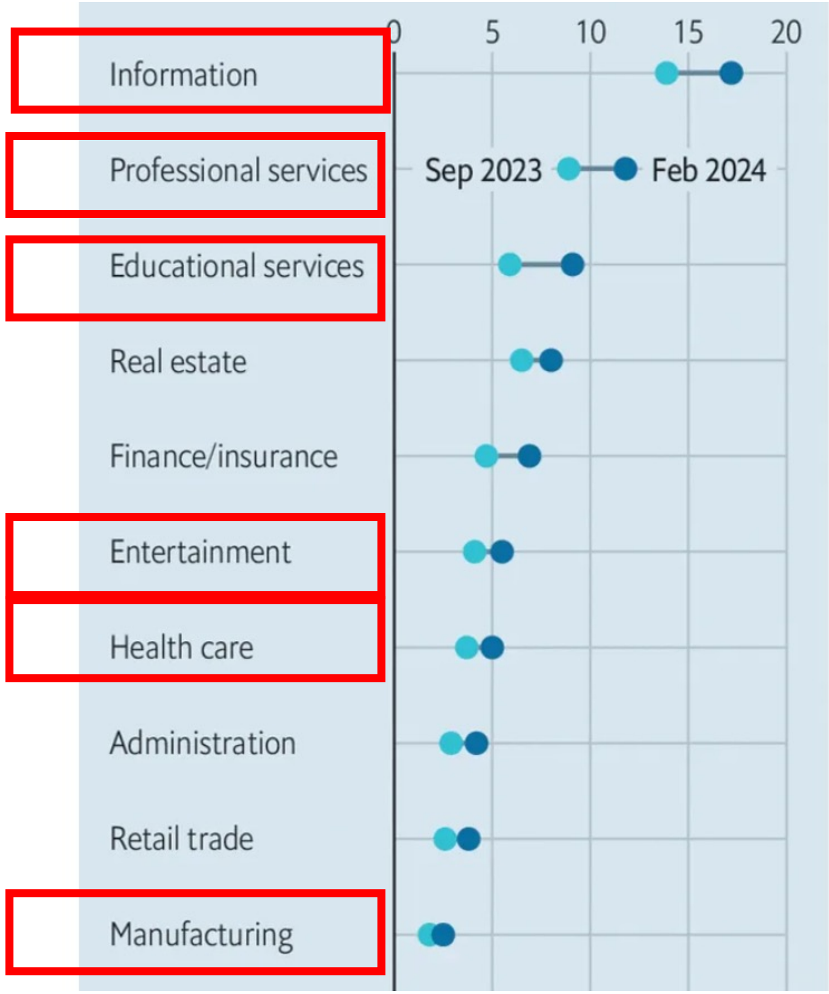
Where Do They Get The Pieces?

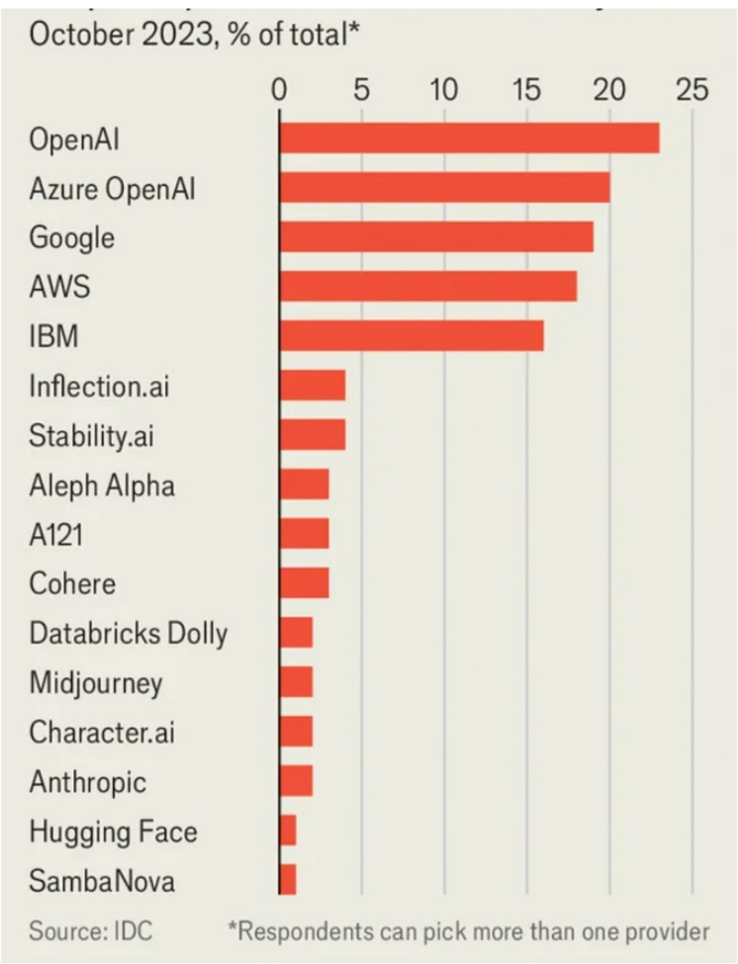


Sources: Andreessen Horowitz; *The Economist*

- Open Source from Hugging Face, Github, etc.
- Foundation models and One-Stop-Shops
  - OpenAI / Microsoft - GPT
  - Google - Gemini
  - Meta - Llama
- Then train or tune to application
- Partnerships

Industries and AI Tech Suppliers




October 2023, % of total*

Source: IDC

*Respondents can pick more than one provider

Structure of I-AI Industry

	Existing Applications	New Applications
Mature (e.g. MSFT, Siemens)	Energy Oil and Gas Drug Discovery	AIOps Service Avatar Defense
Startup (e.g. Pay-i, Inworld)	Medicine	Genomics Cost Professional Services

AI Startups

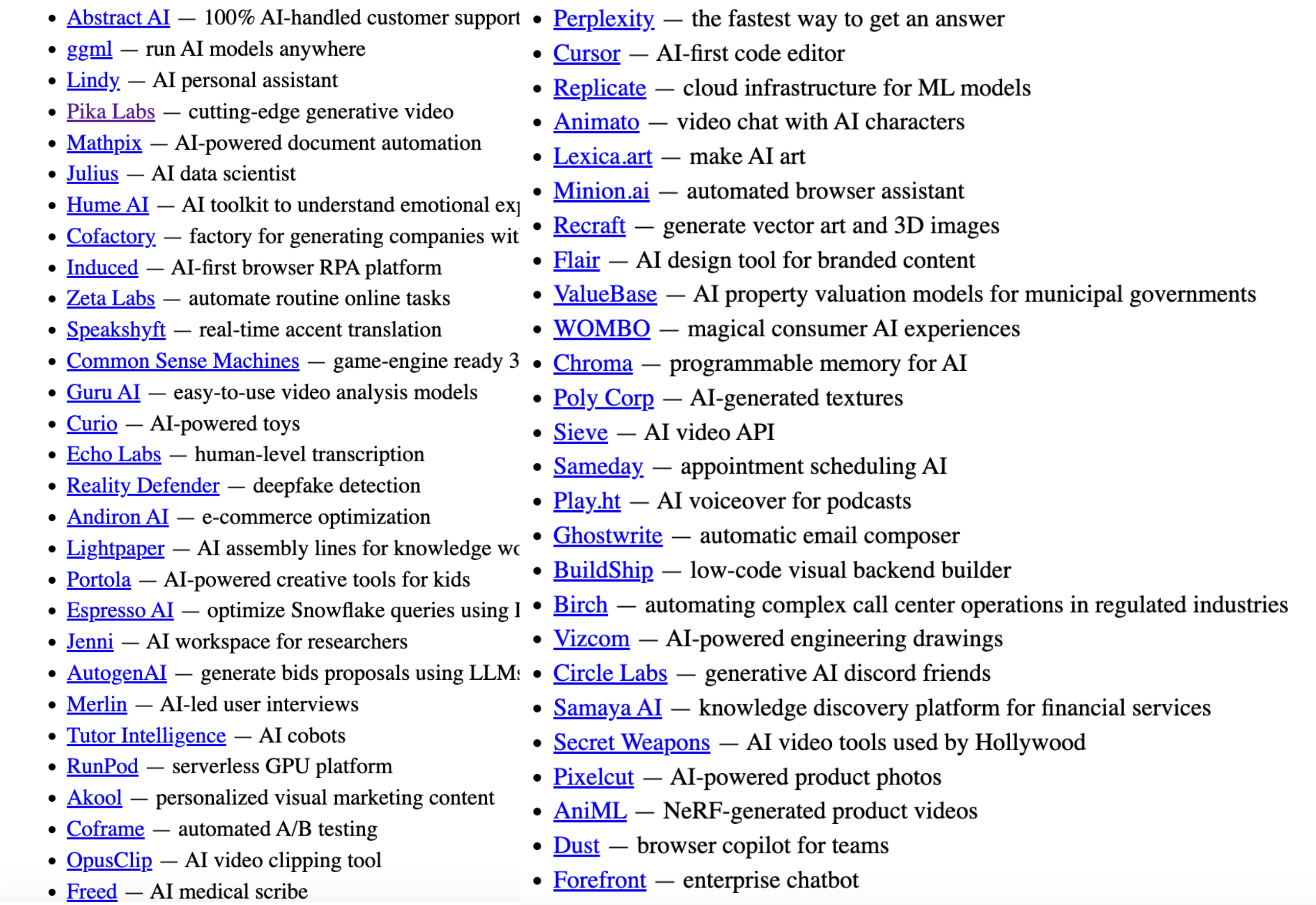
Y

Startx

AI Grant

Hundreds of AI Startups

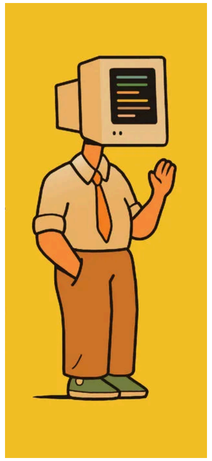
Pay-I and Inworld



- Abstract AI — 100% AI-handled customer support
- ggml — run AI models anywhere
- Lindy — AI personal assistant
- Pika Labs — cutting-edge generative video
- Mathpix — AI-powered document automation
- Julius — AI data scientist
- Hume AI — AI toolkit to understand emotional ex
- Cofactory — factory for generating companies wit
- Induced — AI-first browser RPA platform
- Zeta Labs — automate routine online tasks
- Speakshyft — real-time accent translation
- Common Sense Machines — game-engine ready 3
- Guru AI — easy-to-use video analysis models
- Curio — AI-powered toys
- Echo Labs — human-level transcription
- Reality Defender — deepfake detection
- Andiron AI — e-commerce optimization
- Lightpaper — AI assembly lines for knowledge wo
- Portola — AI-powered creative tools for kids
- Espresso AI — optimize Snowflake queries using l
- Jenni — AI workspace for researchers
- AutogenAI — generate bids proposals using LLMs
- Merlin — AI-led user interviews
- Tutor Intelligence — AI cobots
- RunPod — serverless GPU platform
- Akool — personalized visual marketing content
- Coframe — automated A/B testing
- OpusClip — AI video clipping tool
- Freed — AI medical scribe
- Perplexity — the fastest way to get an answer
- Cursor — AI-first code editor
- Replicate — cloud infrastructure for ML models
- Animato — video chat with AI characters
- Lexica.art — make AI art
- Minion.ai — automated browser assistant
- Recraft — generate vector art and 3D images
- Flair — AI design tool for branded content
- ValueBase — AI property valuation models for municipal governments
- WOMBO — magical consumer AI experiences
- Chroma — programmable memory for AI
- Poly Corp — AI-generated textures
- Sieve — AI video API
- Sameday — appointment scheduling AI
- Play.ht — AI voiceover for podcasts
- Ghostwrite — automatic email composer
- BuildShip — low-code visual backend builder
- Birch — automating complex call center operations in regulated industries
- Vizcom — AI-powered engineering drawings
- Circle Labs — generative AI discord friends
- Samaya AI — knowledge discovery platform for financial services
- Secret Weapons — AI video tools used by Hollywood
- Pixelcut — AI-powered product photos
- AniML — NeRF-generated product videos
- Dust — browser copilot for teams
- Forefront — enterprise chatbot

Disrupting Mature Companies

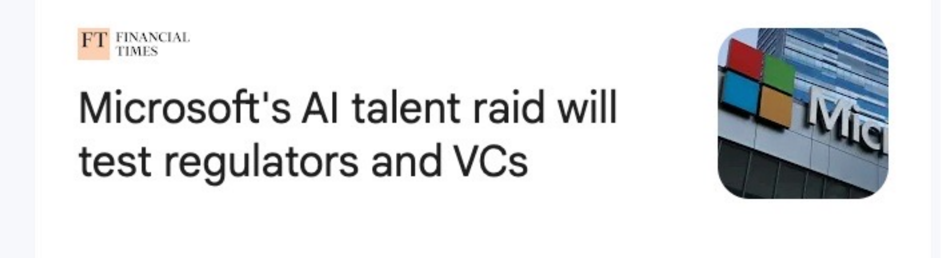
Copilots



NERI

Organizing in an I-AI world – Bain McKinsey

- New ways to address existing applications
- Employees need new skills
- You need to reorganize
- Partnerships (e.g. Microsoft/Siemens)
- New Suppliers (Azure)



FT FINANCIAL TIMES

Microsoft's AI talent raid will test regulators and VCs

I-AI Industry Challenges:

- Convergence – Application boundaries change
- Mission Critical- Who owns what?
- Partnerships and consolidation
  - Cloud
  - Foundation models

Speakers!

Name	Company	Title	Topic
Dan ONeill	Stanford	Professor	Overview
Timothy Chou	Self	Board Member	Medical AI
Gerhard Kress	Siemens	SVP Digital Business	AI for Industry Digitization
Scott Penberthy	Google	Man. Director	AI Genomics
Sarah Elk	Bain & Company	Partner	AI Organization
Lapo Mori	McKinsey	Partner	AI for Process Industry
David Tepper	Pay-i	CEO and Founder	AI Cost Management
Kylan Gibbs	Inworld	CEO and Founder	AI as Interface
Thomas Higginbotham	C3.ai	Sr. Director	AI for Aerospace and Defense
Mathew John	Microsoft	Sr. Director	AIOps

Example Applications

- AIOps
- Genomics
- Human Interface
- Cost
- Process Industry

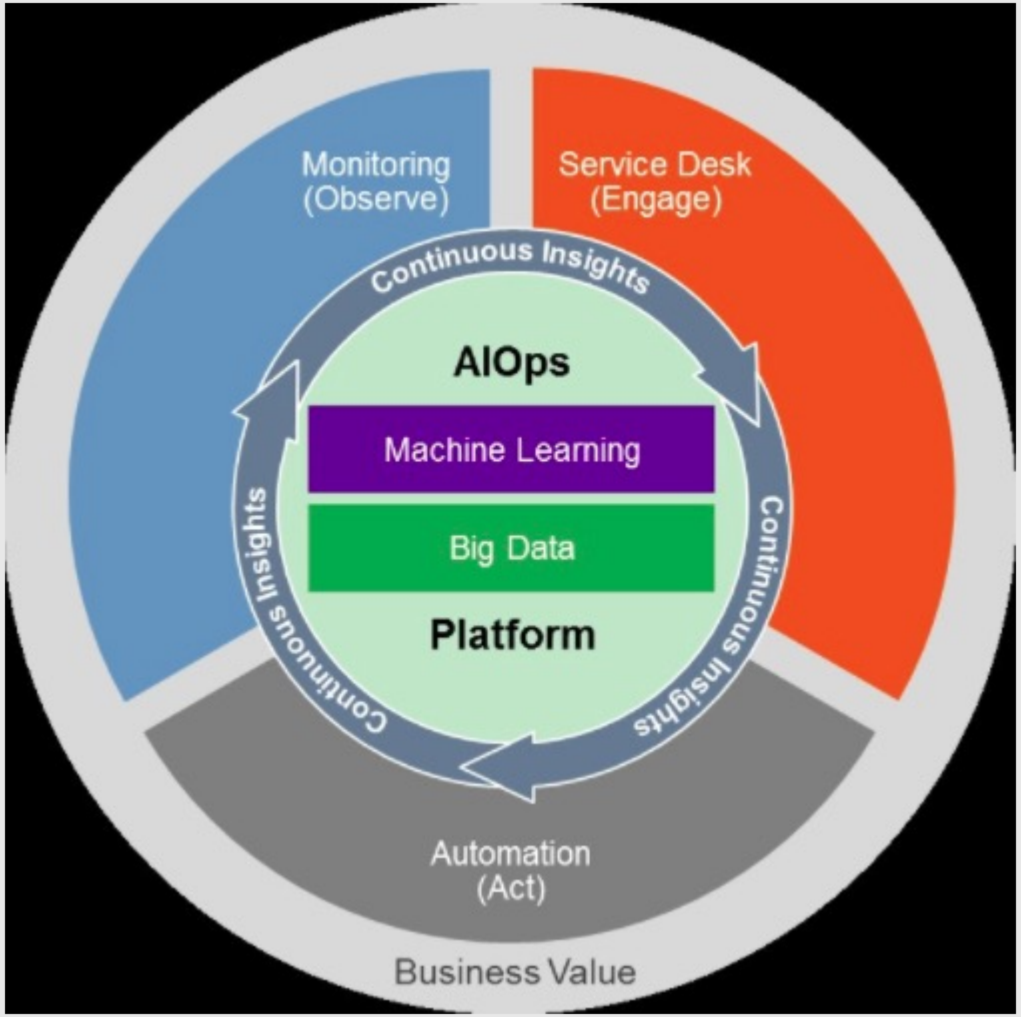
AIOps – AI for Datacenter Operations

Big data and ML driven IT operation automation process

Adoption has increased with the uptick of digital transformation

Business value

- Higher efficiency
- Higher Service quality
- Lower COGS



Source: Gartner

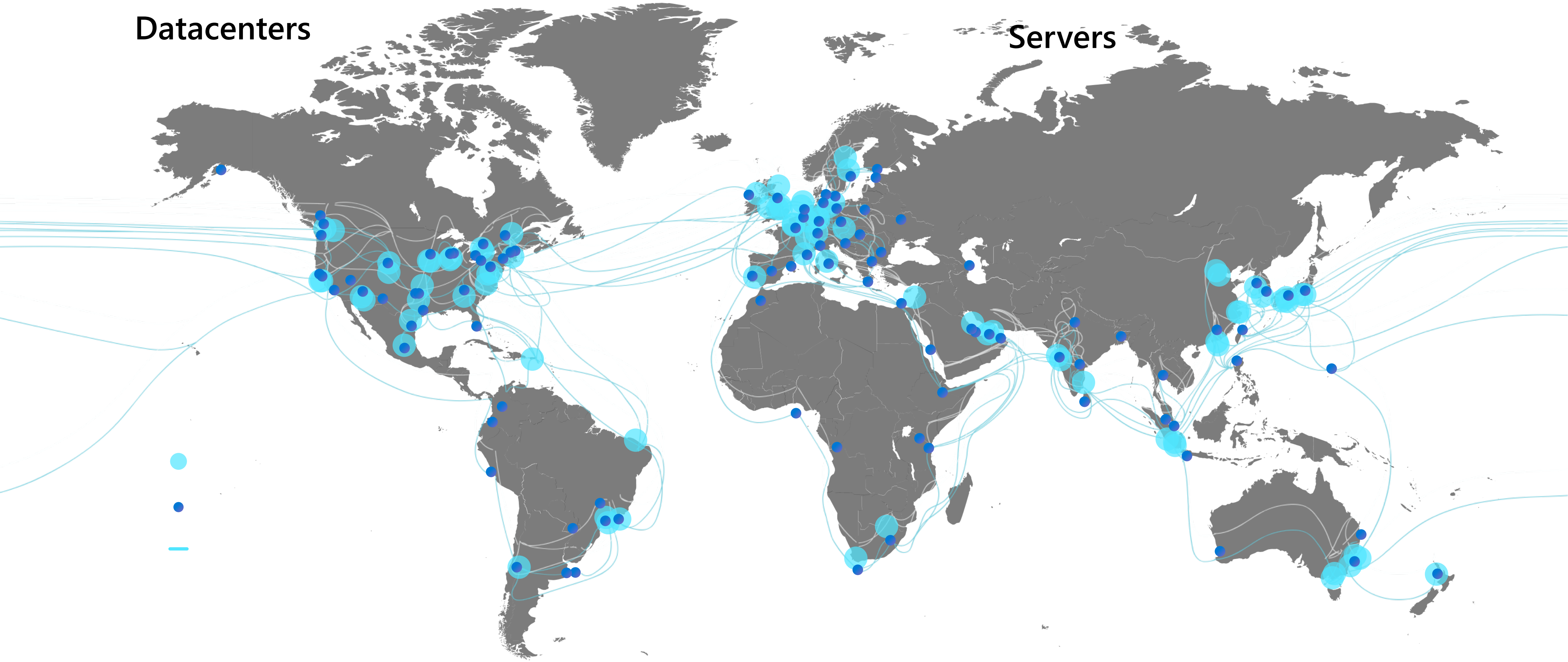
Azure

200+

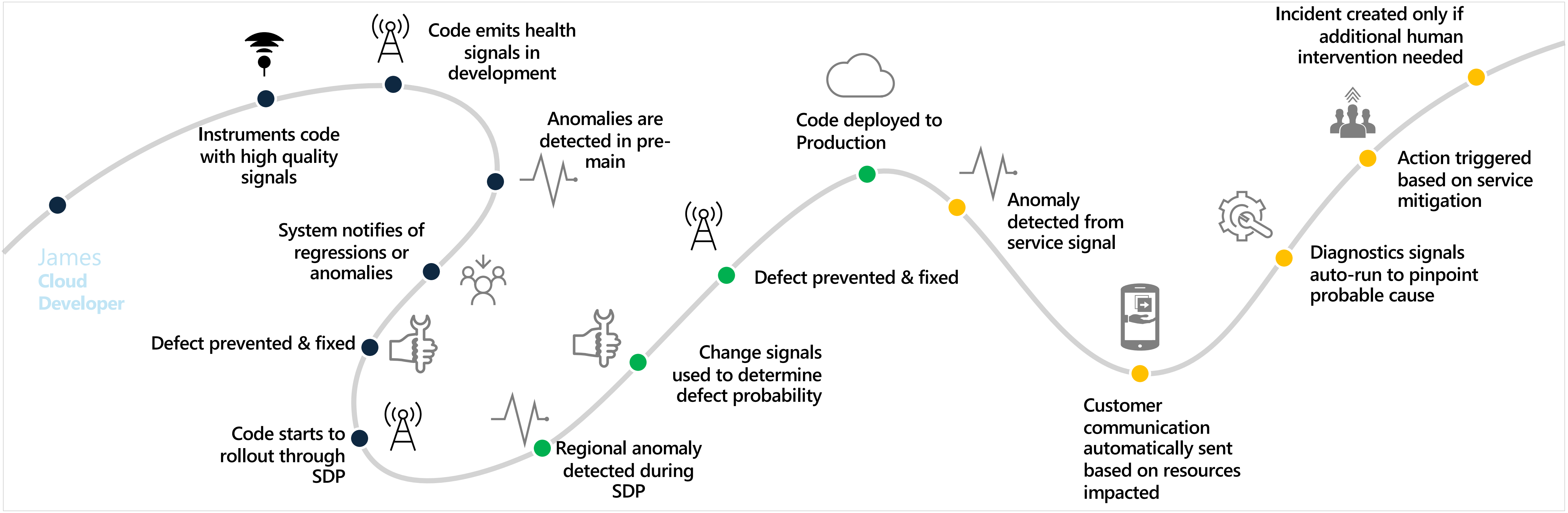
Datacenters

4M+

Servers



Why AIOps? Complexity and Downside



James
Cloud
Developer

Instruments code with high quality signals

Code emits health signals in development

Anomalies are detected in pre-main

System notifies of regressions or anomalies

Defect prevented & fixed

Code starts to rollout through SDP

Regional anomaly detected during SDP

Change signals used to determine defect probability

Defect prevented & fixed

Code deployed to Production

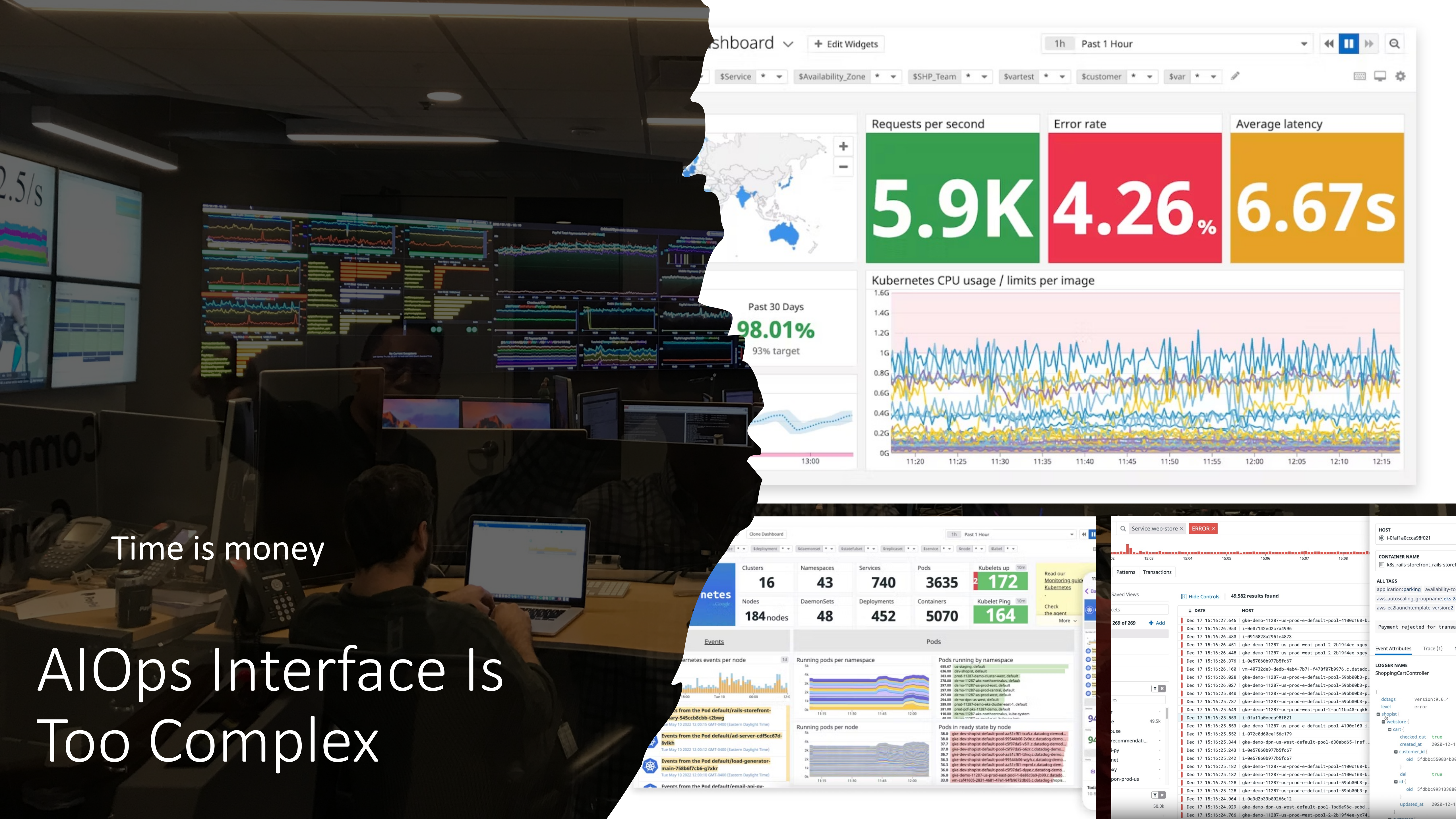
Anomaly detected from service signal

Customer communication automatically sent based on resources impacted

Diagnostics signals auto-run to pinpoint probable cause

Action triggered based on service mitigation

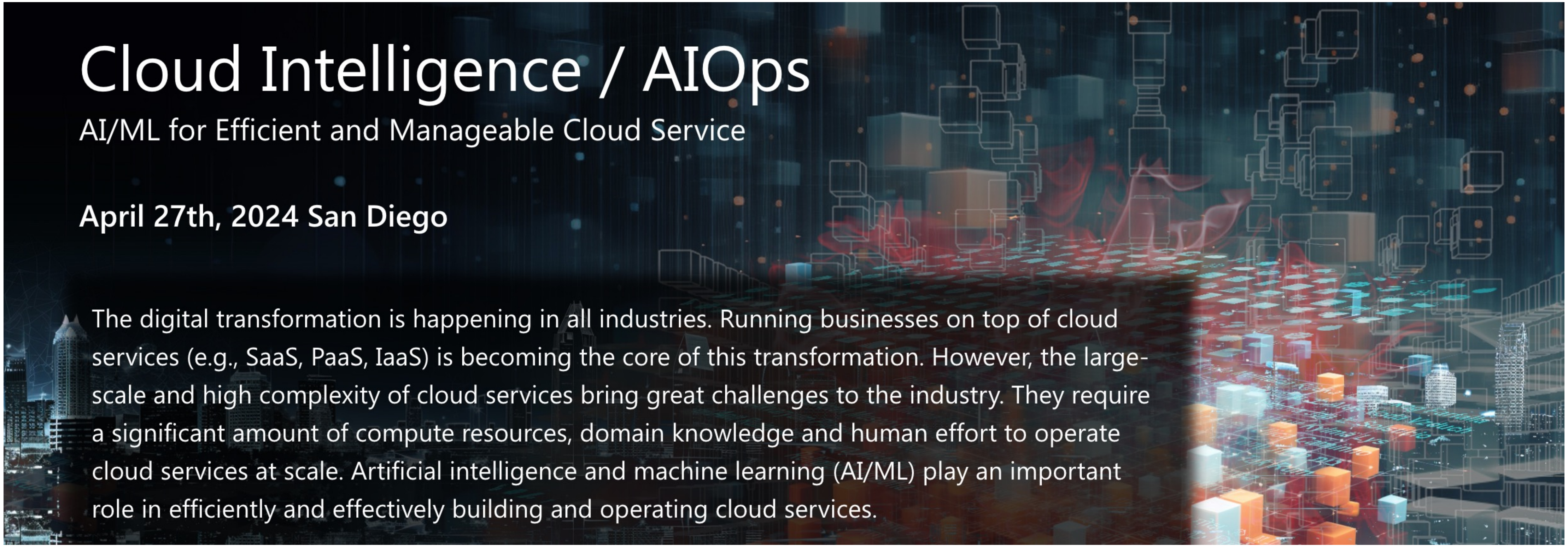
Incident created only if additional human intervention needed



Time is money

AIOps Interface Is Too Complex

Backup



Cloud Intelligence / AIOps

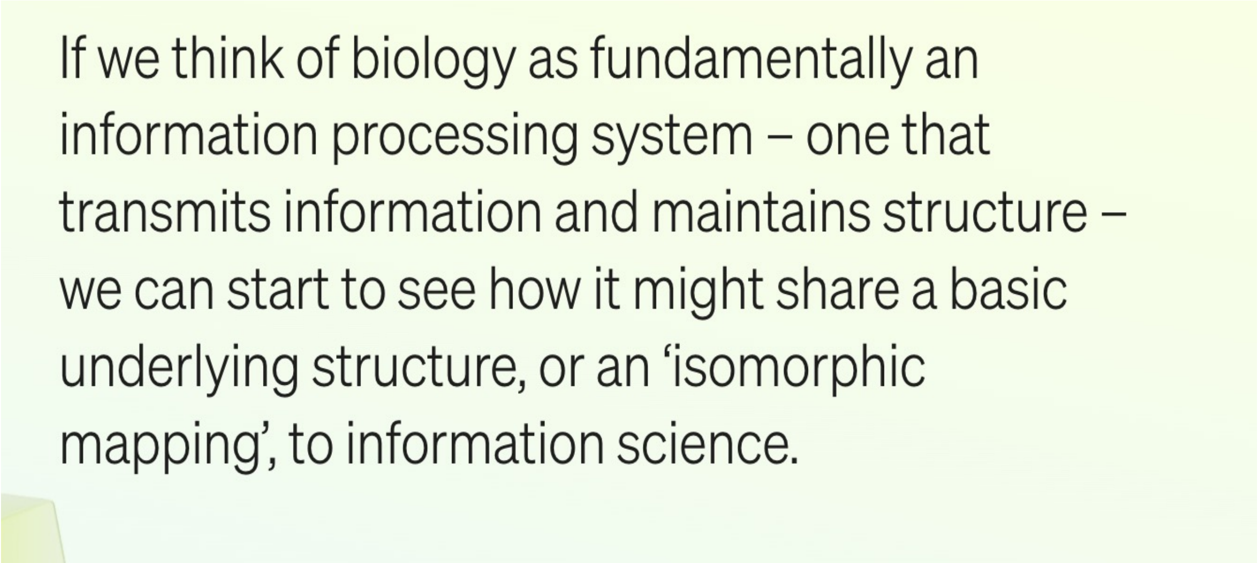
AI/ML for Efficient and Manageable Cloud Service

April 27th, 2024 San Diego

The digital transformation is happening in all industries. Running businesses on top of cloud services (e.g., SaaS, PaaS, IaaS) is becoming the core of this transformation. However, the large-scale and high complexity of cloud services bring great challenges to the industry. They require a significant amount of compute resources, domain knowledge and human effort to operate cloud services at scale. Artificial intelligence and machine learning (AI/ML) play an important role in efficiently and effectively building and operating cloud services.

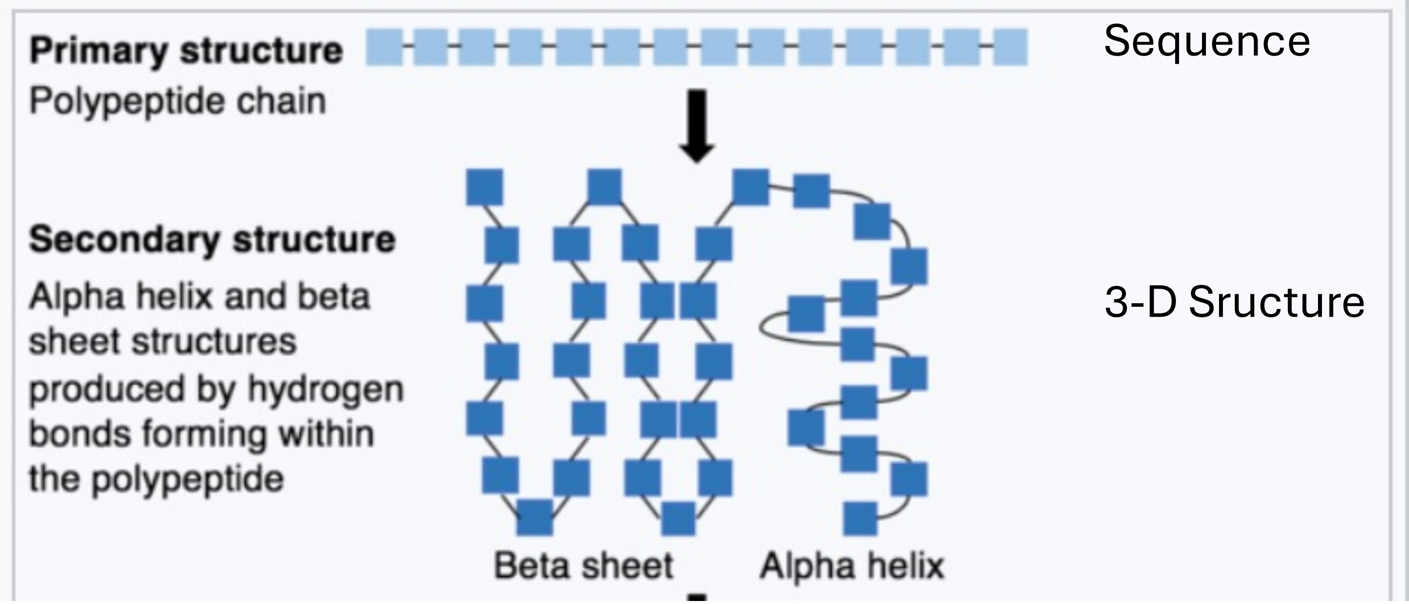
AI-Genomics

- Testing
    - Blood tests for cancer
        - Freenome
        - GRAIL
    - Blood tests for fetal genetic issues
    - Blood tests for epigenetic problems
- Drug discovery
    - Isomorphic Labs


If we think of biology as fundamentally an information processing system – one that transmits information and maintains structure – we can start to see how it might share a basic underlying structure, or an ‘isomorphic mapping’, to information science.

AlphaFold 2

AlphaFold is an AI system developed by DeepMind that predicts a protein’s 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.


Primary structure
Polypeptide chain
Secondary structure
Alpha helix and beta sheet structures produced by hydrogen bonds forming within the polypeptide
Beta sheet
Alpha helix

Sequence

3-D Sructure

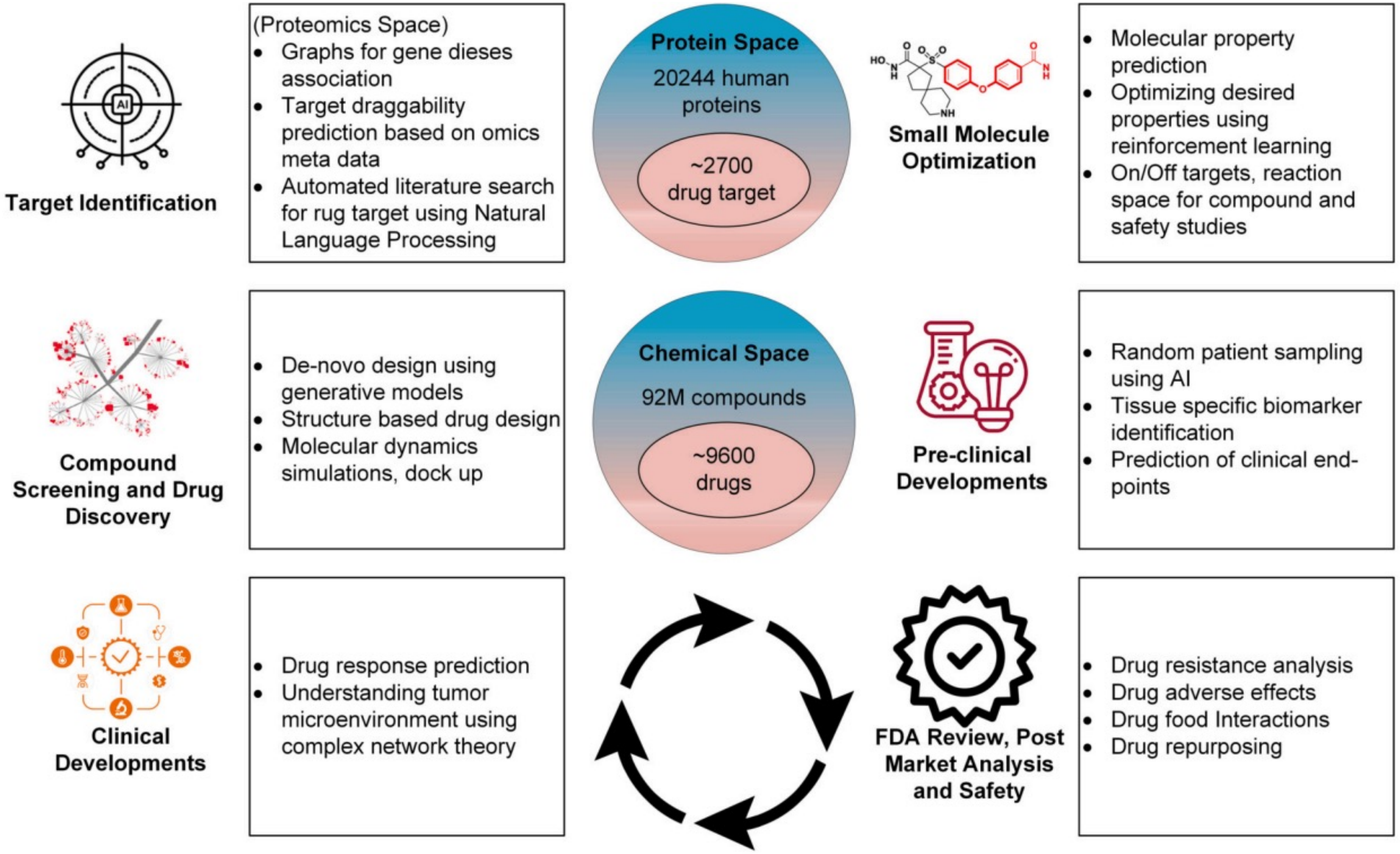
Combinatorics!

I-AI for Drug Discovery Is Complex

Proteins

Molecules

Phenotypical



Target Identification

(Proteomics Space)
- Graphs for gene dieses association
- Target draggability prediction based on omics meta data
- Automated literature search for rug target using Natural Language Processing

Protein Space

20244 human proteins

~2700 drug target

Small Molecule Optimization

- Molecular property prediction
- Optimizing desired properties using reinforcement learning
- On/Off targets, reaction space for compound and safety studies

Compound Screening and Drug Discovery

- De-novo design using generative models
- Structure based drug design
- Molecular dynamics simulations, dock up

Chemical Space

92M compounds

~9600 drugs

Pre-clinical Developments

- Random patient sampling using AI
- Tissue specific biomarker identification
- Prediction of clinical end-points

Clinical Developments

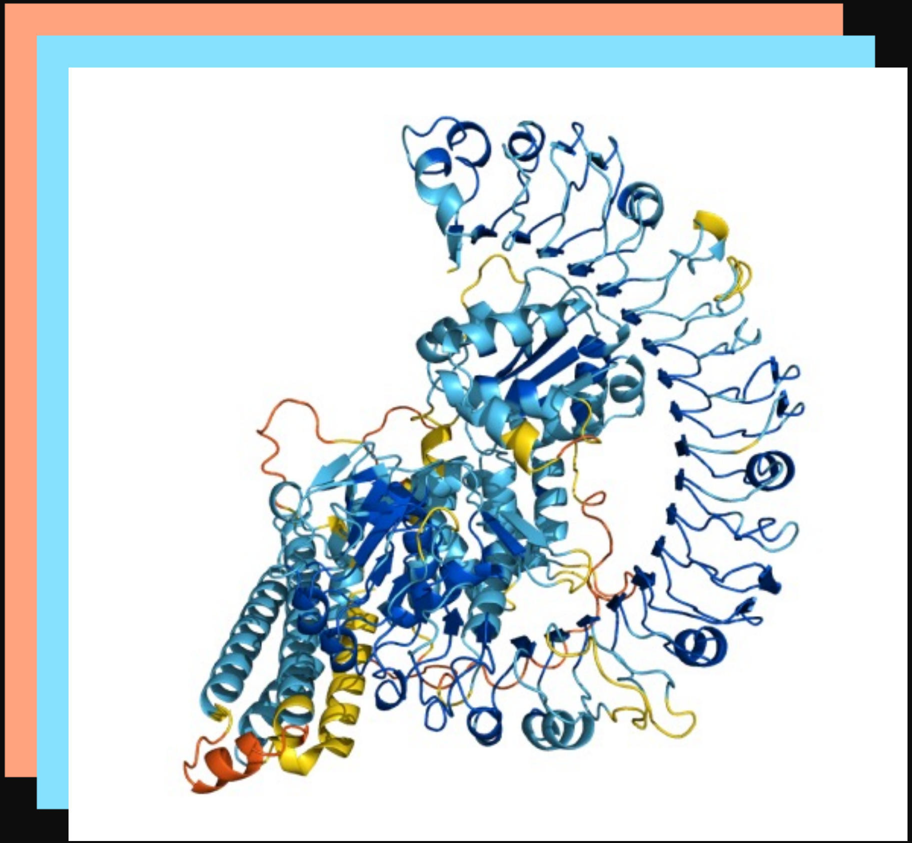
- Drug response prediction
- Understanding tumor microenvironment using complex network theory

FDA Review, Post Market Analysis and Safety

- Drug resistance analysis
- Drug adverse effects
- Drug food Interactions
- Drug repurposing

The Human Interface Is Complex

Drug
discovery:
Locks and keys



Note: For energy reasons there can exist several different native configurations

I-AI Interface Problem

- How will humans be able to effectively interact with AI systems?
    - AIOps
    - Genomics
    - Robotics
- Converged systems
    - Education & training
    - Professional Services
    - Entertainment
        - Actors strike

Ideal:

You declare what is desired and have the system take actions to create it!

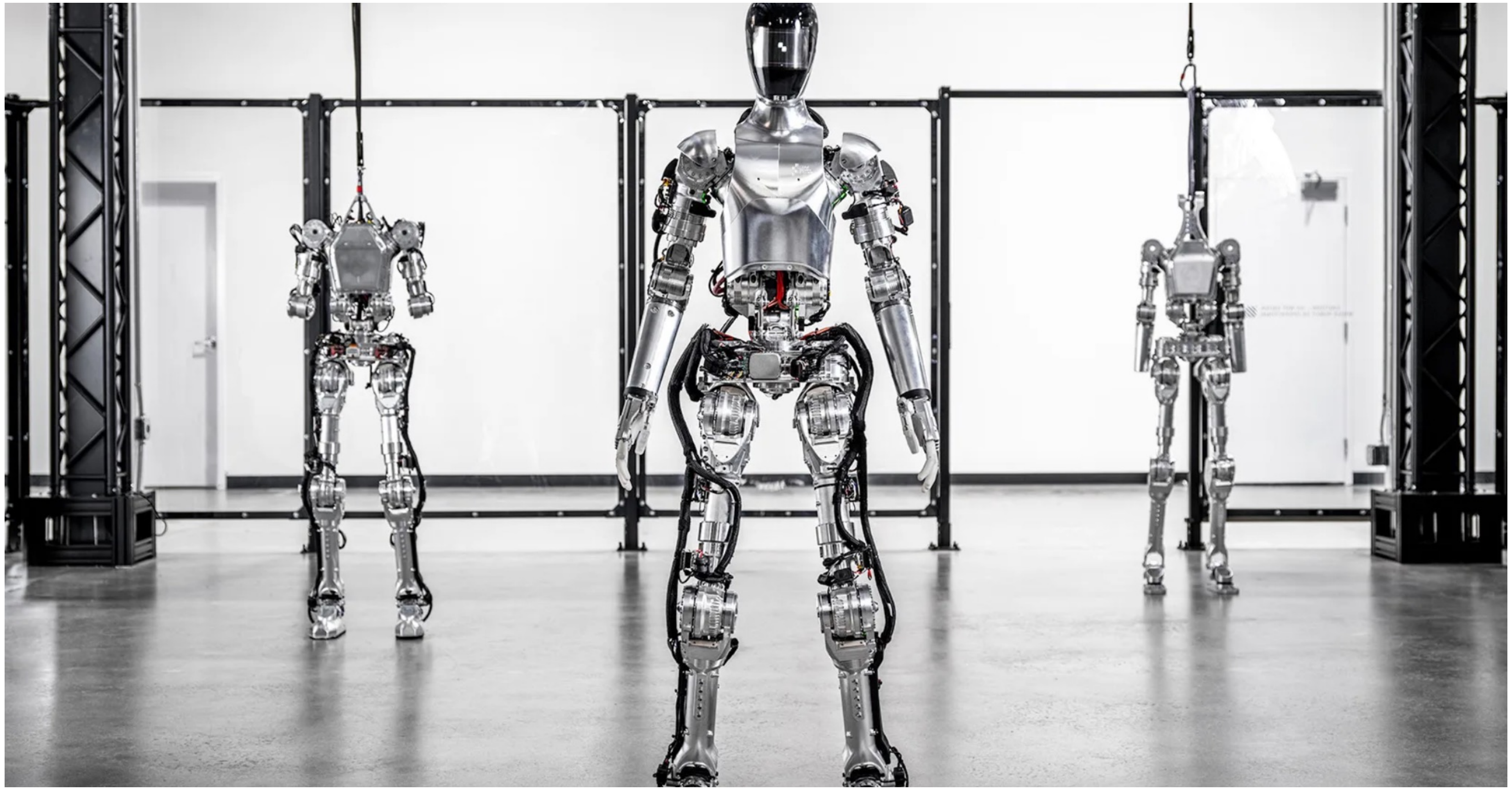
I-AI Interface Vision

Prompt: I need fire fighters to control a wildfire. They must be heat resistant, coordinated among themselves, and able to use human fire fighting equipment.

First, visualize them for me

Second, build them for me

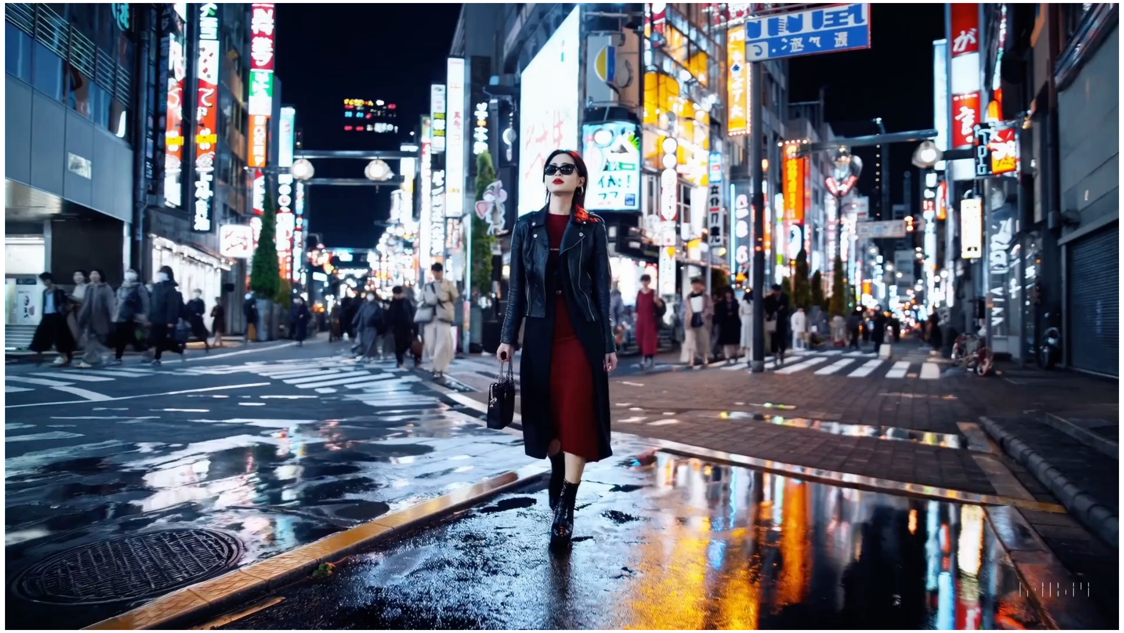
Third, deploy them to the fire



Today:

SORA is a step in this direction

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.



Costs Are A Challenge

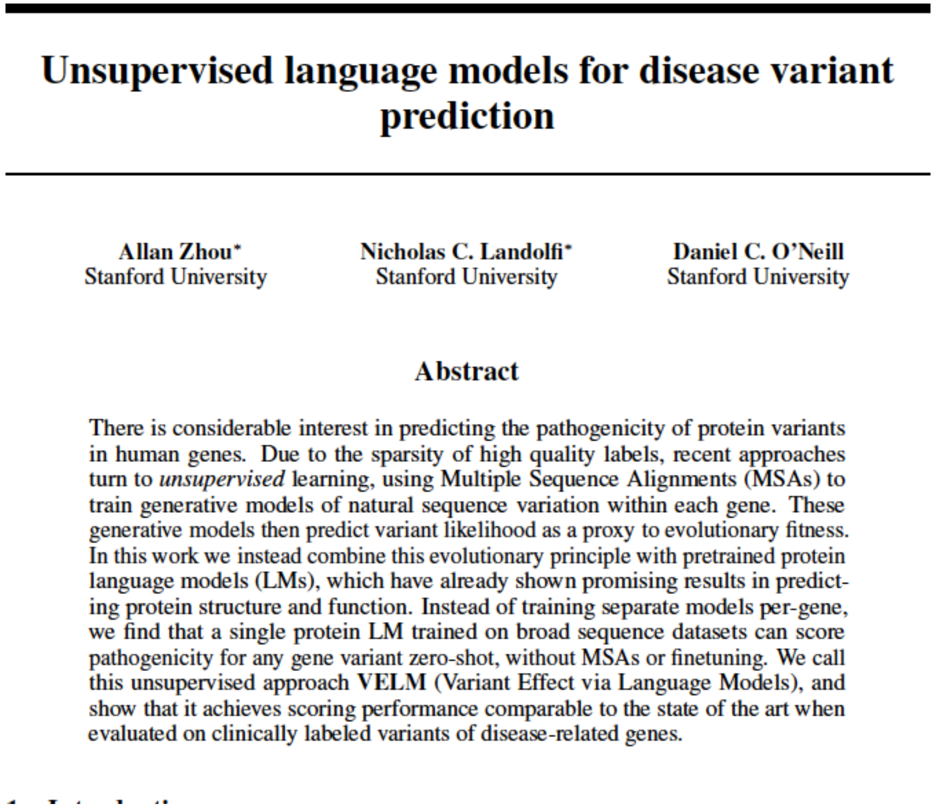
- Direct training costs of foundation models
    - Computer time $10’s-100’s million
    - Cost of getting the training data? Zero to infinite
- Per use costs are also high
    - 1000x marginal cost for an AI search vs. a traditional search

Pay-I startup

API	Max Cost
gpt-3.5-turbo	$0.008
gpt-3.5-turbo-0301	$0.008
gpt-3.5-turbo-0613	$0.008
gpt-3.5-turbo-16k	$0.064
gpt-3.5-turbo-16k-0613	$0.064
gpt-4	$0.48
gpt-4-0314	$0.48
gpt-4-0613	$0.48
gpt-4-32k	$3.84
gpt-4-32k-0314	$3.84
gpt-4-32k-0613	$3.84

Cost Example: A NeurIPS Paper

- $20,000 to **test** a few proteins on GCP (Stanford HAI grant)
- $Millions for foundation model



Unsupervised language models for disease variant prediction

Allan Zhou* Stanford University

Nicholas C. Landolfi* Stanford University

Daniel C. O'Neill Stanford University

Abstract

There is considerable interest in predicting the pathogenicity of protein variants in human genes. Due to the sparsity of high quality labels, recent approaches turn to *unsupervised* learning, using Multiple Sequence Alignments (MSAs) to train generative models of natural sequence variation within each gene. These generative models then predict variant likelihood as a proxy to evolutionary fitness. In this work we instead combine this evolutionary principle with pretrained protein language models (LMs), which have already shown promising results in predicting protein structure and function. Instead of training separate models per-gene, we find that a single protein LM trained on broad sequence datasets can score pathogenicity for any gene variant zero-shot, without MSAs or finetuning. We call this unsupervised approach VELM (Variant Effect via Language Models), and show that it achieves scoring performance comparable to the state of the art when evaluated on clinically labeled variants of disease-related genes.

Predictive Maintenance Example

Maintenance might makes over half of heavy asset OpEx

Very expensive assets might be inoperational because needed maintenance parts are not there

Predictive Maintenance promises to anticipate failures and the need for replacement parts in advance

Predictive Maintenance is most common I-AI use case discussed across several industries

Speakers!

Name	Company	Title	Topic
Dan ONeill	Stanford	Professor	Overview
Timothy Chou	Self	Board Member	Medical AI
Gerhard Kress	Siemens	SVP Digital Business	AI for Industry Digitization
Scott Penberthy	Google	Man. Director	AI Genomics
Sarah Elk	Bain & Company	Partner	AI Organization
Lapo Mori	McKinsey	Partner	AI for Process Industry
David Tepper	Pay-i	CEO and Founder	AI Cost Management
Kylan Gibbs	Inworld	CEO and Founder	AI as Interface
Thomas Higginbotham	C3.ai	Sr. Director	AI for Aerospace and Defense
Mathew John	Microsoft	Sr. Director	AIOps

Summary

- I-AI is everywhere
- I-IA is different
    - Complex models
    - Manage downsides
- Changing traditional industries
- Creating 100’s of startups
- Creating applications that can’t be done any other way

- Tons of challenges and opportunities
    - New applications
    - Technical unknowns
    - Performance limitations
    - Interface to humans
    - Many more

Questions