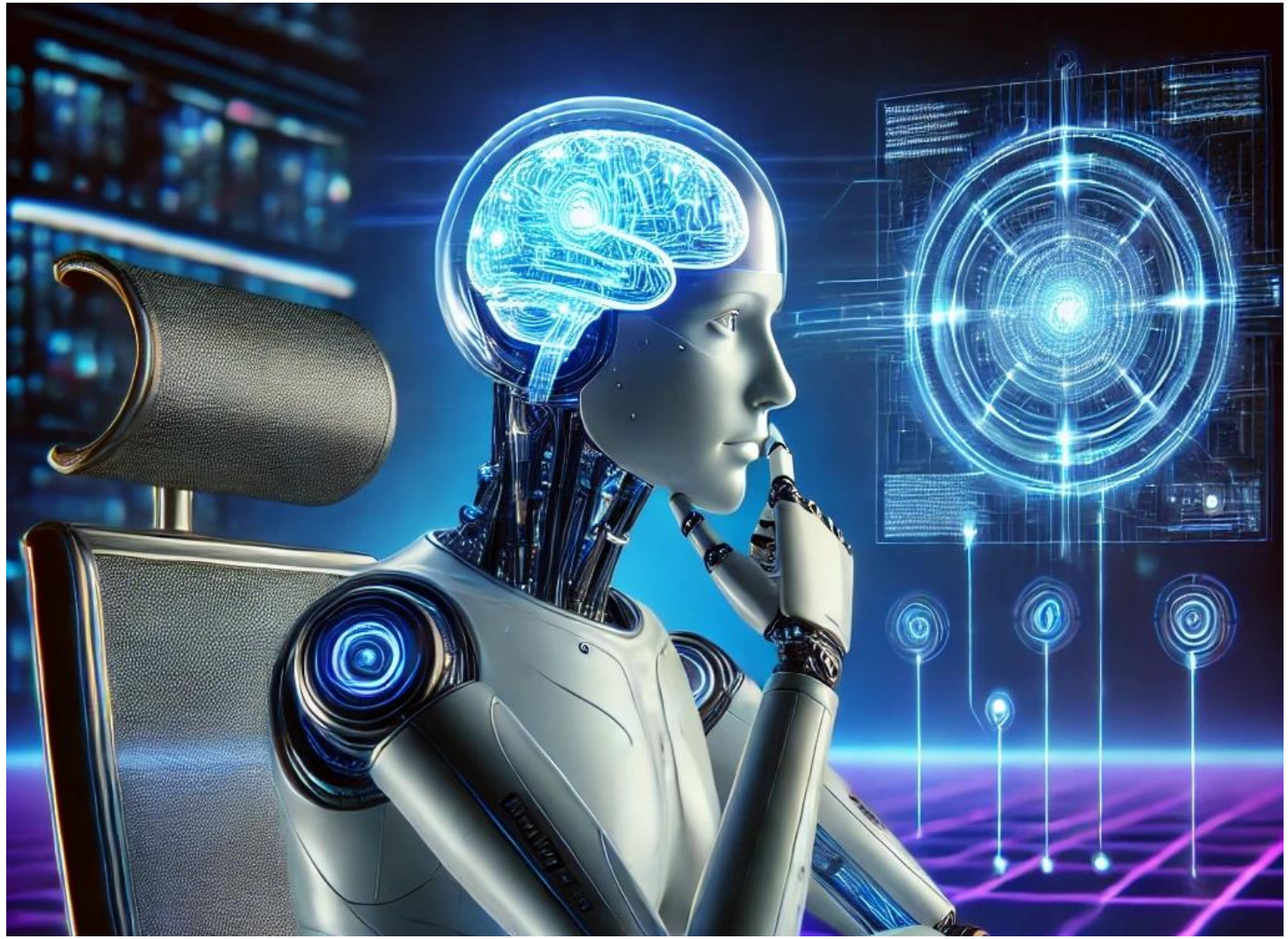
# CTO Forum: Cloud Cost Optimization: A Hands-on Generative AI Exercise

**Prof Iavor Bojinov**

# Plan

1. Recap (what we discussed last time)
2. AI & ML deep dive
3. Hands on data analysis exercise
4. Discussion, learnings, and next steps.



Harvard Business School

4 Hands-on sessions

## 1. Agents 101: Building Chatbots

Custom GPTs: Combining data and prompts for scaled impact

## 3. AI & ML with Gen AI

Leveraging generative AI to for data analytics, ML, and AI

## 2. Agents 201: Trigger based Agents

Building Simple Automation

## Preparing for the future

Ensuring your organization is planning for the next 5 years



Harvard Business School

# Provide Hands-On Training on the *Bleeding Edge* of Gen AI

# In the news

What has changed over the last few months?

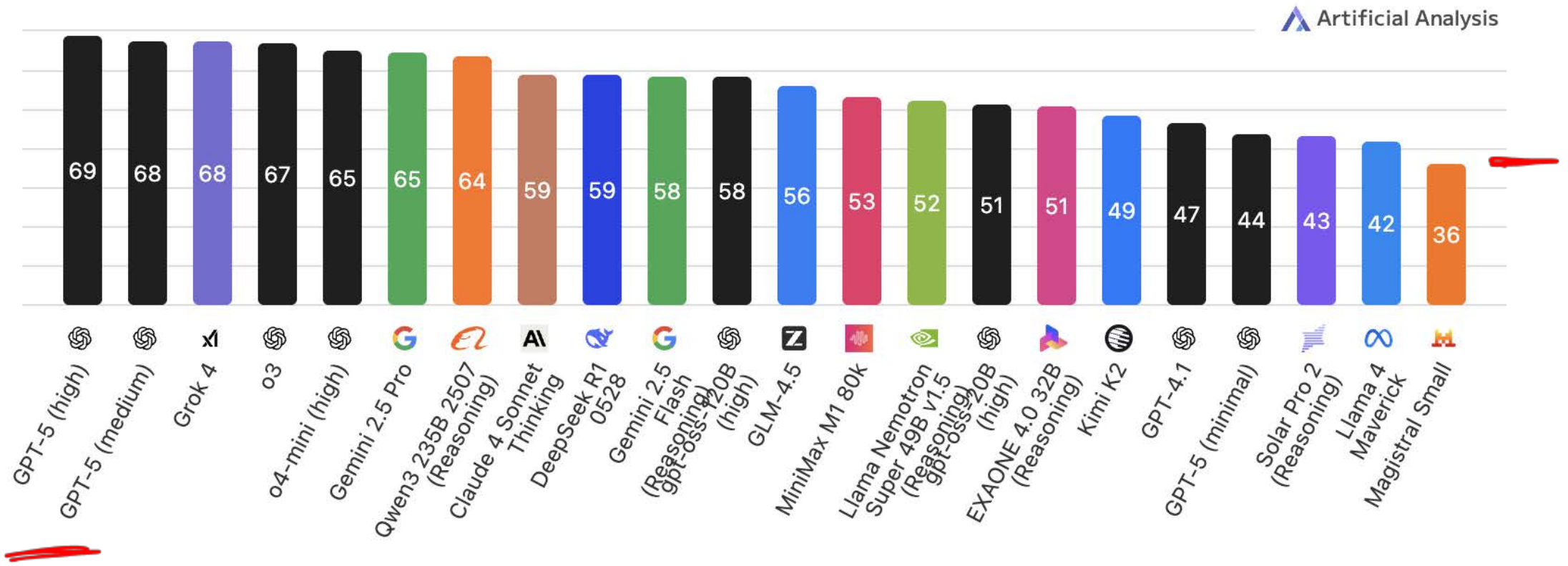


Harvard Business School

# State of the art today

## Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v2.2 incorporates 8 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, IFBench, AA-LCR



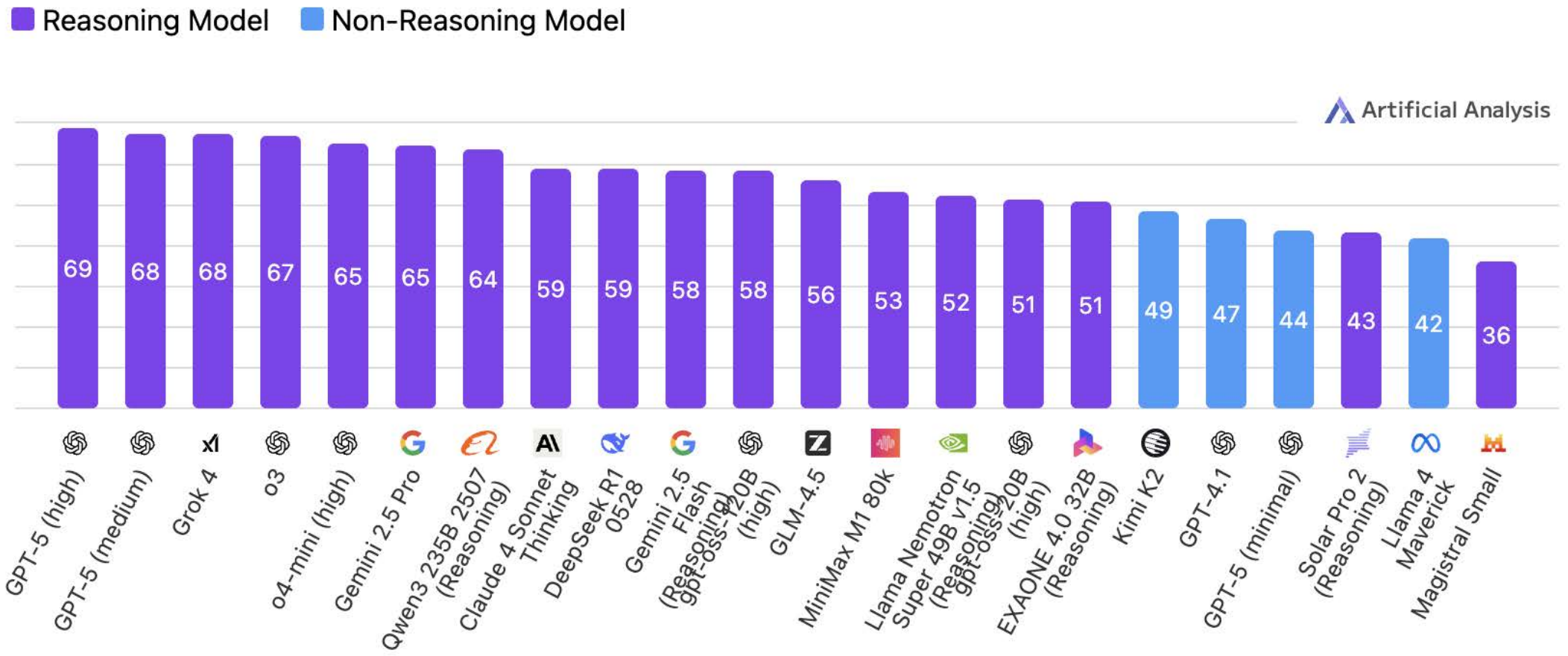



Harvard Business School

# It’s all about reasoning

## Artificial Analysis Intelligence Index by Model Type

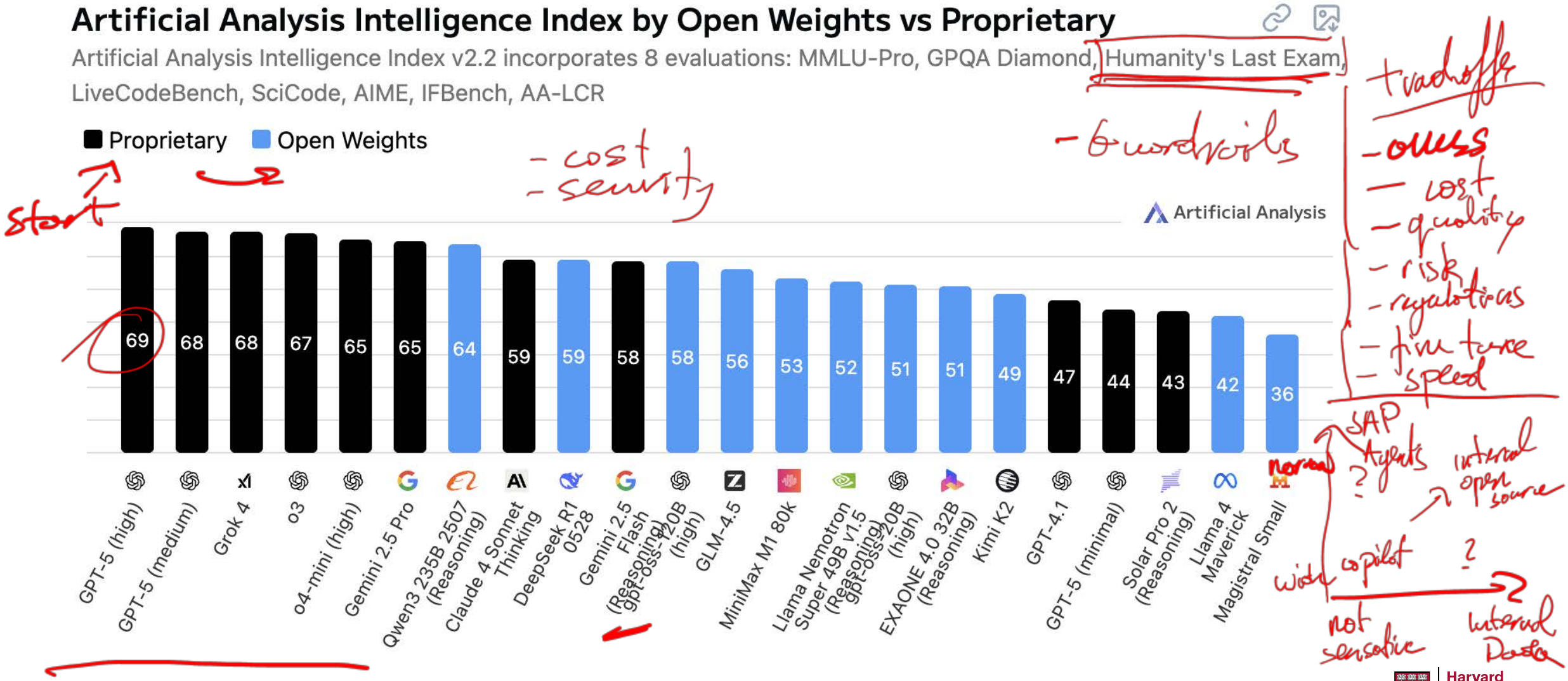
Artificial Analysis Intelligence Index v2.2 incorporates 8 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, IFBench, AA-LCR






Harvard Business School

# Growth in Open Weights






Harvard Business School

# Examples of Humanity's last exam

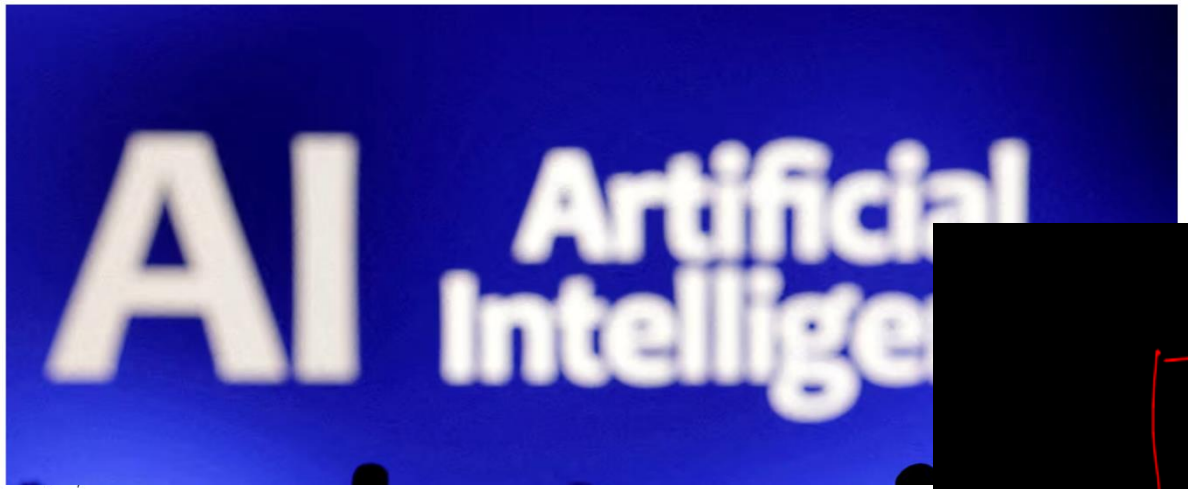

Reuters World Business Markets Sustainability Legal Commentary Technology Investigations More

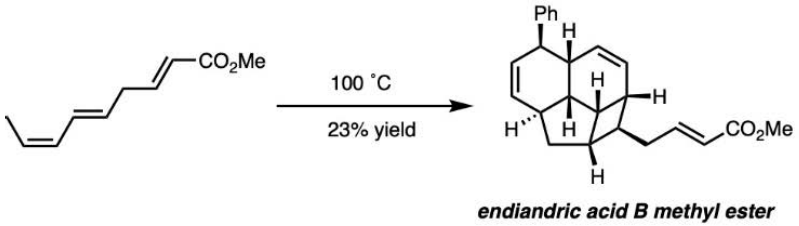
entation of a Roman inscription, originally found
. Provide a translation for the Palmyrene script.
of the text is provided: RGYNᵓ BT ḤRY BR ᶜTᵓ

## AI experts ready 'Humanity's Last Exam' to stump powerful tech

By **Jeffrey Dastin** and **Katie Paul**

September 16, 2024 10:42 PM GMT+2 · Updated September 16, 2024




*endiandric acid B methyl ester*

own is a thermal pericyclic cascade that

delooping of the symmetric group $\Sigma_7$ on 7 letters under the unique 0-simplex $*$ of $\mathbf{B}_\bullet\Sigma_7$.

How many natural cotransformations are there between $F$ and $G$?

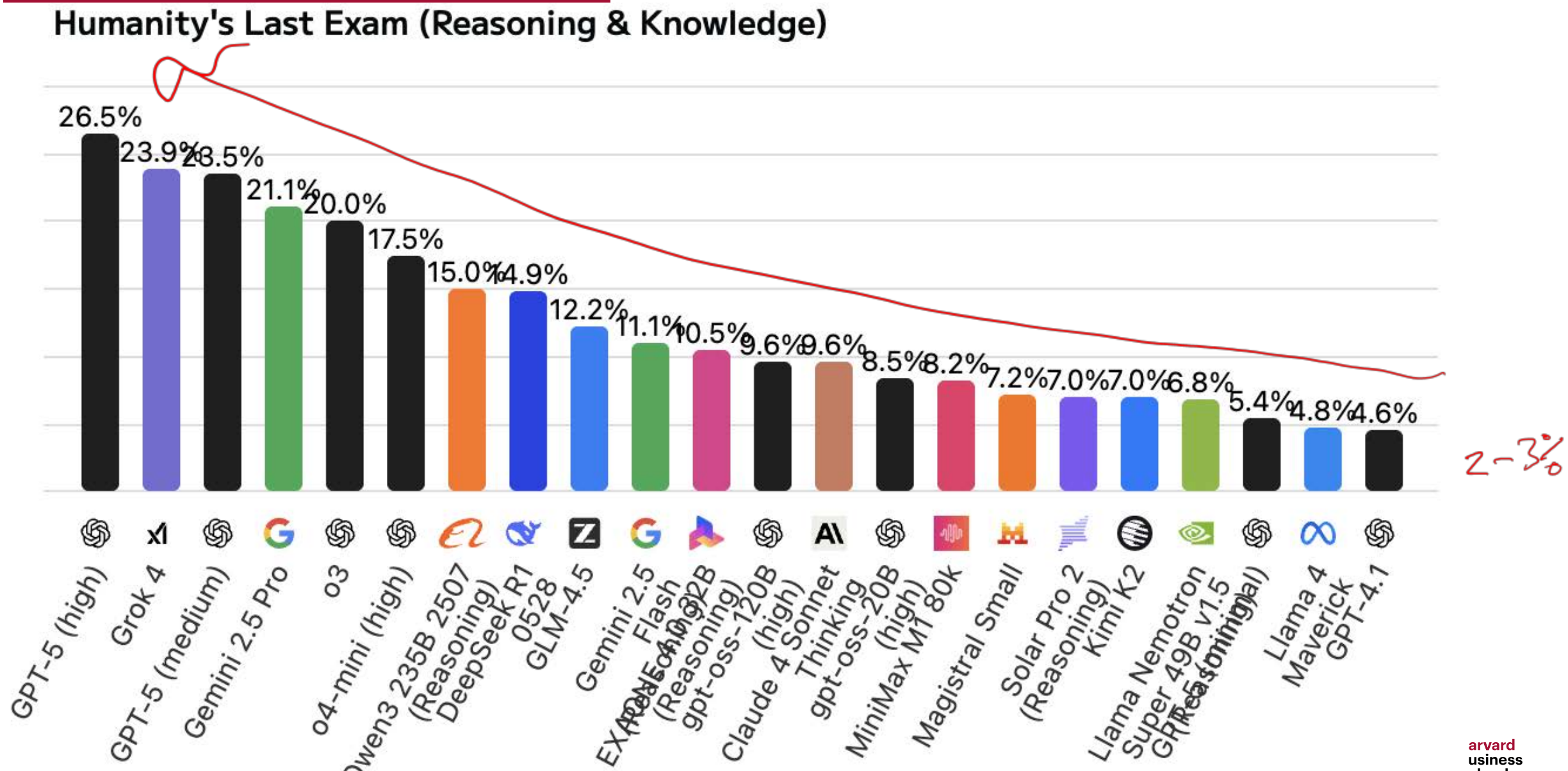
THE SHIFT

## When A.I. Passes This Test, Look Out

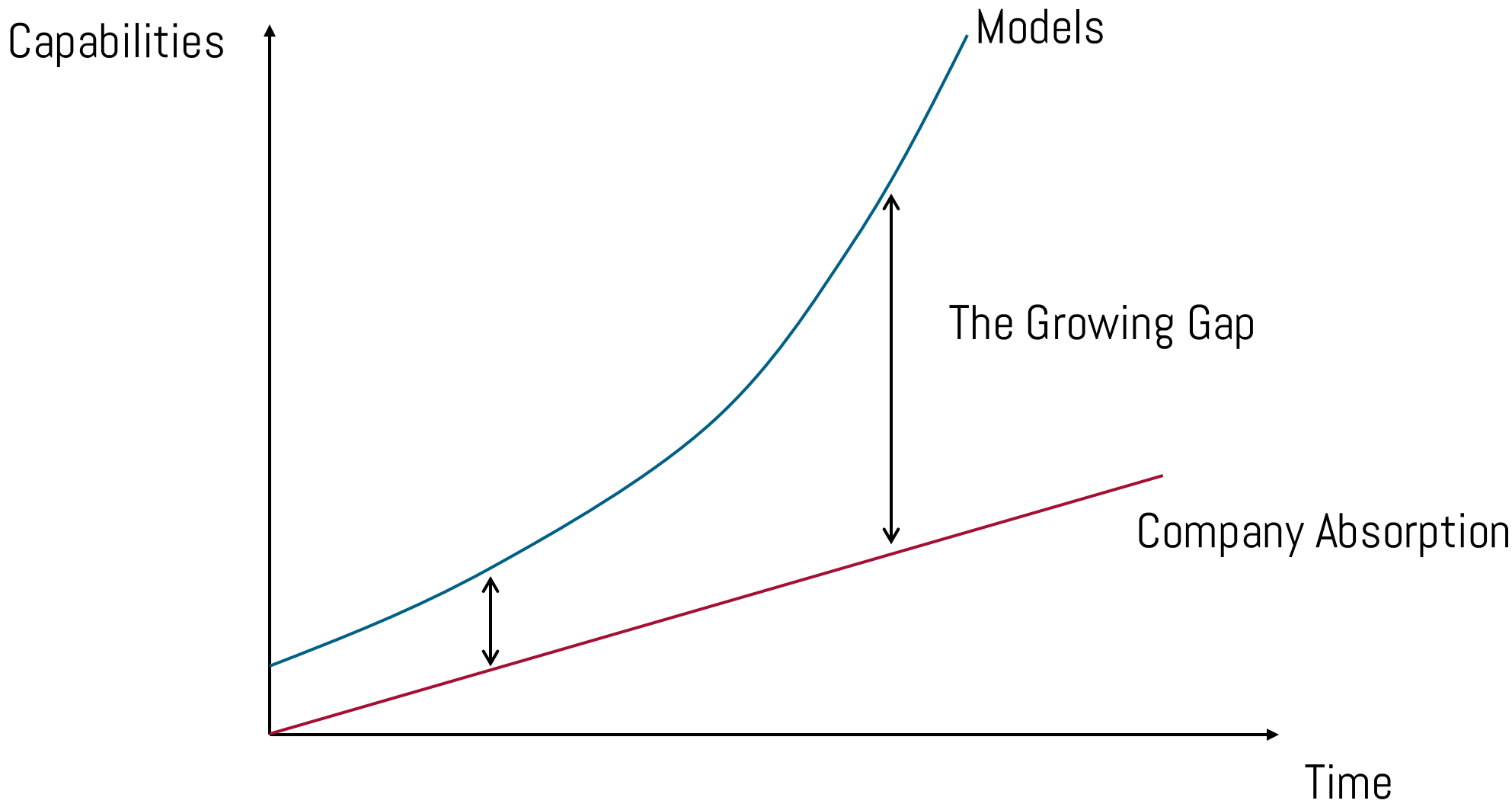
The creators of a new test called "Humanity's Last Exam" argue we may soon lose the ability to create tests hard enough for A.I. models.

n are the number of atoms on each component).

# Let’s see how the new models are doing…



# Growing gap






Harvard Business School

# Recap

What did we do last few time?

# Agents

| Feature | Chatbots | Agents |
|---|---|---|
| **Autonomy** | Responds when prompted | Can operate independently |
| **Proactiveness** | User-driven | Task-driven, can take action on its own |
| **Customization** | Adjusted via instructions & integrations | Programmed with logic, APIs, and workflows |
| **Use Case** | Chatbots, assistants | Automation, decision-making, multi-step tasks |
| **Memory** | Limited recall within chat (but growing!) | Can store & recall structured knowledge |
| **Risk** | Low, still requires human input and direction | High because of the higher level of autonomy |

# Session 0 (October):
# Creating AI-first snack company

1. Analyze current trends + identify an opportunity
2. Create a name, logo, and packaging prototype
3. Generate a recipe that scales
4. Identify customer segment and location of the initial launch
5. Create a targeted marketing campaign
6. Summarize the discussion & create a power point slide to pitch your company!

# 5 Use Cases of Generative AI

## Information Compression

The foundational ability to learn and store vast information in an efficient model. It's the "super-powered search engine" most people are familiar with.

### Content Creation

Instantly generate novel text, images, designs, and code, breaking through creative bottlenecks and accelerating production.

### Information Transformation

Restructure data and change its format. Translate a complex report into a simple memo or adapt communications for different audiences.

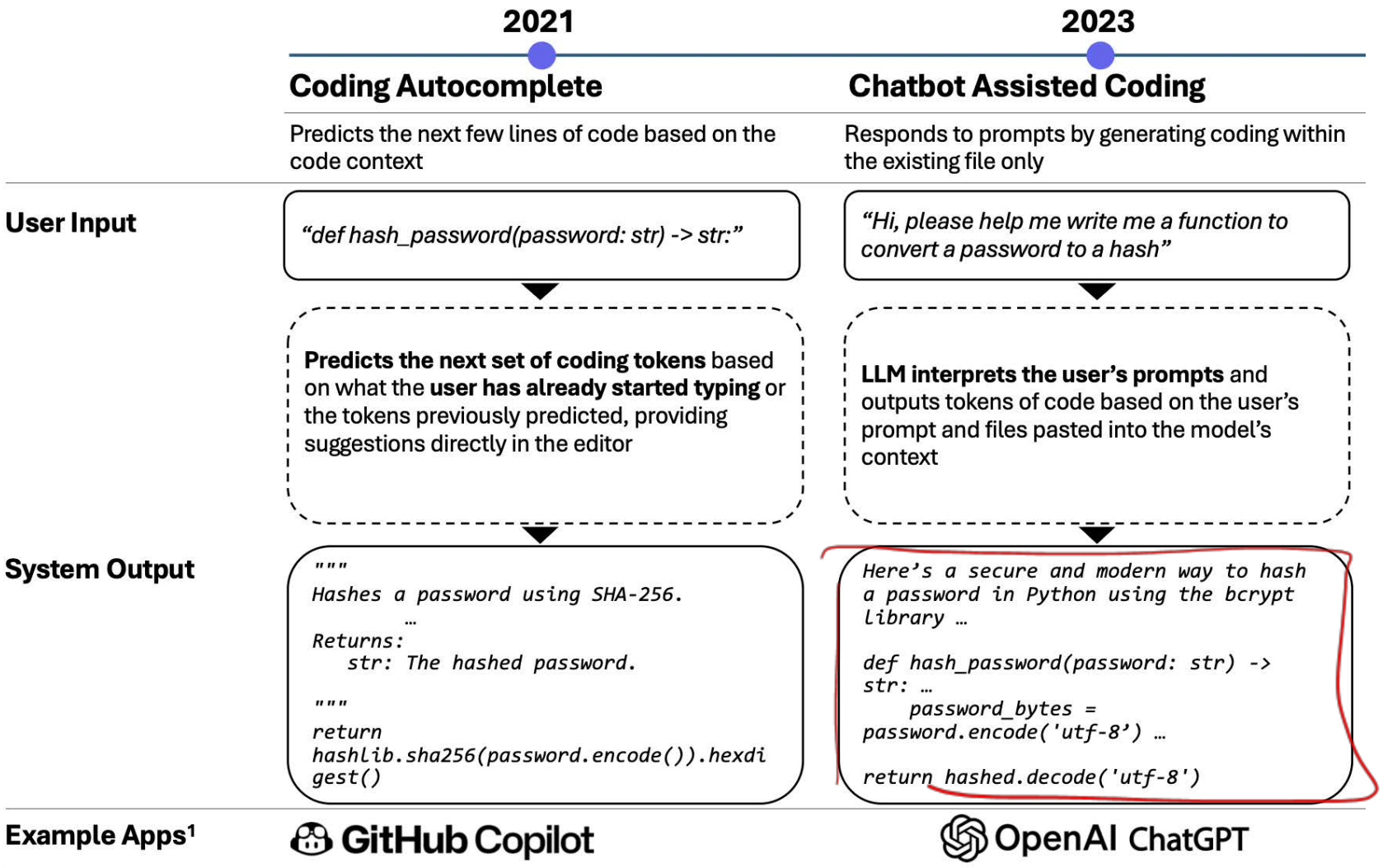
### Simulation & Prediction

Model future market scenarios, predict operational failures before they happen, and de-risk investments by simulating outcomes.

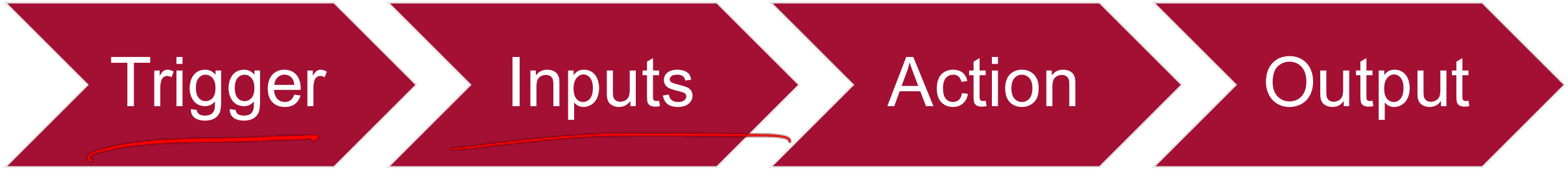
### Complex Reasoning

Tackle multi-step strategic challenges, from optimizing a global supply chain to planning a complex market entry strategy.

# Coding Example

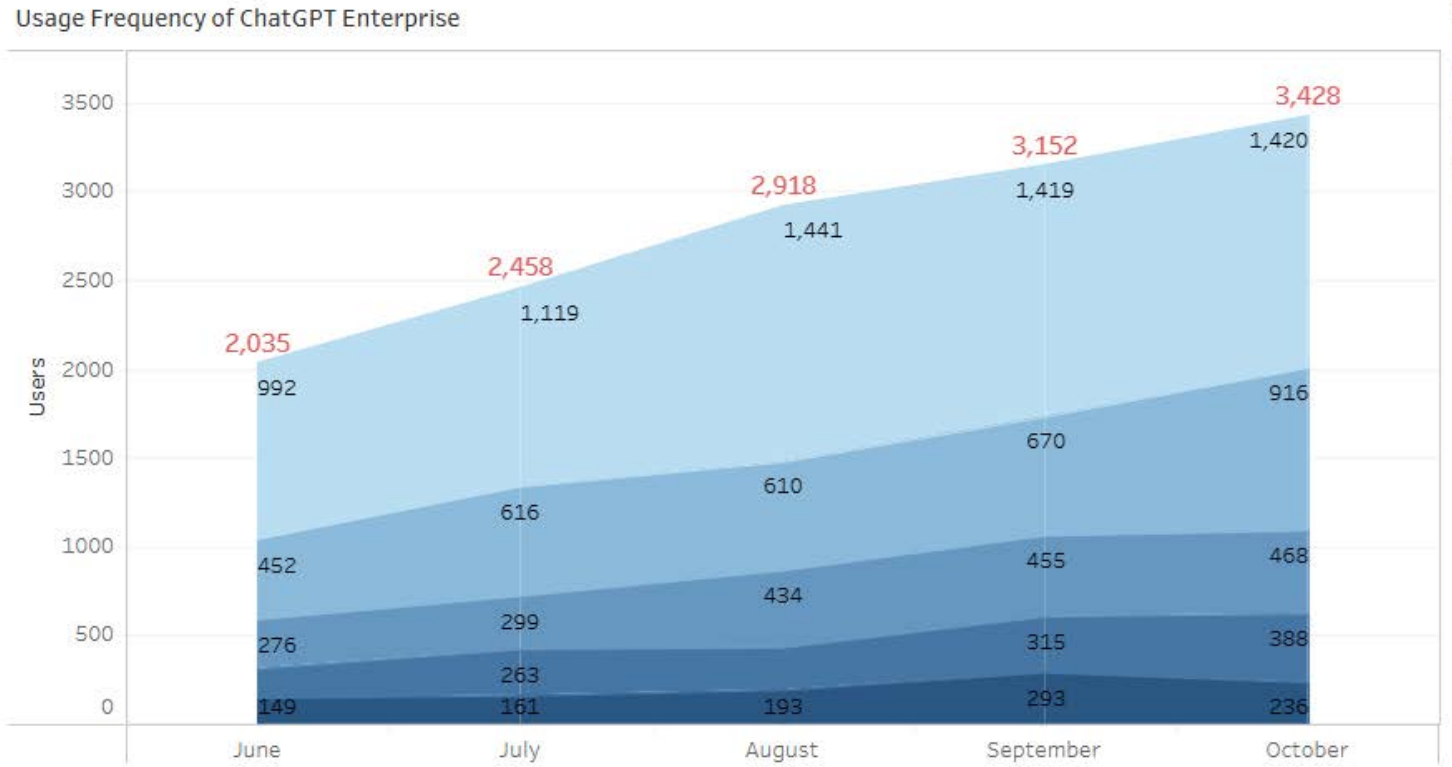


# 4 Key Ingredients for Agents




| Trigger | Inputs | Action | Output |
| --- | --- | --- | --- |
| Specific Time, Action, Prompt, etc. | Predefined Part of trigger Human input | Search, write, analyze, use tools, etc. | Email, agent, order, next step, etc. |

# Motivation for the GPT Evaluator: Modern's Gen AI Strategy




Harvard Business School

DRAFT
JANUARY 17, 2024

IAVOR BOJINOV
KARIM LAKHANI
ANNIKA HILDEBRANDT
JAMES WEBER

## Moderna: Democratizing Artificial Intelligence

In late December 2024, Vice President of AI Products and Innovation Brice Challamel met with CEO Stéphane Bancel, Chief People and Digital Technology Officer Tracey Franklin, and Chief Information Officer Brad Miller to review the adoption of generative AI at Moderna. Over the past year, the biotechnology company had provided all employees with access to OpenAI's ChatGPT Enterprise and encouraged them to incorporate the tool into their daily work. GPTs—Generative Pre-trained Transformers—were a powerful form of artificial intelligence that could reshape a variety of standard business processes. One of ChatGPT's key functionalities allowed users to create and share custom GPTs, each fine-tuned with specific instructions and data to deliver more accurate, relevant responses for particular use cases. (See **Exhibit 1** for Moderna's top 20 GPTs and **Exhibit 2** for GPT usage activity.)

From the outset, Moderna had been a digital-first, AI-focused company. Bancel famously described it as "a technology company that happens to do biology." By 2024, Moderna aimed to obtain 10 new drug approvals within three years. Bancel believed that sustained AI-driven innovation would enable the company's nearly 6,000 employees to keep pace with rival pharmaceutical firms employing more than 100,000 people. To foster this innovation, the company encouraged employees to develop, publish, and maintain custom GPTs, embracing a model akin to the Apple App Store or Google Play Store, where employees could share their creations with each other. Yet AI was not without its flaws. Employees were still learning to wield these emerging tools, and GPTs sometimes produced inaccuracies—so-called "hallucinations." Challamel recognized that as a publicly traded and highly regulated pharmaceutical company, GPT errors in critical processes could have serious consequences for Moderna. To balance risk management with speed and innovation, he implemented governance practices for AI use.

As the Moderna leadership team discussed generative AI adoption, concerns about the use and governance of custom GPTs began to resurface. During the meeting, Challamel highlighted the recent spike in usage of the Self-Review GPT, a tool assisting employees with quarterly and annual performance reviews. Franklin expressed concern: "I'm worried that the Self-Review GPT is potentially problematic, as it is augmenting—and to some extent replacing—a critical process in developing employees. Maybe there are some processes and work that should be kept off-limits? How can Moderna lead the way in pioneering human-AI augmentation in all the work that gets done here?" Seizing on the point, Challamel turned the group's attention to a new GPT called DoseID, created by physician and medical writer Lee Quist, which provided drug dosing recommendations for clinical

Professors Iavor Bojinov and Karim Lakhani, Research Associate Annika Hildebrandt, and Case Researcher James Weber (Case Research & Writing Group) prepared this case. It was reviewed and approved before publication by a company designate. Funding for the development of this case was provided by Harvard Business School and not by the company. HBS cases are developed solely as the basis for class discussion. Cases are not intended to serve as endorsements, sources of primary data, or illustrations of effective or ineffective management.

# Agents 101: Building Custom Chatbots

## Key Components

1. **Base Model Selection** – Choose GPT-4 or future variants.
2. **Custom Instructions** – Define behavior, tone, and constraints.
3. **Knowledge Augmentation** – Upload documents, databases, and private datasets.
4. **Tool & API Integrations** – Extend capabilities (e.g., code execution, web access, external APIs).

GPT Risk Matrix

| | | | | |
|---|---|---|---|---|
| Severity | 3 - Critical | B | C | C |
| | 2 - Medium | A | B | C |
| | 1 - Low | A | A | B |
| | | 1 - Individual | 2 - Team | 3 - Company |
| | | Impact | | |

A – Low criticality

B – Medium criticality

C – High criticality

Harvard Business School

DRAFT
FEBRUARY 4, 2025

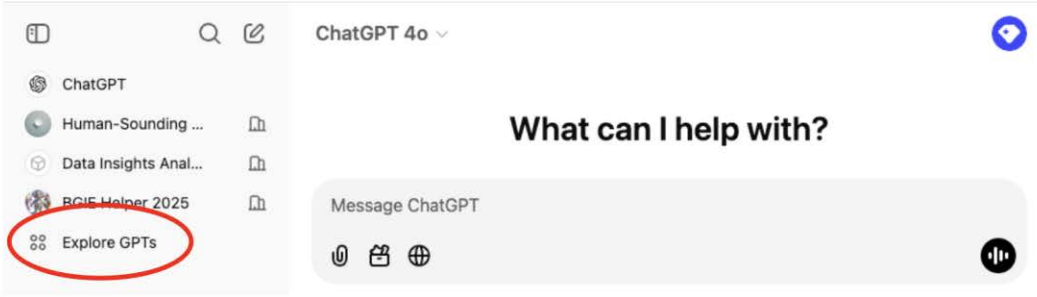
IAVOR BOJINOV
ANNIKA HILDEBRANDT

## Building a Custom GPT and AI Agent Evaluator

Custom GPTs can quickly and easily be built using ChatGPT. These models are the precursors to agents as they can combine multiple sources of information with custom instructions; however, they are still intended to be worker companions as they lack the autonomy to act. Nevertheless, they provide a powerful tool that enables employees to create a wide range of custom applications, from personal benefits assistants to coding support. One of the benefits of custom GPTs is they can be shared broadly with others, either in your enterprise or with other ChatGPT users.

This decentralized approach that allows everyone to create and share agents raises some vital governance questions. To address governance issues, you will be creating a custom GPT that creates governance guidelines for the creators of GPTs. Your evaluator will help GPT builders understand what standards custom GPTs and AI agents must adhere to.

The instructions below are tailored for the ChatGPT premium accounts with access to build GPTs. Begin by navigating to chat.com and logging in or creating an account. In the left-hand navigation bar, select 'Explore GPTs' to reach the GPT home page.




### Task 1: Creating your first GPT – Basic Configuration

Being by creating your first GPT and configuring the details.

1) On the GPT home page, select the create button in the top right to be brought to the configuration page.

# Agents 201: Trigger-based Agents


Build an agent that searches Google News for company-specific data breaches and sends a daily report.

1. Work as a breakout group by designating 1 person to be the primary “writer.” Have that person screen share.
2. Everyone else should work in parallel using similar prompts and steps.
3. The final output requires showcasing a powerful and useful AI Agent.
4. You’ll have about 45-1 hour to work on this.
5. When you finish the basic agent, build on more complicated features and tailor it to your setting.

Harvard Business School

DRAFT
MAY 12, 2025

IAVOR I. BOJINOV
ANNIKA HILDEBRANDT

## Creating a Custom AI Agent for Cybersecurity Incident Updates

### Introduction

Custom AI agents can leverage generative AI to automate different workflows. These agents can enable employees to create a wide range of custom applications, from asset creation to A/B testing.

To test the power of AI agents, you will be creating an AI that alerts you to possible cybersecurity incidents at your organization. The agent will search for news articles relating to cybersecurity incidents at your company, identify the most recent results, and create and send an email with the findings. While this is a relatively simple example, it should demonstrate the power of AI agents and inspire you to create AI agents for your organizational use cases.

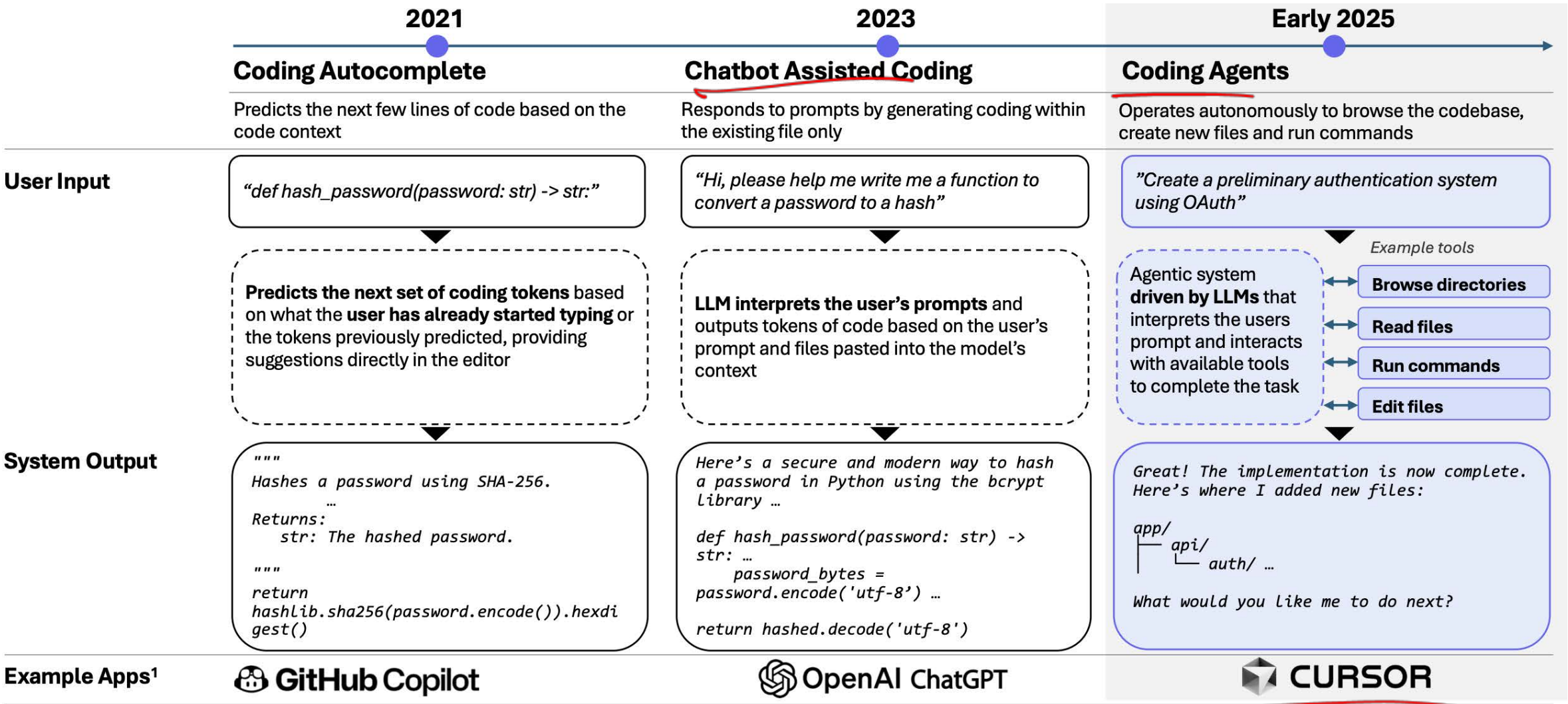
The instructions below are tailored for the MindStudio AI platform. Begin by navigating to mindstudio.ai and creating a free account. With a free account, you are able to build up to 3 custom agents a month and run workflows 1000 times (although please note that usage limits are subject to change).

### Preliminaries: Familiarizing yourself with Workflows and Variables

Most agentic AI platforms leverage what are known as workflows. Workflows define a flow of different tasks that should be completed in a specific sequence. These workflows may consist of both AI and non-AI components, often called blocks. For example, a block may create text based on some inputs using generative AI. Other blocks may not use AI, completing tasks such as sending an email or even running some traditional Python code. These workflows combine different technologies to create a powerful automated process.

Business School

# Coding with Agents






Harvard Business School

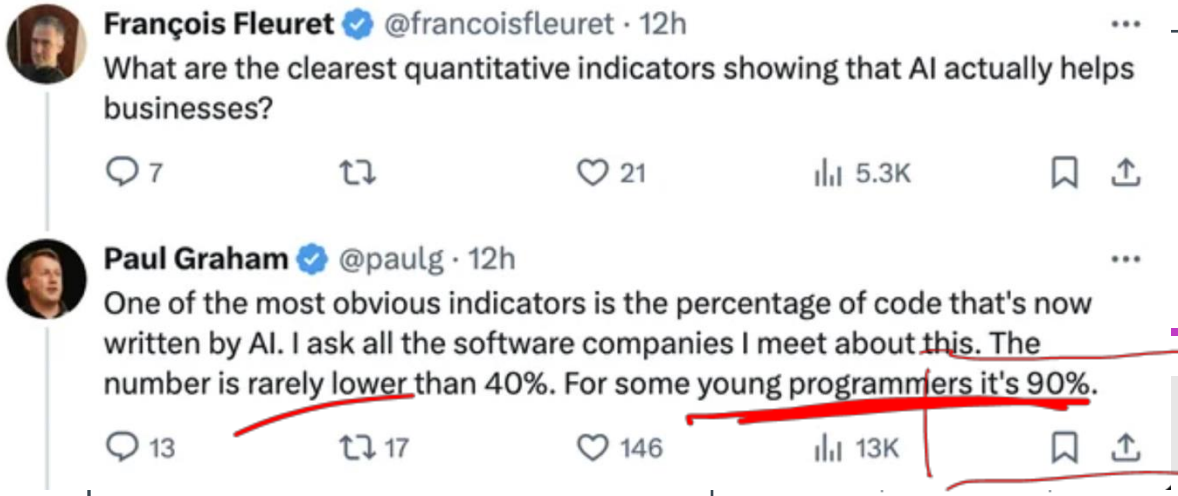
# Operations: More efficient processes, better outcomes

Change Management

TECH

## Satya Nadella says as much as 30% of Microsoft code is written by AI

PUBLISHED TUE, APR 29 2025·9:33 PM EDT | UPDATED TUE, APR 29 2025·9:58 PM EDT



François Fleuret @francoisfleuret · 12h

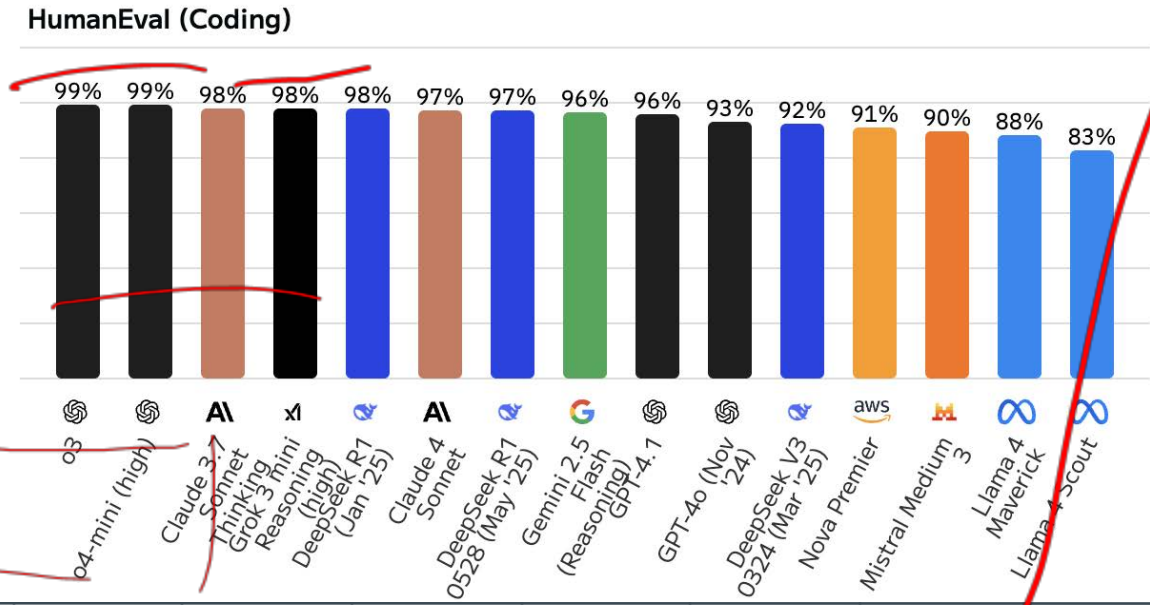
What are the clearest quantitative indicators showing that AI actually helps businesses?

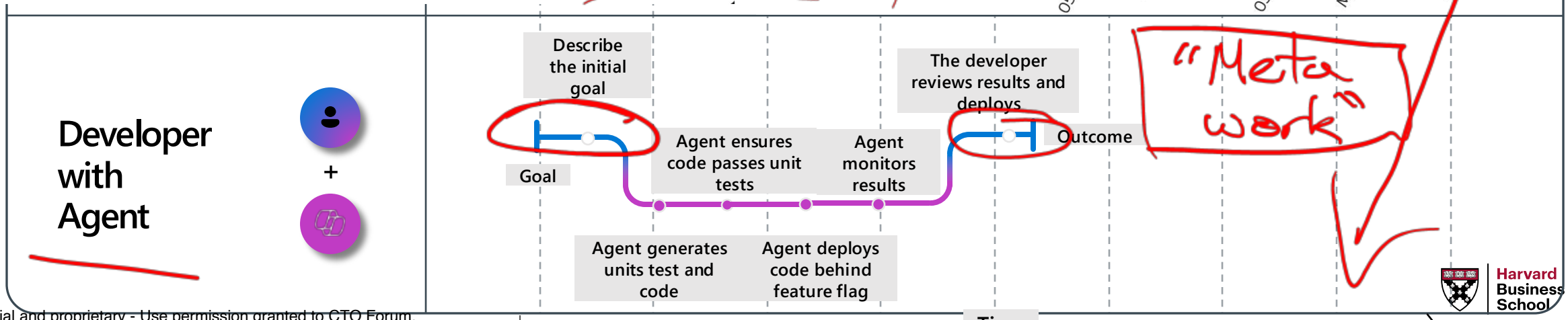
7 21 5.3K

Paul Graham @paulg · 12h

One of the most obvious indicators is the percentage of code that's now written by AI. I ask all the software companies I meet about this. The number is rarely lower than 40%. For some young programmers it's 90%.

13 17 146 13K




Harvard Business School

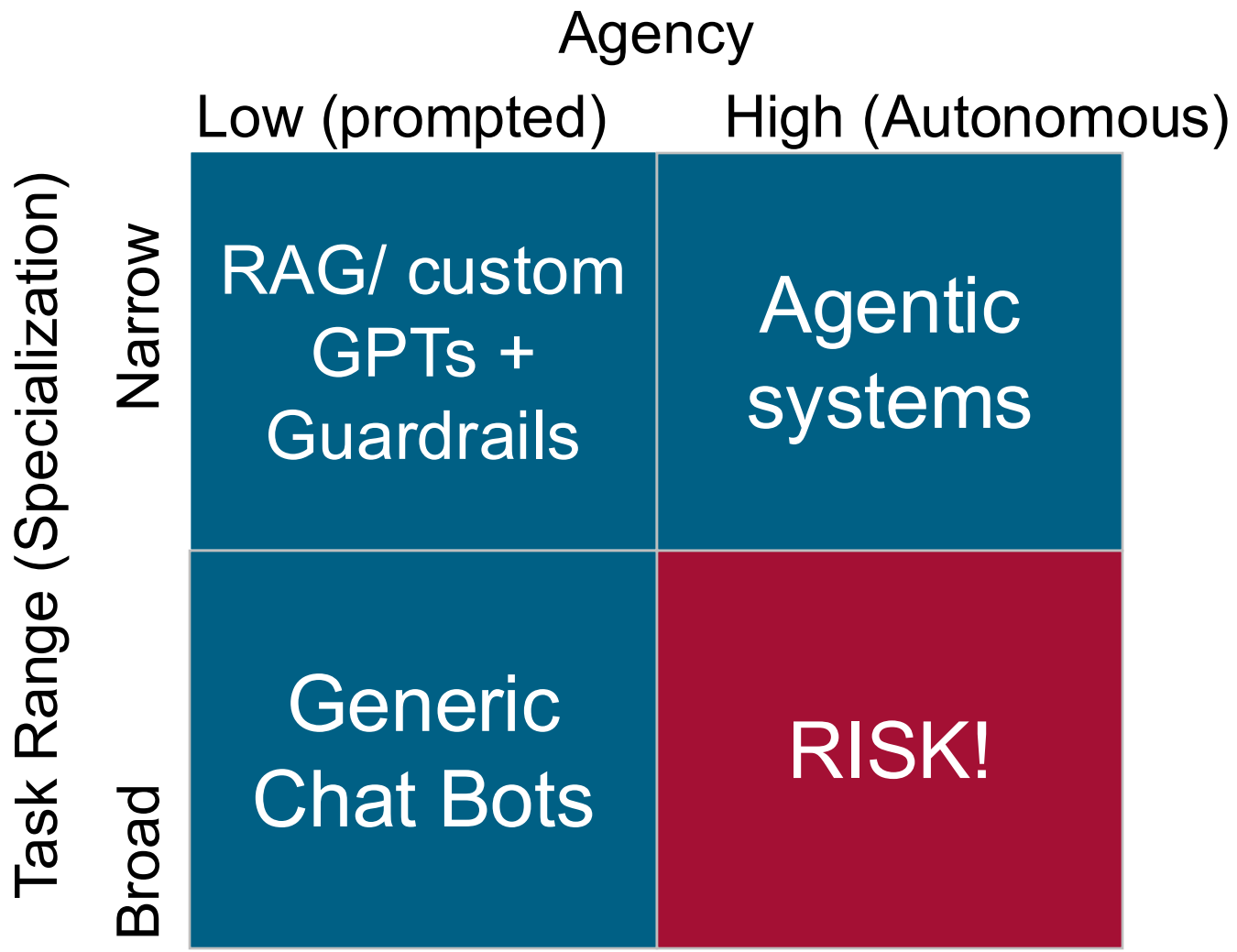
# Recap

## Frameworks

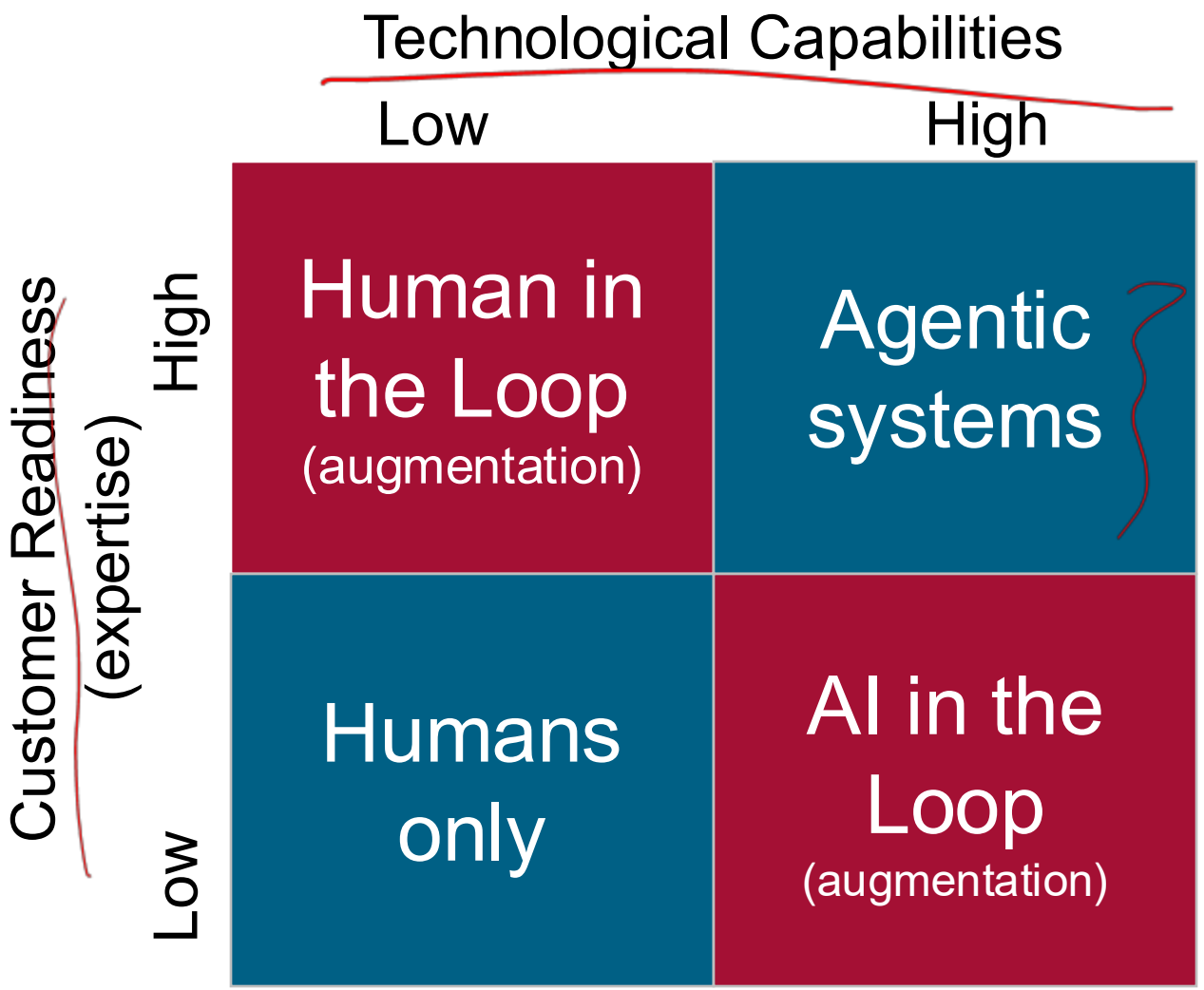


Harvard Business School

# Range vs Agency trade-off



# Customers & Employees not ready…



# Three modes of democratizing AI

## CONSUMER MODEL

## CONTROLLED DEMOCRATIZATION

## PRODUCER MINDSET

# Consumer Model

Employees have access to centrally managed tools (internal/external).

**Tools:** GitHub Copilot, LLMs through sandbox, etc.

**When:** Low level of employee readiness

**Examples:** Most Companies (P&G, Pernod Ricard, etc.)

**Pros:**

- Helps employees get accustomed to new tools
- Captures basic efficiencies in operations
- Centralized process redesign work

**Cons:**

- No AI network effect
- Limited focus on self-improvement
- Challenges in adoption as limited value

Harvard Business School

# Controlled Democratization

Employees have access to centrally managed tools (internal/external) and are empowered to build their own automations for specific functional patterns (such as RAGs or translators).

**Tools:** Either custom-built or generic tools like Microsoft Copilot Studio and ChatGPT Enterprise with extensive control over development and sharing.

**Functional Patterns:** Summarization, internal Q&A through RAGs, etc.

**When:** Medium level of employee readiness

**Examples:** A few Companies (JP Morgan, etc.)

**Pros:**

- Enable employees to build and share some customization
- Basic network effects
- Begin to transform individual work

**Cons:**

- Challenges in adoption as limited value
- Requires changing mindset: consumer → producer.

Harvard Business School

# The 6 most common use democratized functional patterns

1. Summarization of documents
2. Translation
3. Document Review
   - Legal, security, etc.
4. Data Analysis
5. Content Generation
   - Emails, marketing briefs, images, etc.
6. Knowledge Repository through Retrieval-Augmented Generation (RAG)
   - Benefits, answering questions, etc.

# Producer Mindset

Employees have access to centrally managed tools (internal/external) and are empowered to build their own automations for **ANY** use cases (such as RAGs or translators) with a centralized effort to transform all working processes.

**Tools:** Either custom-built or generic tools like Microsoft Copilot Studio and ChatGPT Enterprise, with limited control over development and sharing.

**When:** High level of employee readiness

**Examples:** A few companies (Moderna, etc.)

**Pros:**

- Widespread innovation through a producer mindset
- Significant network effects
- Fundamental rewiring of the company's operating model

**Cons:**

- Managing complexity (e.g., 1,000s of custom GPTs at Moderna)

# Keep your AI projects on track (HBR):

**SELECTION**

Prioritizing & sequencing effectively

**DEVELOPMENT**

Accelerate through the AI Factory

**EVALUATION**

Does it really

**ADOPTION**

Release and drive growth

**MANAGEME**

Monitor, manag
improv


HBR / Spotlight Article / Keep Your AI Projects on Track

Keep Your AI Projects on Track

Most go off course. To make sure yours succeed, consider these five steps. by Iavor Bojinov

From the Magazine (November–December 2023) / Reprint S23063

Michael Brandon Myers

When I worked as a data scientist at LinkedIn in 2018 and 2019, AI was of interest only to a small team of people in the data science organization with advanced degrees in statistics or computer science. AI —and especially its newest star, generative AI—is now a central theme in corporate boardrooms, leadership discussions, and casual exchanges among employees eager to supercharge their productivity. The is so fundamental that "Data Science for Managers," a co create to teach MBA students how to develop, lev is now a first-year requirement at Harvard

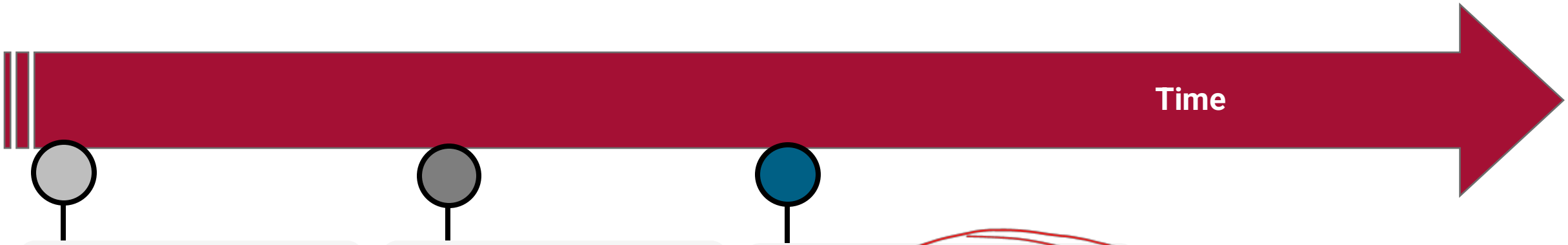
https://hbr.org/2023/11/keep-your-ai-projects-on-track

# AI for democratizing Data Analytics & ML

AI/ML

Harvard Business School

# AI Innovation Has Progressed In Waves




**Wave 1: Cybernetics (1950s-70s)**

- Field of AI research was founded and key **concepts were first described, including early neural networks**

**Wave 2: Trained experts (1980s-90s)**

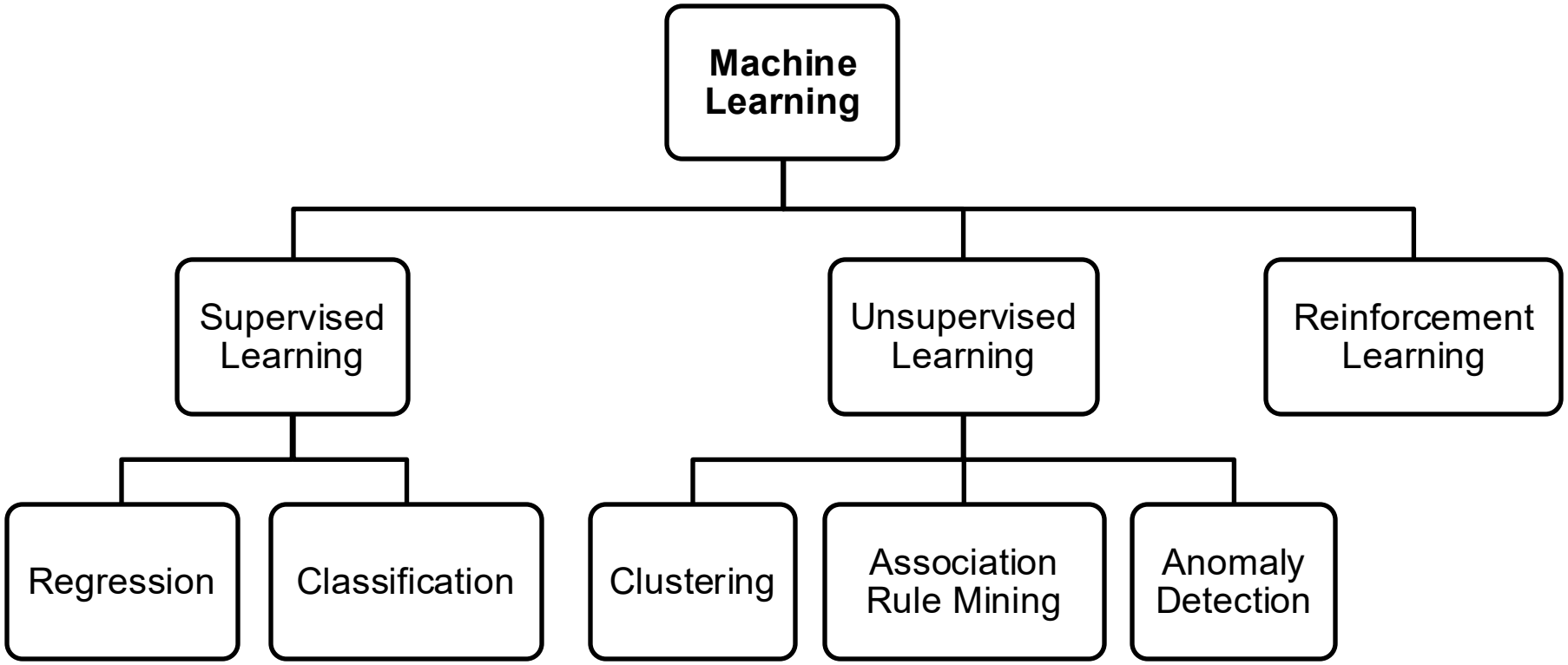
- Expert systems were developed to **solve problems about a specific domain of knowledge**

**Wave 3: Good learners (2000s-)**

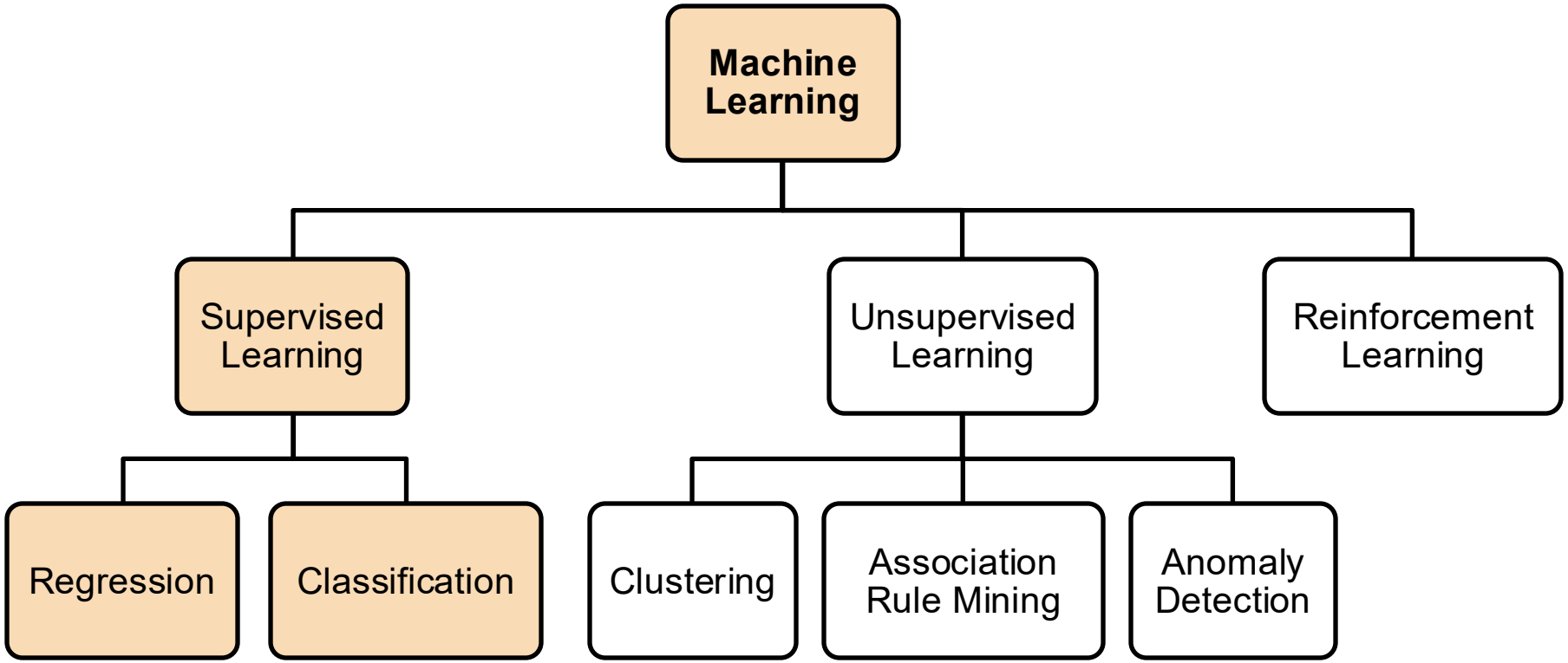
- Progress in artificial neural networks, “deep” learning, CNNs, RNNs led to **notable achievements in text, speech, language and vision processing**

# Good learners: Machine Learning

Machine learning is a set of algorithms and statistical methods that are trained on data to produce some sort of predictions.

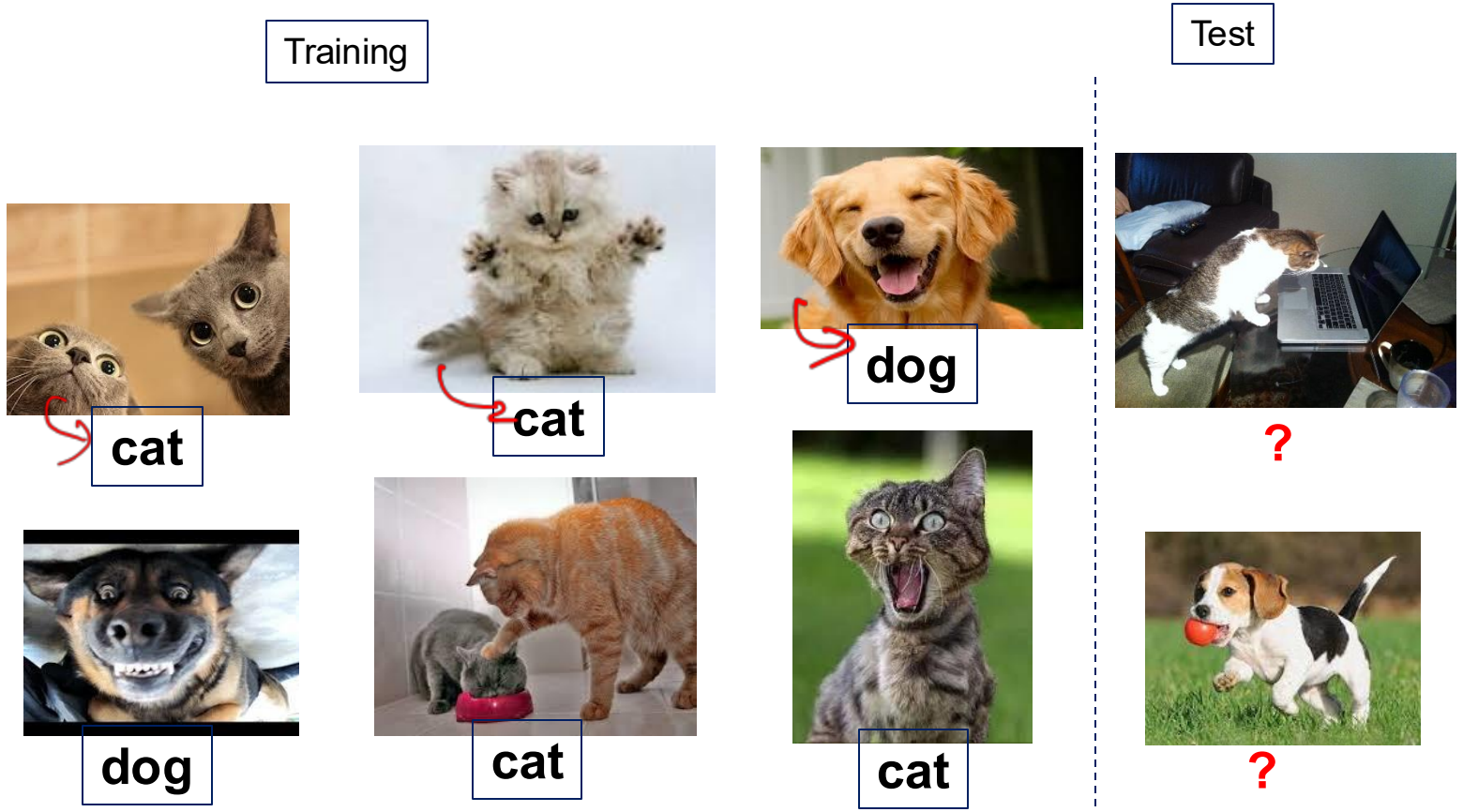


# Machine Learning



# Supervised Learning

- Predict outcome of new observations given set of labeled observations



# Supervised Learning

- Predict outcome of new observations given set of labeled observations

Training

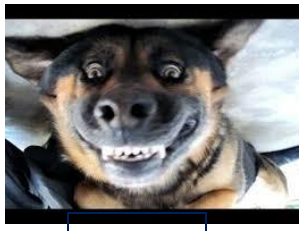
cat

cat

dog

Test

Cat



dog

cat

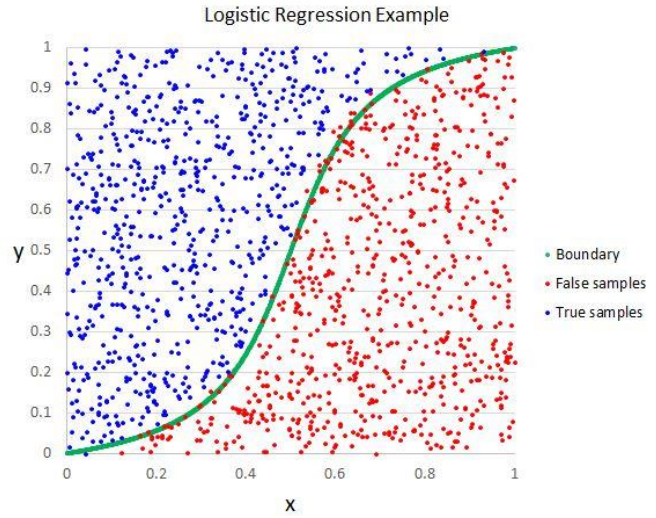
cat

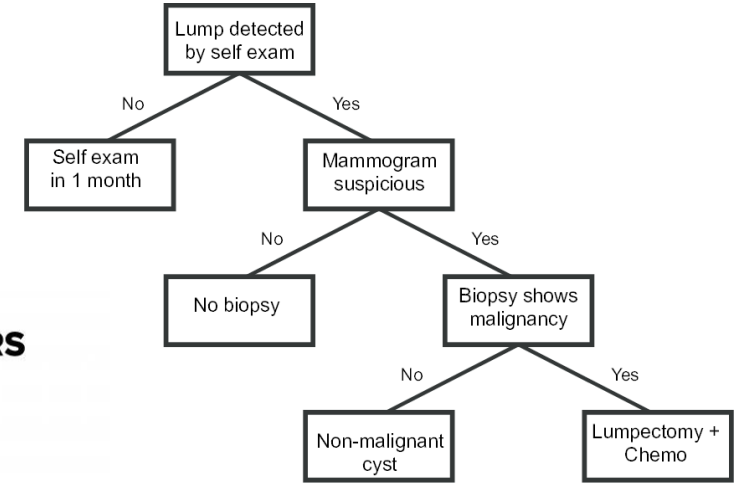
Dog

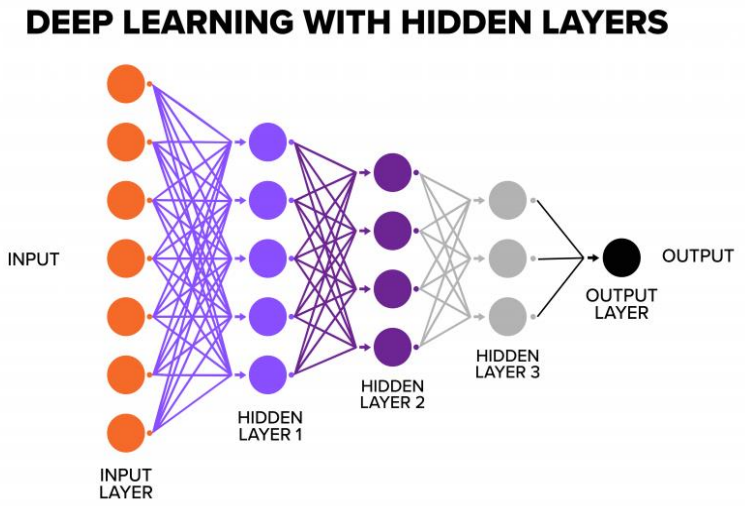
HARVARD | BUSINESS | SCHOOL

# Methods

- **Linear Regression**
- Logistic Regression
- **Decision Trees**
- Random Forests
- Gradient boosting
- Neural Networks
- Support Vector Machines
- …and so many others!

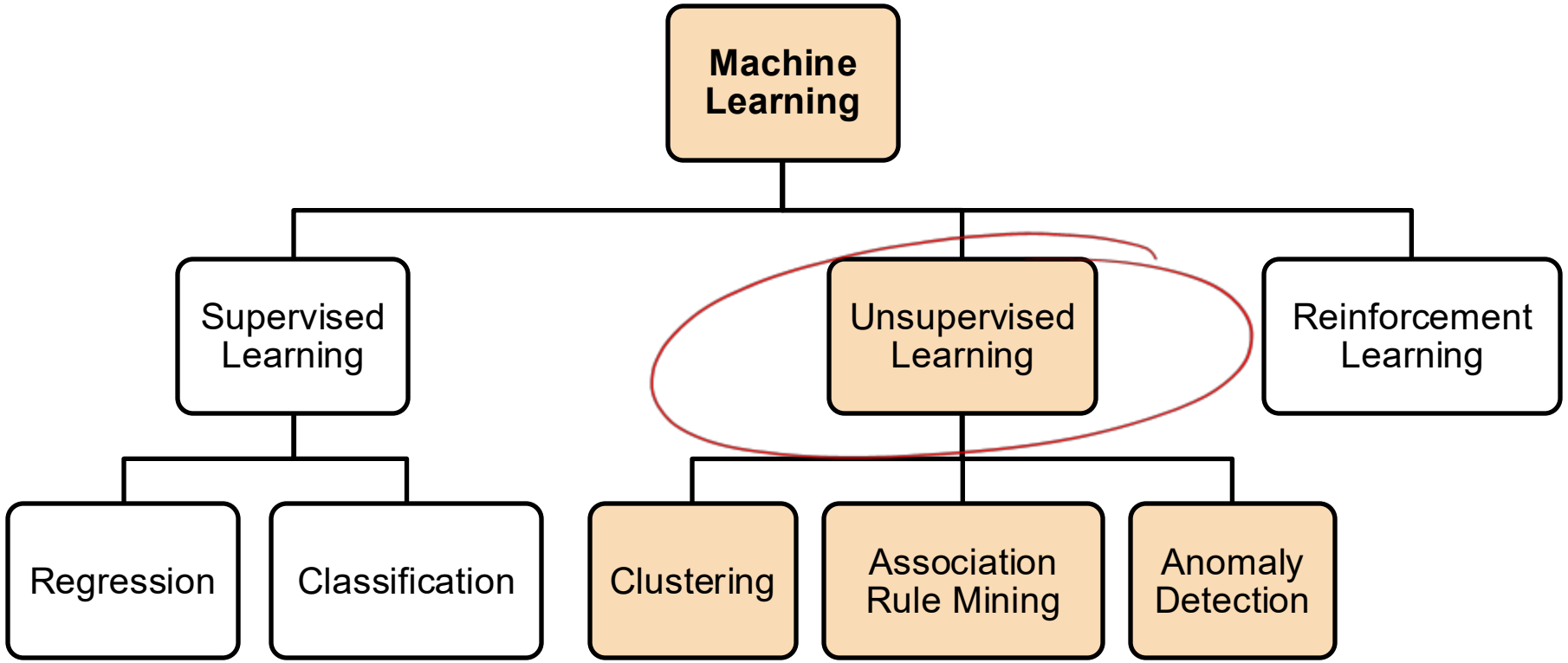



D | BUSINESS | SCHOOL

# Machine Learning Tasks



# Unsupervised Learning

- Predict patterns/groupings of data given set of *unlabeled* observations

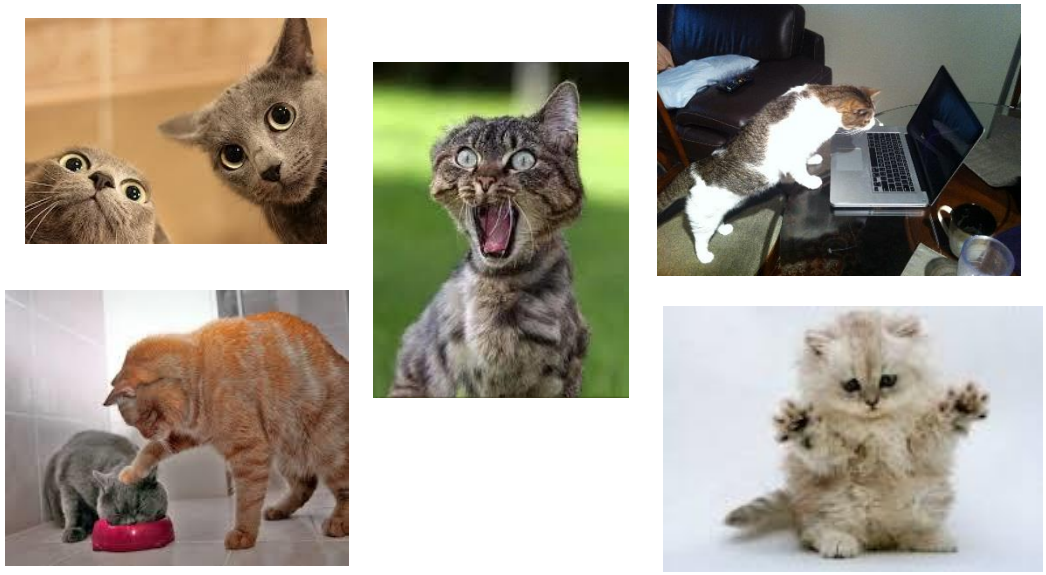







# Unsupervised Learning

- Predict patterns/groupings of data given set of *unlabeled* observations

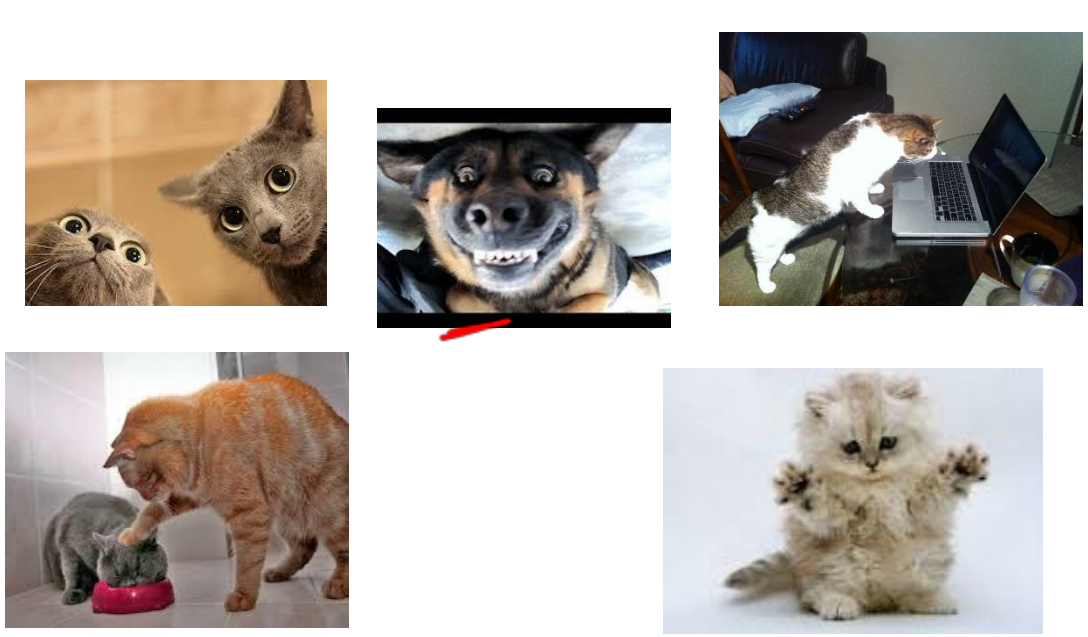


**cats**

**dogs**

# Unsupervised Learning

- Predict patterns/groupings of data given set of *unlabeled* observations

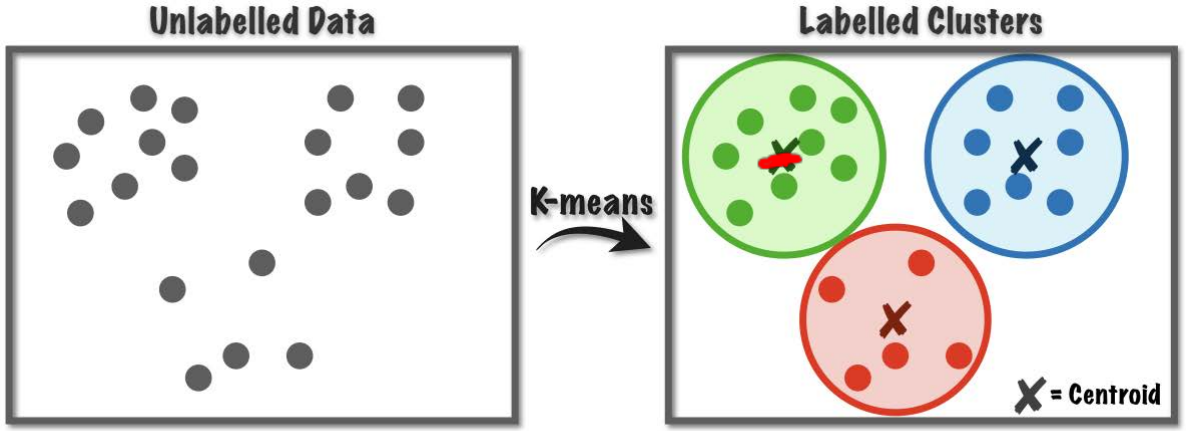


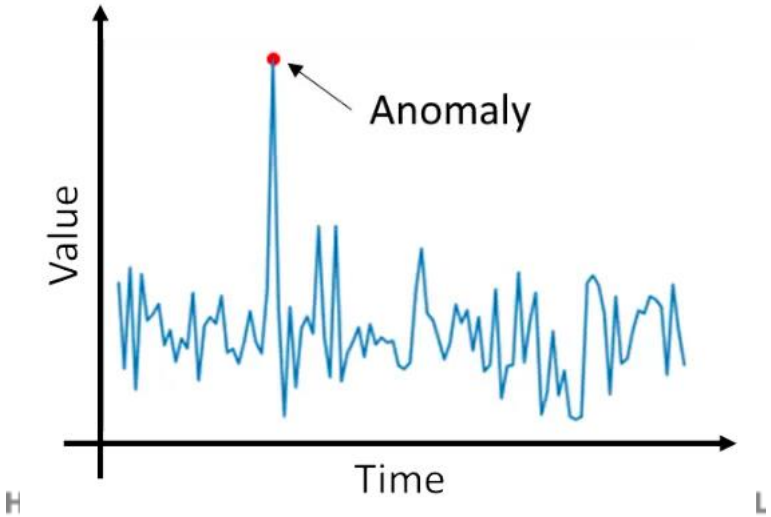
**indoor**

**outdoor**

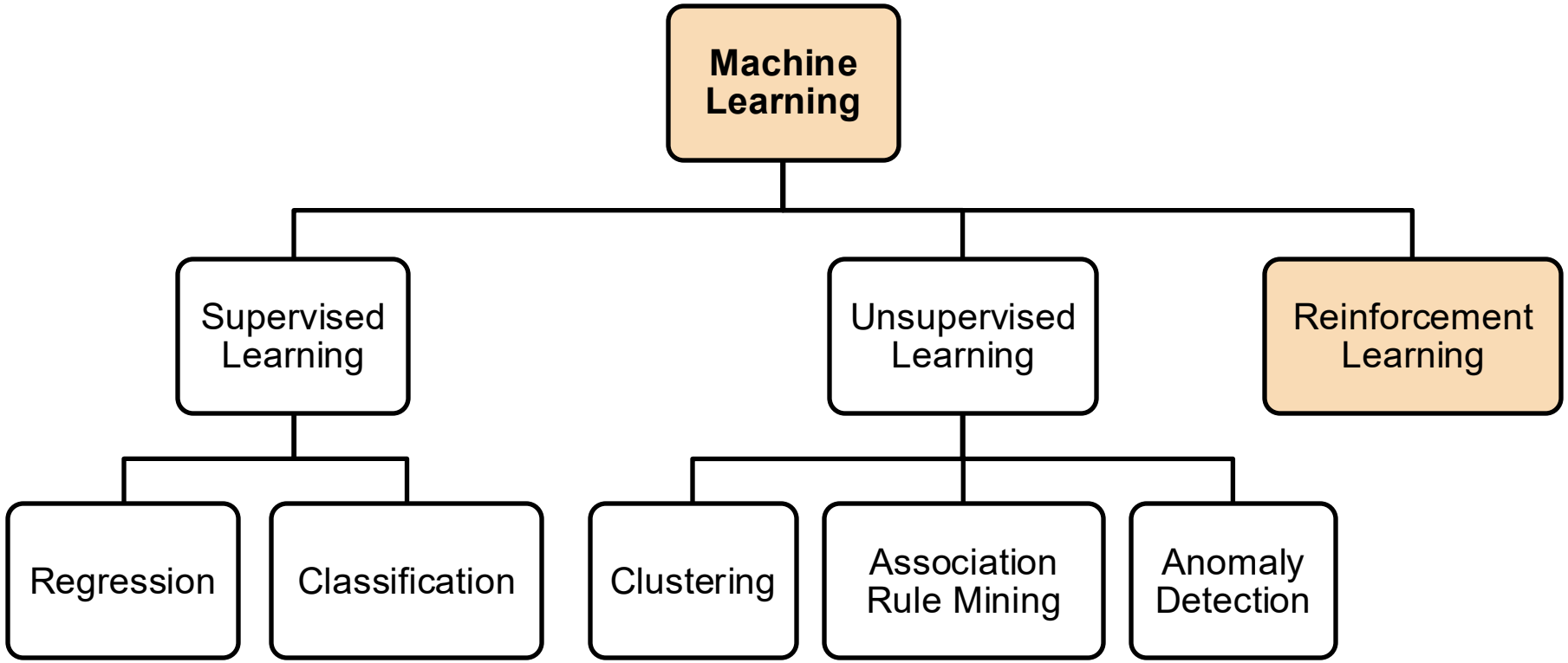
# Machines (i.e., methods)

- Clustering
  - **K-Means**
  - Hierarchical
- Association Rule Mining
  - Apriori
- Anomaly Detection
  - Highly dependent on type of anomaly and data
  - Often based on other machine learning techniques



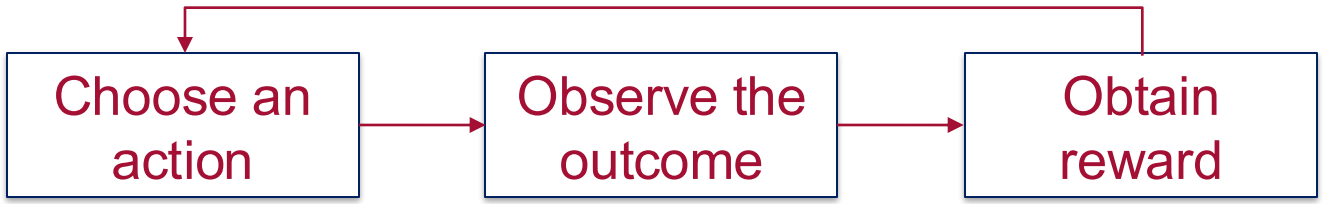



# Machine Learning Tasks



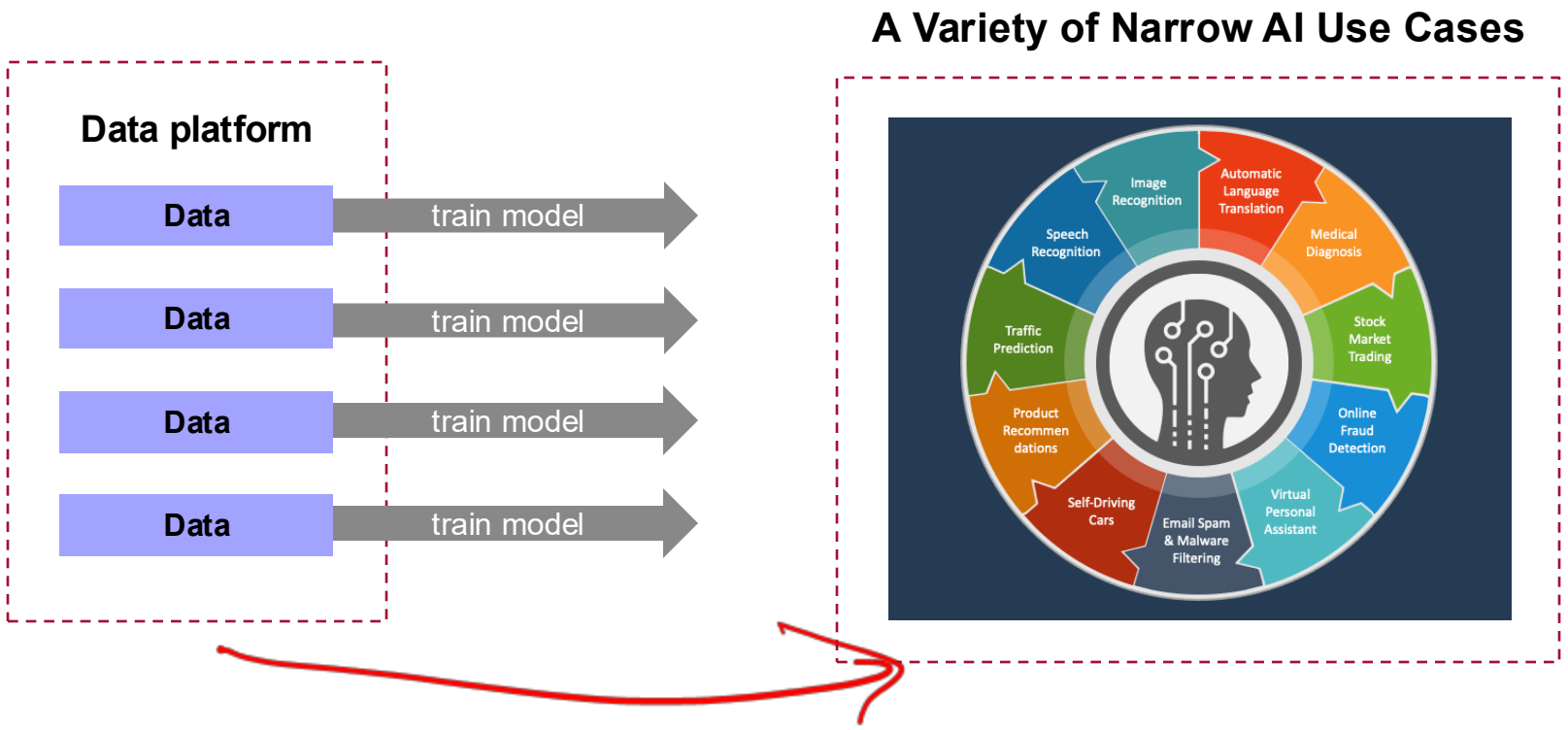
# Reinforcement Learning

- In a sequential fashion, i.e., for each time period…




- Unknown (and noisy) environment, i.e. you don't know the outcome for each action
- Example: Self driving cars, Netflix recommendation, etc.
- Goal: Maximize cumulative reward
- Exploration vs. Exploitation tradeoff:
  - Exploration - take an action to learn more about the environment
  - Exploitation - take an action to maximize expected reward

# Traditional “narrow" AI causes the proliferation of basic, single-purpose, recurrence-based models



# AI Innovation Has Progressed In Waves

Time

## Wave 1: Cybernetics (1950s-70s)

- Field of AI research was founded and key **concepts were first described, including early neural networks**

## Wave 2: Trained experts (1980s-90s)

- Expert systems were developed to **solve problems about a specific domain of knowledge**

## Wave 3: Good learners (2000s-)

- Progress in artificial neural networks, “deep” learning, CNNs, RNNs led to **notable achievements in text, speech, language and vision processing**

## Wave 4: Generative systems (2020s ⟶)

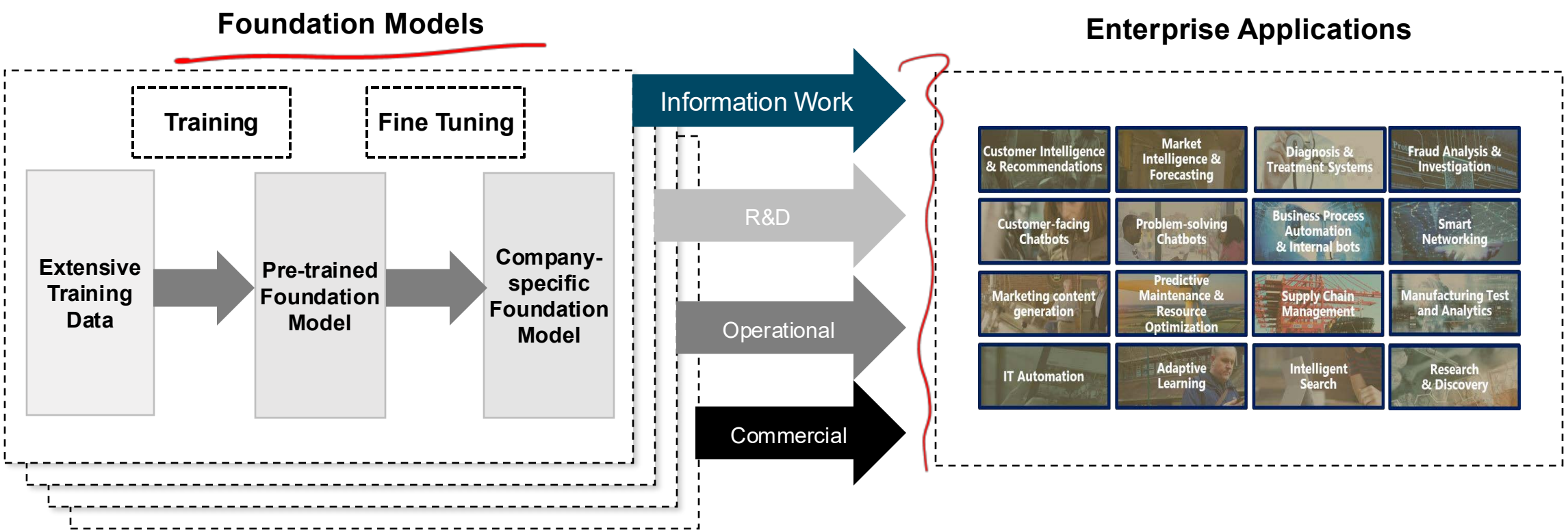
- These systems can adapt contextually, **learn and function with minimal supervision, and massively augment human intelligence**

cat

dog


Harvard Business School

# These innovations are enabling highly capable, pre-trained Foundation Models



# Two fundamentally different approaches to operational AI

Traditional Machine Learning


“Overtrained”, Generative, Attention based AI

# Today's Session: Data Analysis in Julius

Focus on data & ML

1. Perform EDA to understand the data and the cloud cost growth problem
2. Use Clustering Analysis to uncover systematic patterns
3. Deep dive into department differences
4. Build a predictive model to act as an early warning system
5. Develop a recovery plan

Harvard Business School

DRAFT
JULY 29, 2025

IAVOR BOJINOV

## Cloud Cost Optimization: A Hands-on Generative AI Exercise

### Introduction

You've just stepped in as the new Chief Technology Officer (CTO) at FlavorFusion Foods, a consumer-packaged-goods company that has grown at a dizzying pace. Unfortunately, its cloud infrastructure costs have grown even faster, skyrocketing from $380,000 to a staggering $2.4 million per month in just six months. The CFO is alarmed, and the board is demanding answers.

Your mission is to use generative AI as your trusted analyst to investigate this crisis. You will need to analyze six months of cloud cost data, identify the key drivers of the cost explosion, and build an actionable recovery plan to present to the executive team. To help you, you can leverage a generative AI tool with data analysis capabilities.

*Company Context*

- **FlavorFusion Foods:** A rapidly growing CPG company that scaled from a regional startup to a national brand in 18 months.
- **Infrastructure:** The company operates a complex environment including a high-traffic e-commerce platform, a sprawling supply chain system, multiple data analytics platforms, and core corporate systems (ERP, marketing automation, etc.).
- **The Challenge:** Cloud spend has reached an unsustainable $2.4 million per month, with non-production environments shockingly accounting for 62% of the total.

*Task 1: Understand the Crisis with Exploratory Data Analysis*

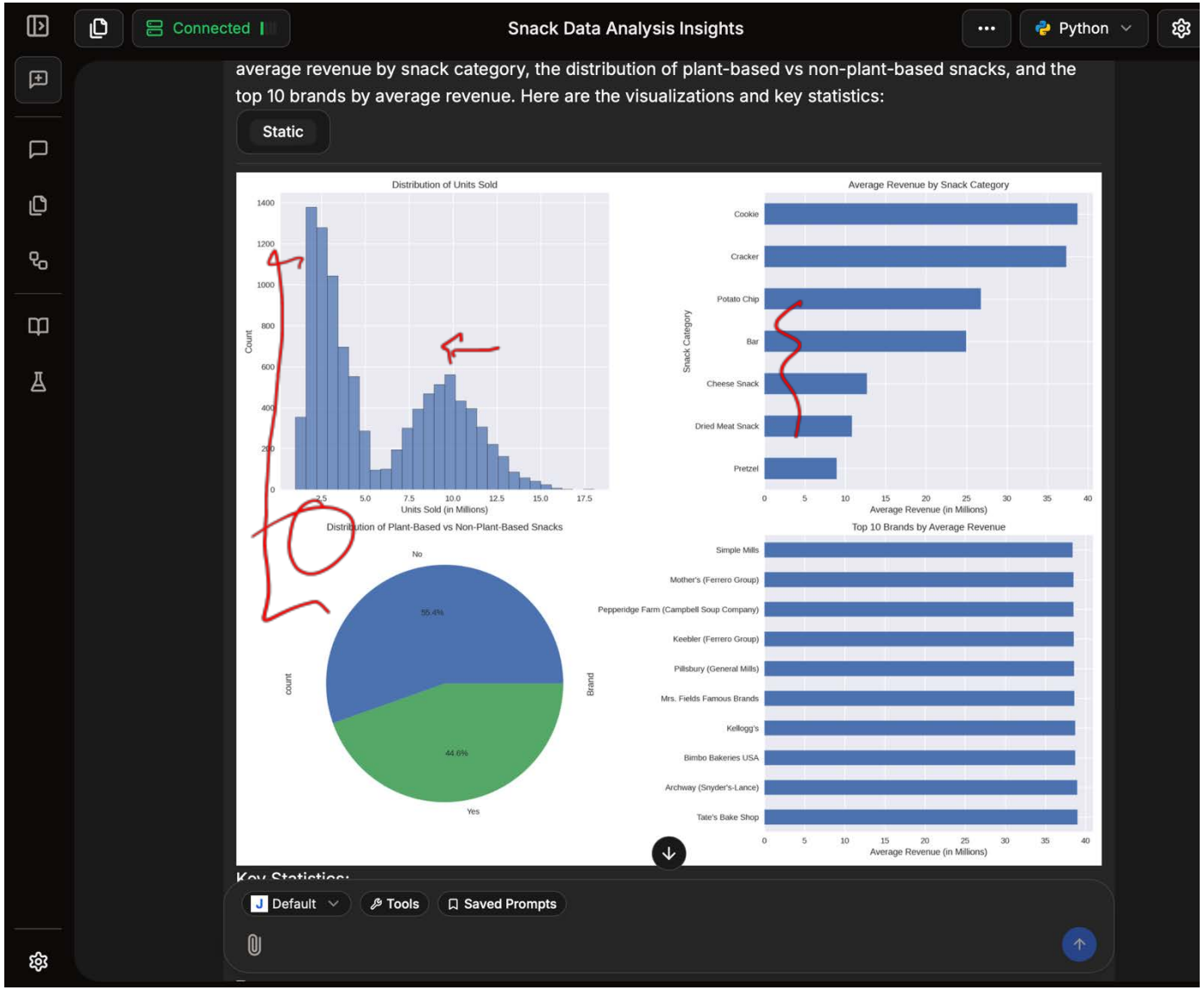
First, you need to get a handle on the data and confirm the scale of the problem. Exploratory Data Analysis (EDA) is like being a detective with a new case file. You start by summarizing the evidence, looking for initial clues, and getting a feel for the situation before you identify your main suspects.

1. Begin by loading the provided `cloud_costs_daily.csv` data file into your AI assistant.
2. Ask your AI assistant to perform an initial analysis to understand the data's structure and check for any quality issues.
3. Work with your assistant to visualize the spending trends to see the problem for yourself.

Professor Iavor Bojinov prepared this exercise as the basis for class discussion rather than to illustrate either effective or ineffective handling of an administrative situation.

# Julius AI

- Build specifically for data analysis
- Multi-model (but mostly ChatGPT)
- Pre-defined workflows
- Editable code
- Explains code
- Stable






Harvard Business School

julius.ai

Mail - Bojinov, Iav... | Concur | Harvard... | OneLogin | Quick Links - myH... | Dashboard | Wikipedia | Check Harvard Lib... | Inbox – iav.bojinov... | Dashboard

Julius

Product | Use Cases | Resources | Security | Community | Pricing | Log In | Sign Up

Register August 12: Notebooks deep dive webinar >

# The AI data analyst that works for you

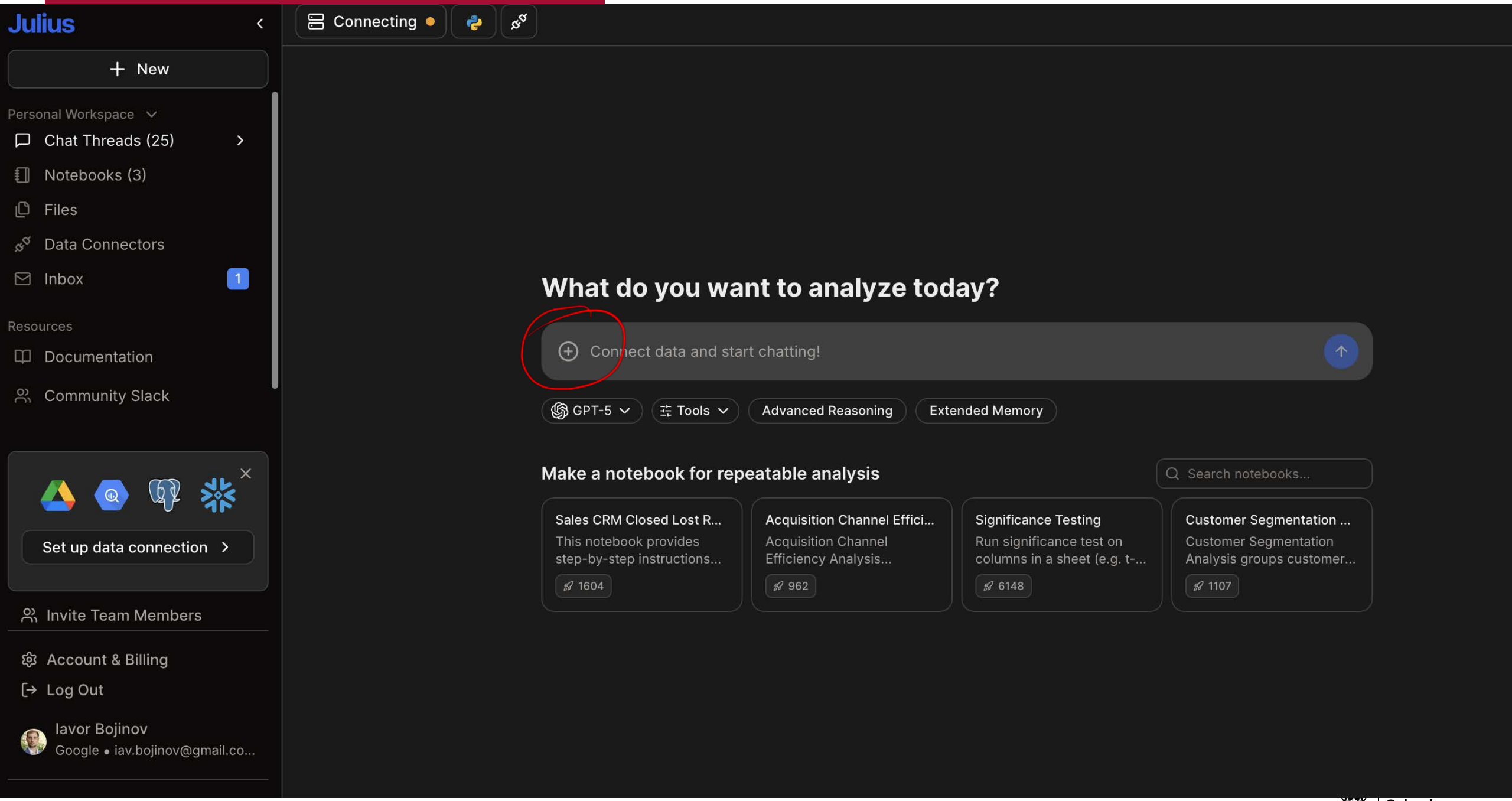
Connect your data, ask questions in plain English, and get insights in seconds. No coding required.

Try Julius free >

Loved by 2,000,000+ users and trusted by individuals at

PRINCETON UNIVERSITY BCG zapier

## Intuitive and easy to use



# How to work together for the next 30 minutes

1. Work as a group of 3 people (pre-assigned) to produce a single analysis & recommendation.
2. Everyone else should work in parallel using similar prompts.
3. You can also distribute tasks, but if you switch "chats," the context will be lost.
4. Try to push Julius & feel free to use other tools (like M365 Copilot Analyst Agent & ChatGPT)

Harvard Business School

DRAFT
JULY 29, 2025

IAVOR BOJINOV

## Cloud Cost Optimization: A Hands-on Generative AI Exercise

### Introduction

You've just stepped in as the new Chief Technology Officer (CTO) at FlavorFusion Foods, a consumer-packaged-goods company that has grown at a dizzying pace. Unfortunately, its cloud infrastructure costs have grown even faster, skyrocketing from $380,000 to a staggering $2.4 million per month in just six months. The CFO is alarmed, and the board is demanding answers.

Your mission is to use generative AI as your trusted analyst to investigate this crisis. You will need to analyze six months of cloud cost data, identify the key drivers of the cost explosion, and build an actionable recovery plan to present to the executive team. To help you, you can leverage a generative AI tool with data analysis capabilities.

*Company Context*

- **FlavorFusion Foods:** A rapidly growing CPG company that scaled from a regional startup to a national brand in 18 months.
- **Infrastructure:** The company operates a complex environment including a high-traffic e-commerce platform, a sprawling supply chain system, multiple data analytics platforms, and core corporate systems (ERP, marketing automation, etc.).
- **The Challenge:** Cloud spend has reached an unsustainable $2.4 million per month, with non-production environments shockingly accounting for 62% of the total.

*Task 1: Understand the Crisis with Exploratory Data Analysis*

First, you need to get a handle on the data and confirm the scale of the problem. Exploratory Data Analysis (EDA) is like being a detective with a new case file. You start by summarizing the evidence, looking for initial clues, and getting a feel for the situation before you identify your main suspects.

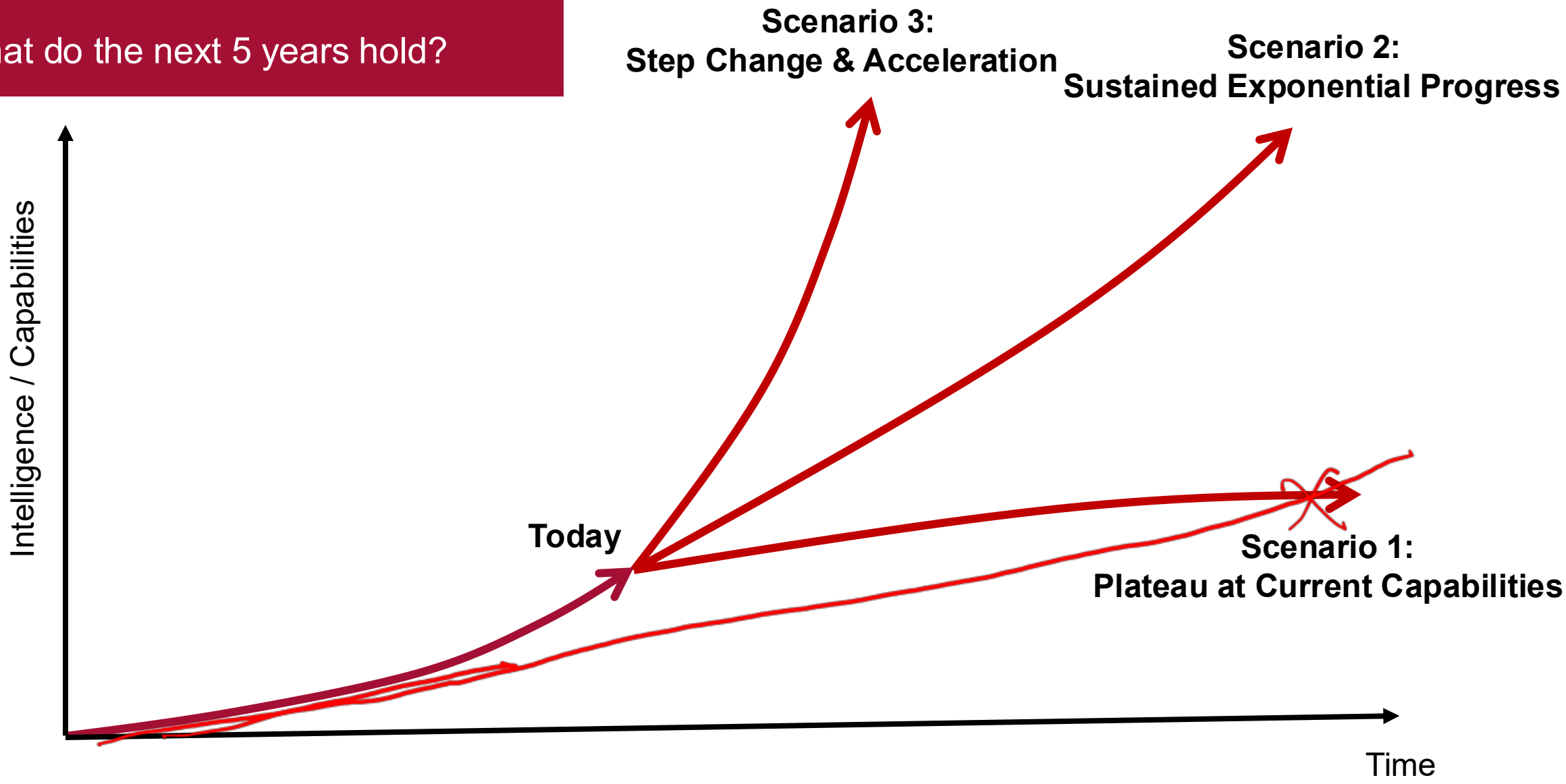
1. Begin by loading the provided `cloud_costs_daily.csv` data file into your AI assistant.
2. Ask your AI assistant to perform an initial analysis to understand the data's structure and check for any quality issues.
3. Work with your assistant to visualize the spending trends to see the problem for yourself.

Professor Iavor Bojinov prepared this exercise as the basis for class discussion rather than to illustrate either effective or ineffective handling of an administrative situation.

# What are your key takeaways from the exercise?

- Stochastic ⇝ small differences
- Next prompt.
- Data democratization
- Data quality!!
- Automation & proactive
- EDA for "free"
- Need context!

# What do the next 5 years hold?



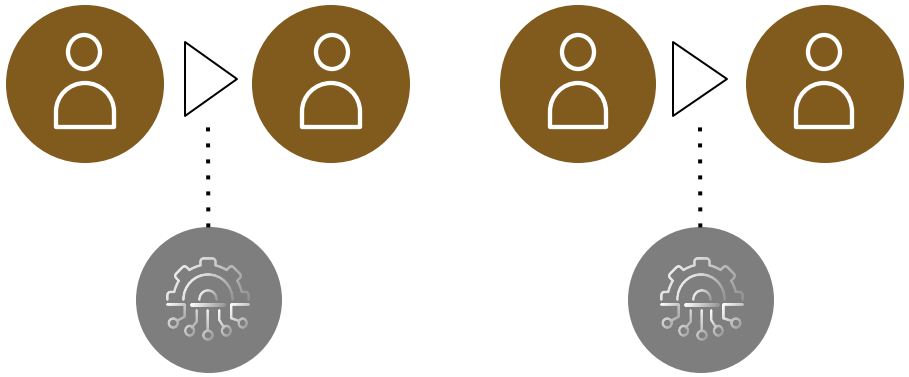



Harvard Business School

# Agents transform AI from a tool to a teammate.



Harvard Business School

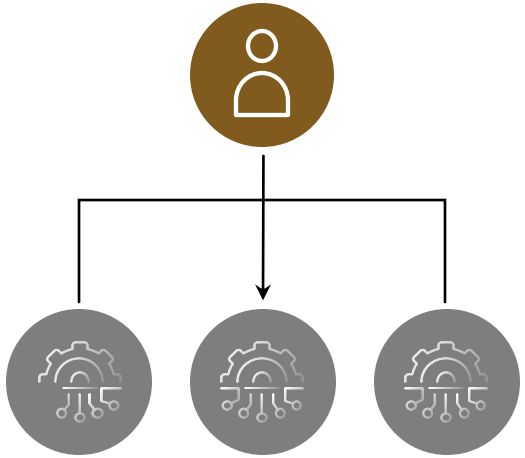


## Process Redesign:

Reimagining workflows end-to-end for and with AI

## Value Creation

Faster time to value
Improved quality of outcomes



## Producer Mindset:

Autonomous agents empower every individual and team.

## Value Creation

Increased scalability
Significant cost reductions

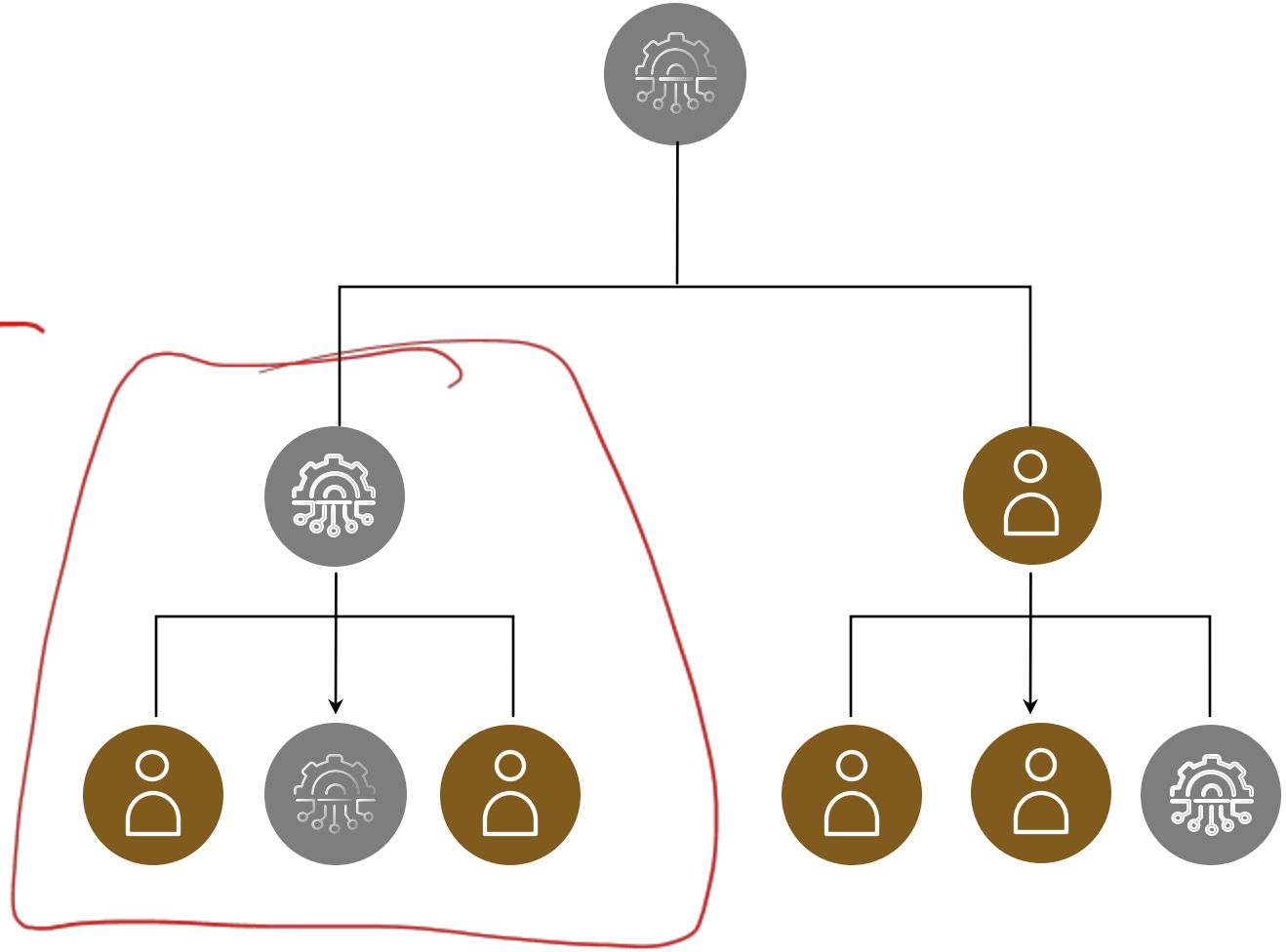
# Agentic AI Poised to Impact Organizations & Work

## Future Teams

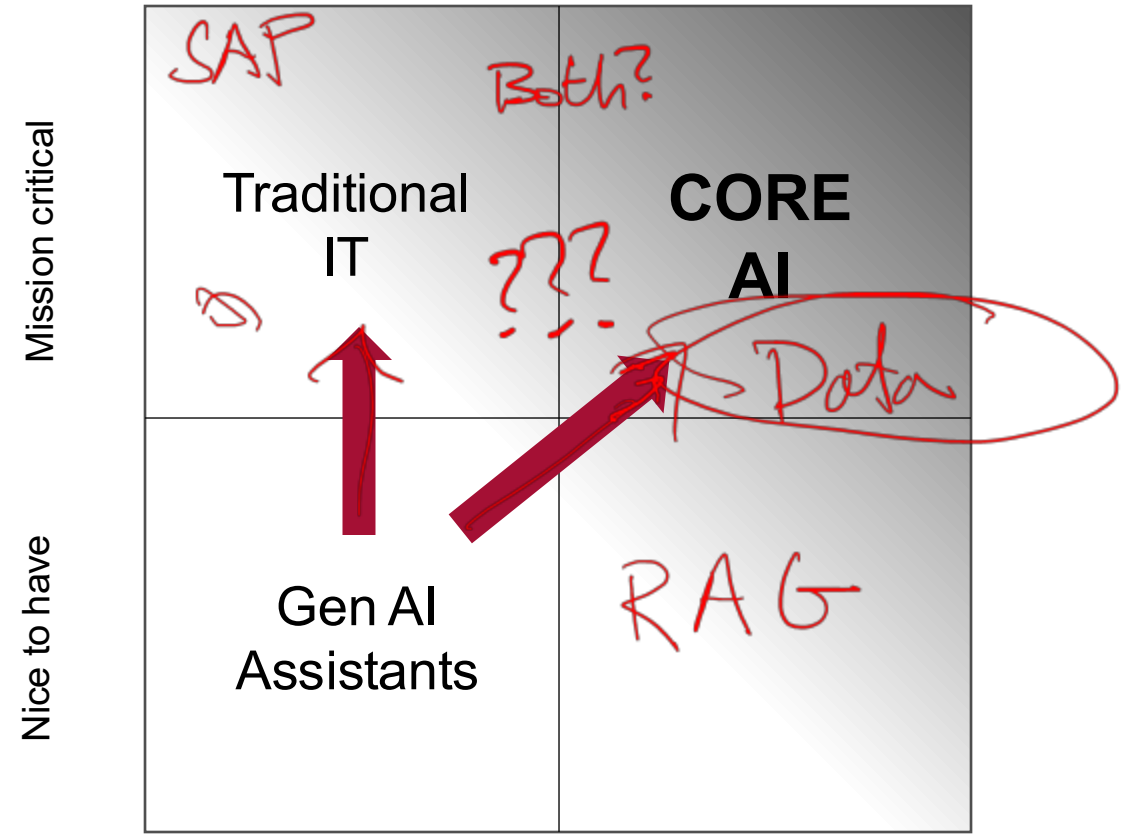
Mix of Human and Agent-Led Teams

## Value Creation

Optimized value creation, capture & delivery



# Can AI make a strategic difference to the firm?



# Thank you!