

DEEL 12

CUSTOM vs OFF- THE-SHELF AI



Custom vs Off-The-Shelf AI

- Introduction & examples
- Approaches to AI Solutions
- What are you buying with off-the-shelf?
- Conclusions
- Exercise

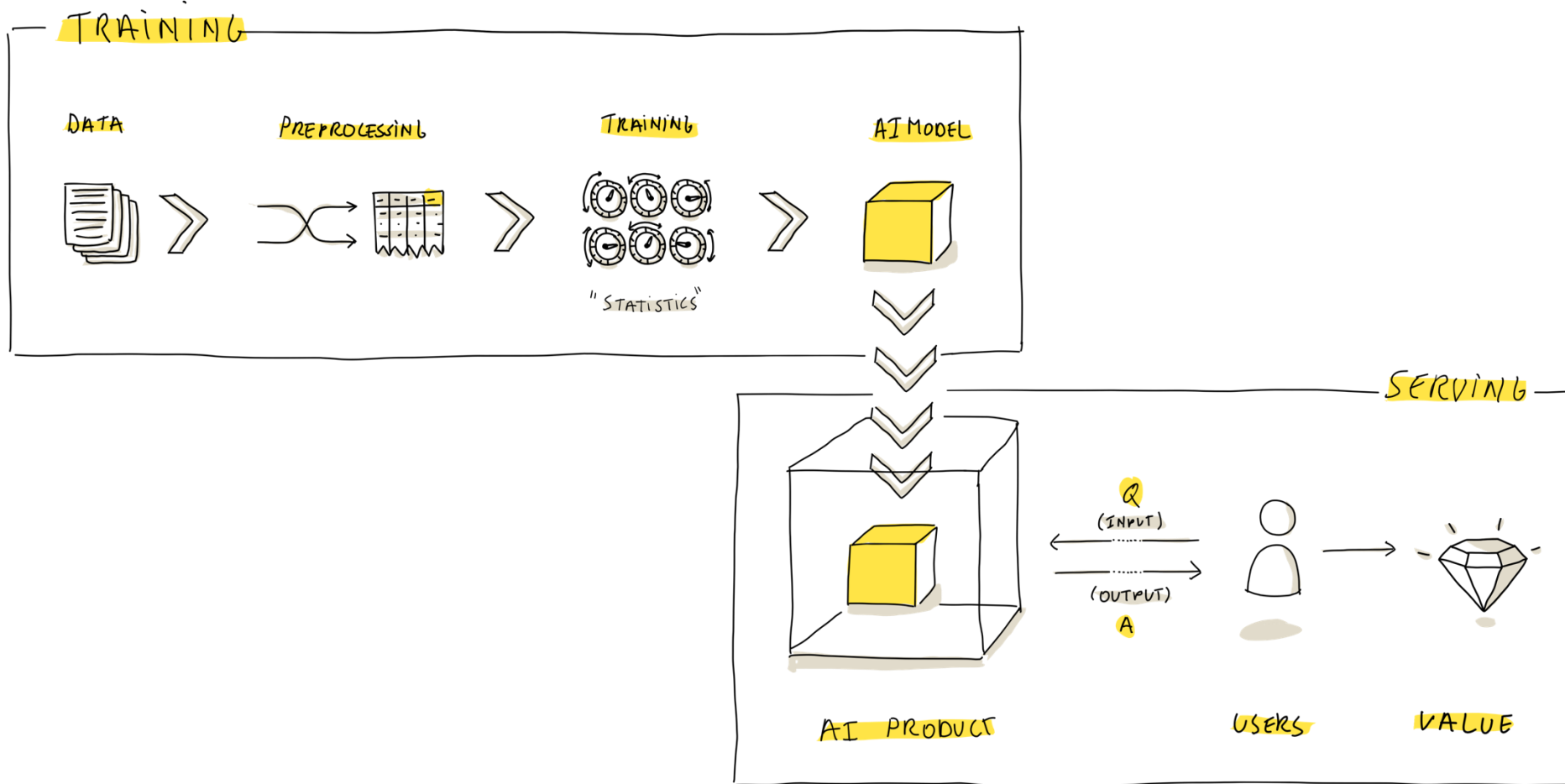


Custom vs Off-The-Shelf AI

- **Introduction & examples**
- Approaches to AI Solutions
- What are you buying with off-the-shelf?
- Conclusions
- Exercise

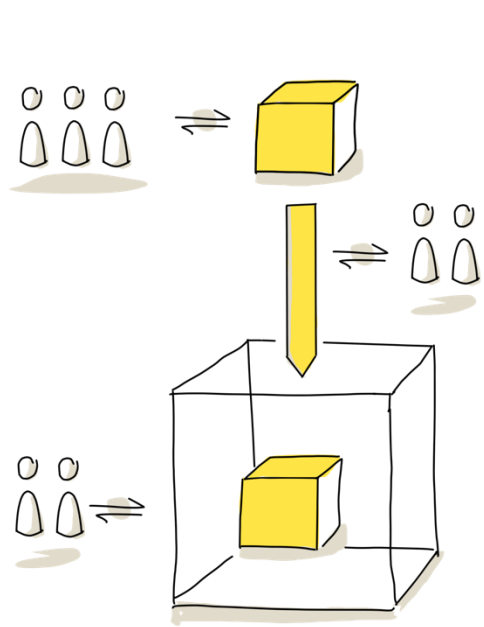


Recap: AI Product

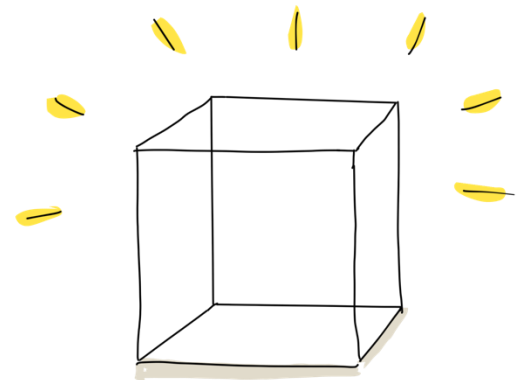
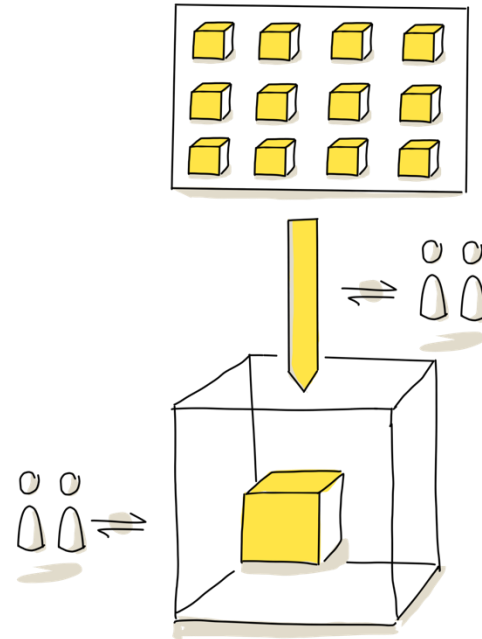
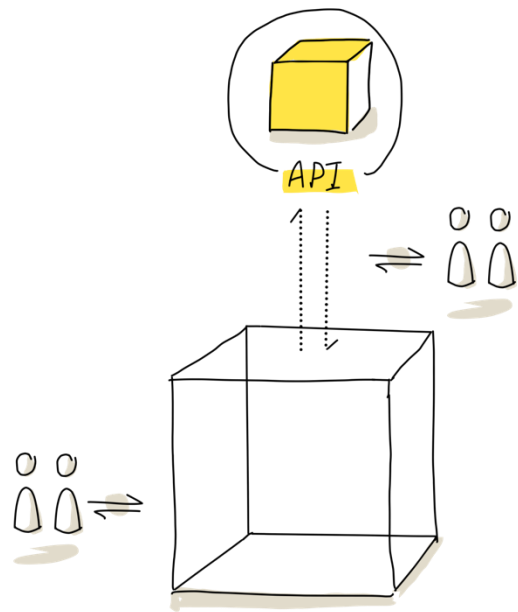




Recap: AI Product



CUSTOM AI



OFF-THE-SHELF AI

← MAKE

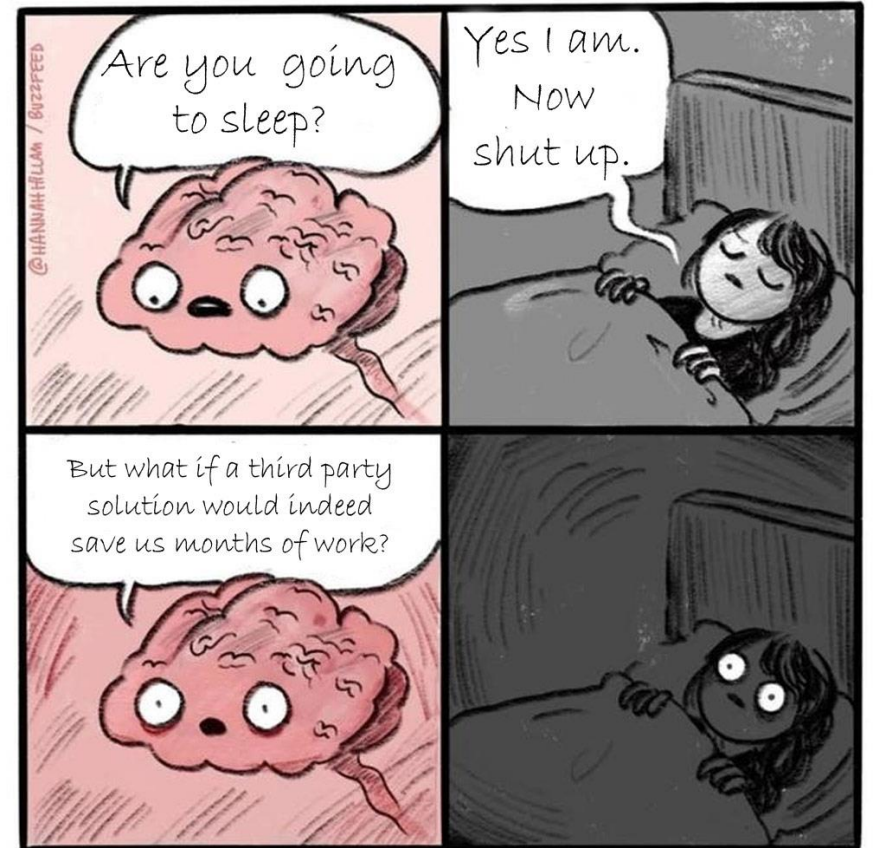
BUY →

This Part

Don't Reinvent



Perfect It





This Lesson

- This lesson is about: no need to do it all yourself
 - We will discuss different approaches later
- Services economy: triggered through cloud (IaaS/PaaS/SaaS)
- For more advanced models: option to do it yourself disappearing – role of hyperscalers



Example: Off-the-shelf AI Jesus

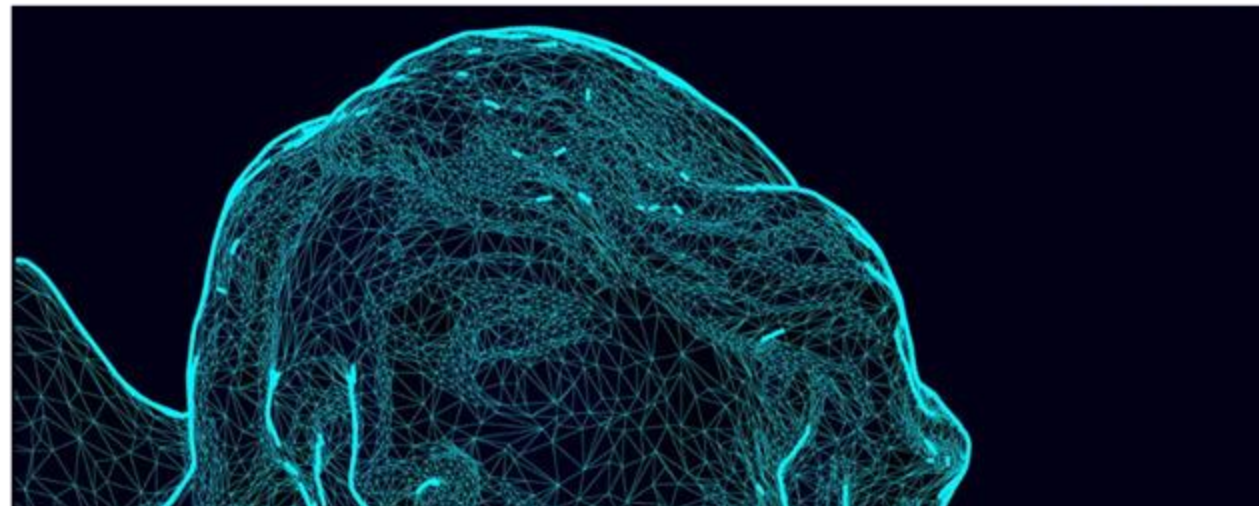


Tech

'AI Jesus' takes confessions at Swiss church

Some churchgoers describe a spiritual experience within the confessional booth, though others labelled it a gimmick

Anthony Cuthbertson • Monday 25 November 2024 15:43 GMT • [23](#) Comments







Example: Off-the-shelf GIS Tooling



Without FlyPix



With FlyPix



Example: Off-the-shelf Legal Docs Analysis

The screenshot displays the LegalFly interface with a dark theme. On the left is a sidebar with navigation options: Dashboard, Agents, Review (selected), Drafting, Discovery, Compare, Translate, Management, Documents, and Playbooks. The main content area shows a document titled "Software Licence Agreement" with a "Time saved: 2 hours 38 mins" indicator. The document text includes: "THIS SOFTWARE LICENCE AGREEMENT (the 'Agreement') dated this 1st day of September, 2023 (the 'Execution Date') BETWEEN: [First Name] [Last Name] of [Company Name] (the 'Vendor') & [First Name] [Last Name] of [Company Name] (the 'Licensee')". Below this is a "Background" section stating: "The Vendor wishes to licence computer software to the Licensee and the Licensee desires to purchase the software licence under the terms and conditions stated below. IN CONSIDERATION OF the provisions contained in this Agreement and for other good and valuable consideration, the receipt and sufficiency of which is acknowledged, the parties agree as follows: Licence 1. Under this Agreement the Vendor grants to the Licensee a non-exclusive and non-transferable licence (the 'Licence') to use CyberSync (the 'Software')." The right-hand panel shows a "Review" tab with a "Playbook 6 items" section. Two items are visible: "Clear terms on data collection, usage, and protection" and "Specifics of licensing rights and restrictions". Each item has a green checkmark, a description, and buttons for "Explain" and "Re-draft". A third item, "Is there a money back guarantee?", is partially visible at the bottom.

LegalFly

Time saved: 2 hours 38 mins

Anonymisation Feedback

Copilot Review Anonymisation

Dashboard

Agents

Review

Drafting

Discovery

Compare

Translate

Management

Documents

Playbooks

Software Licence Agreement

THIS SOFTWARE LICENCE AGREEMENT (the "Agreement") dated this 1st day of September, 2023 (the "Execution Date")

BETWEEN: [First Name] [Last Name] of [Company Name] (the "Vendor")

&

[First Name] [Last Name] of [Company Name] (the "Licensee")

Background

The Vendor wishes to licence computer software to the Licensee and the Licensee desires to purchase the software licence under the terms and conditions stated below.

IN CONSIDERATION OF the provisions contained in this Agreement and for other good and valuable consideration, the receipt and sufficiency of which is acknowledged, the parties agree as follows:

Licence 1. Under this Agreement the Vendor grants to the Licensee a non-exclusive and non-transferable licence (the "Licence") to use CyberSync (the "Software").

Licence

Playbook 6 items

- Clear terms on data collection, usage, and protection**
The agreement clearly outlines the terms of data collection, usage, and protection, specifying that Customer Data will not be used to train the SaaS Solution without Customer's consent.
Explain Re-draft
- Specifics of licensing rights and restrictions**
The agreement provides specific details on licensing rights and restrictions, granting the Customer and its Affiliates a limited, revocable, non-exclusive, worldwide right to access and use the SaaS Solution.
Explain Re-draft
- Is there a money back guarantee?**



Example: Off-the-shelf Planning / Routing



Planning models Developers ▾ Resources ▾ Pricing

Talk to us

Planning Models



Field Service Routing

Reduce fuel costs and boost Field Team productivity with Timefold's Planning AI. Automate complex scheduling to save...

[Read more](#)



Employee Shift Scheduling

Timefold's employee shift scheduling API platform automates planning so you can schedule shifts for thousands of employee...

[Read more](#)



Last Mile Delivery Routing

Timefold's last mile delivery routing Planning AI optimizes the final leg of the delivery process resulting in efficient delivery fleet...

[Read more](#)



Vehicle Routing (VRP)



Maintenance Scheduling



Machine Job Scheduling



Example: Off-the-shelf LLMs

LLaMA
by  **Meta**



Gemini for Google Cloud
AI Assistance

Vertex AI Gen AI Platform

Gemini Models *Ultra, Pro, Nano*

Ultra Scale AI Infrastructure





Custom vs Off-The-Shelf AI

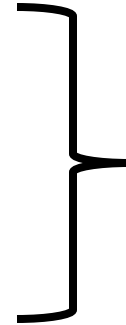
- Introduction & examples
- **Approaches to AI Solutions**
- What are you buying with off-the-shelf?
- Conclusions
- Exercise



1. Hiring Data Scientists

2. AutoML

3. “Off the Shelf” AI



Custom AI



Custom vs Off-The-Shelf AI

- Introduction & examples
- Approaches to AI Solutions
 1. Hiring Data Scientists
 2. AutoML
 3. Off-the-Shelf AI
- What are you buying with off-the-shelf?
- Conclusions
- Exercise

} Custom AI



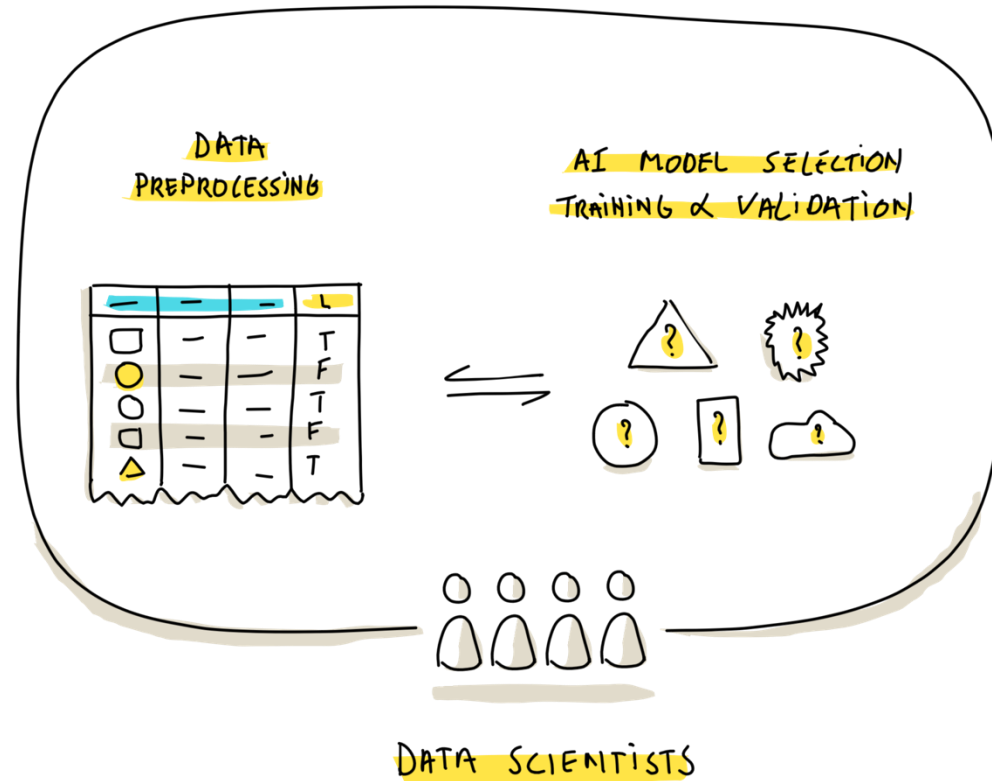
Custom vs Off-The-Shelf AI

- Introduction & examples
- Approaches to AI Solutions
 - 1. Hiring Data Scientists**
 2. AutoML
 3. Off-the-Shelf AI
- What are you buying with off-the-shelf?
- Conclusions
- Exercise

} Custom AI



Classic Approach: Hiring Data Scientists





5-10 Years ago: Data Scientists were THE key resources to run AI projects



Nieuwe opleiding moet tekort aan data scientists terugdringen

Data Big data Analytics Arbeidsmarkt Detachering & opleidingen



De Universiteit van Amsterdam, de Vrije Universiteit en de Hogeschool van Amsterdam slaan de handen ineen met de oprichting van de Amsterdam School of Data Science. De drie kennisinstellingen willen met het grootste opleidingsaanbod van Europa het nijpende tekort aan data scientists op de arbeidsmarkt terugdringen. Kajsa Ollongren, wethouder en locoburgemeester van Amsterdam, verzorgt op 24 maart de officiële opening van het nieuwe opleidingsplatform.



Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)



Andrew J Buboltz, silk screen on a page from a high school yearbook, 8.5" x 12", 2011 Tamar Cohen



Home Page - Select or create a notebook x Decision Tree Regression in Jupyter x +

localhost:8888/notebooks/Decision%20Tree%20Regression%20in%20Jupyter%20Notebook.ipynb

Apps Program to print al... SEND CODE Image result for eg... JSON Formatter &... Earth Engine Pytho... salib tutorial - Goo... Why is the file nam... Sensitivity Analysis... Reading list

jupyter Decision Tree Regression in Jupyter Notebook (autosaved) Python 3 Logout

File Edit View Insert Cell Kernel Widgets Help Notebook saved Trusted Python 3

Run Code

6	0.000041	0.018358	26.80	27.56	642.568	244.449
7	0.000044	0.018319	25.80	26.56	765.452	266.611
8	0.000041	0.018766	26.70	26.48	448.680	269.184
9	0.000037	0.015654	27.00	27.04	201.872	248.631

In [58]: `# Splitting the dataset into training and testing dataset`
`from sklearn.model_selection import train_test_split`
`# Splitting the dataset`
`X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20)`

In [59]: `print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)`
`(192, 6) (48, 6) (192,) (48,)`

In []:

Type here to search Desktop 23°C Cloudy 2:56 AM 12/26/2021



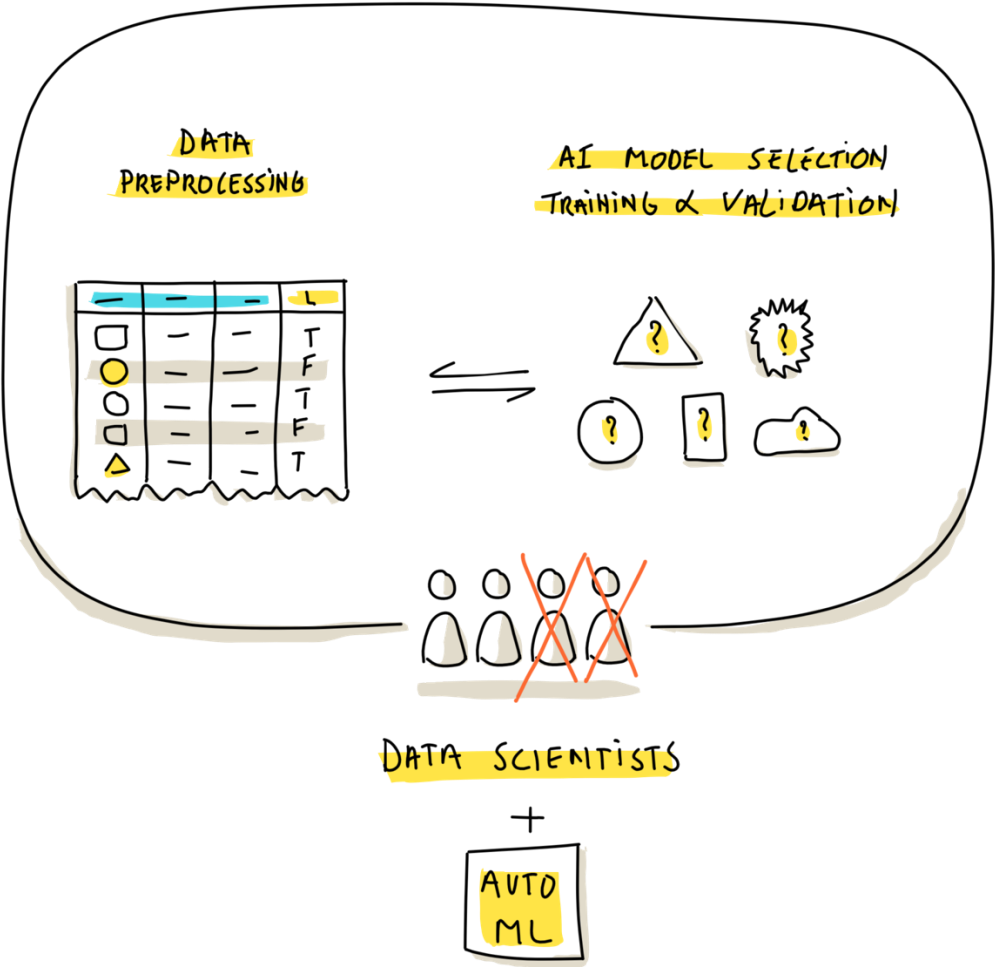
Custom vs Off-The-Shelf AI

- Introduction & examples
- Approaches to AI Solutions
 1. Hiring Data Scientists
 2. **AutoML**
 3. Off-the-Shelf AI
- What are you buying with off-the-shelf?
- Conclusions
- Exercise

} Custom AI

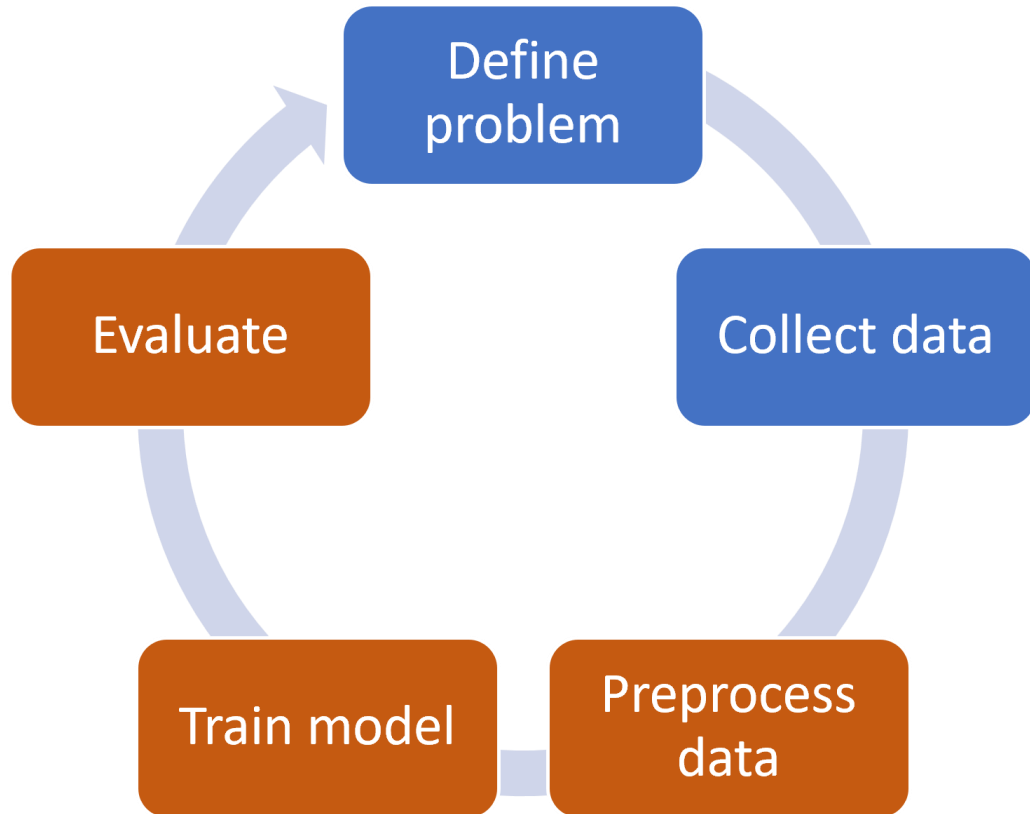


AutoML

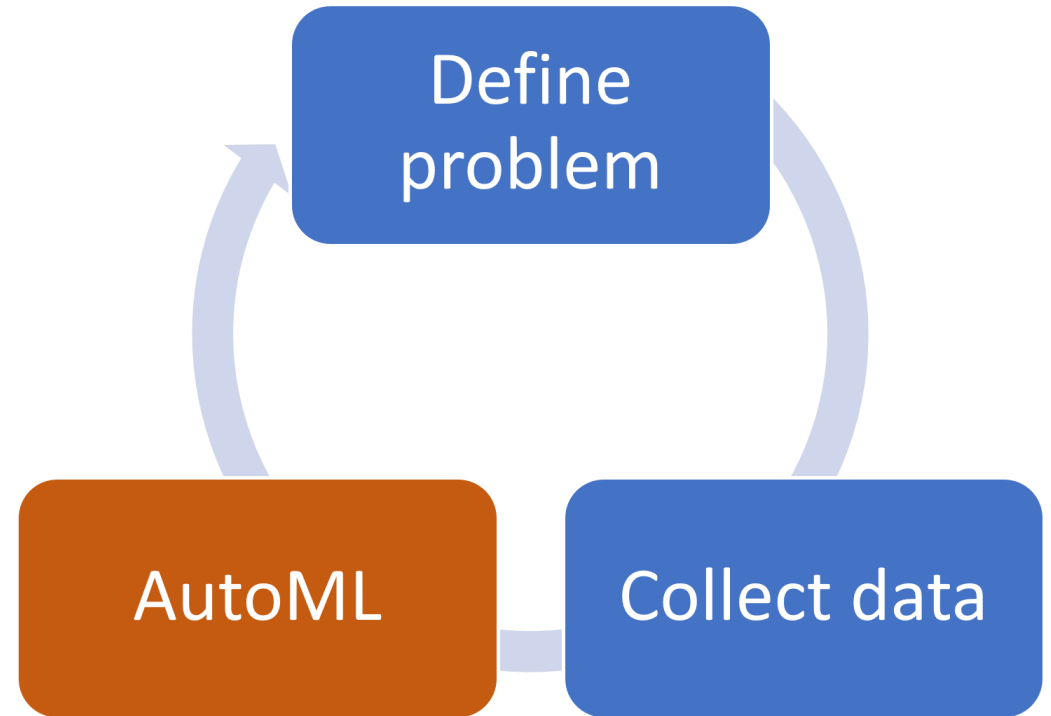


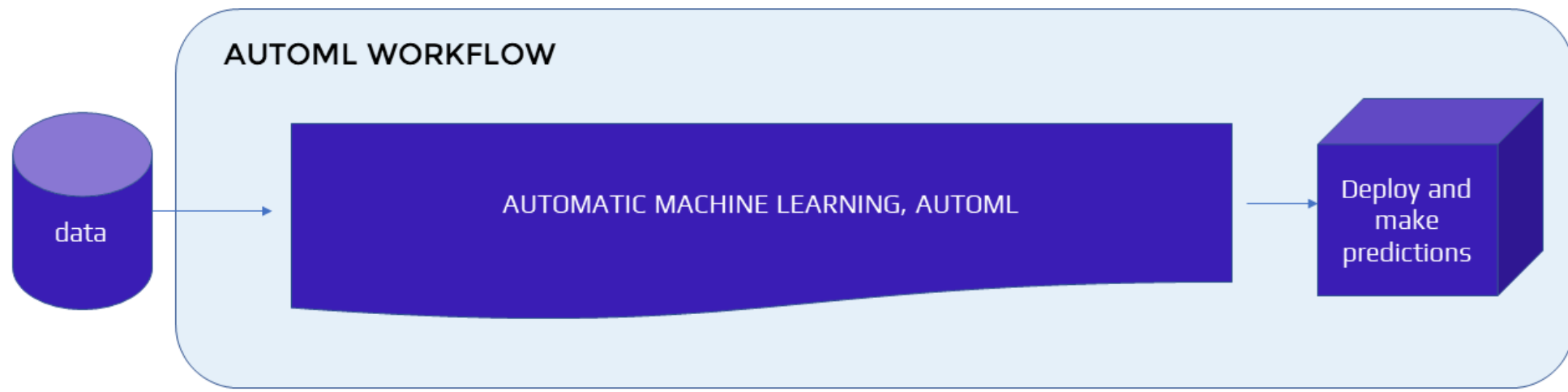
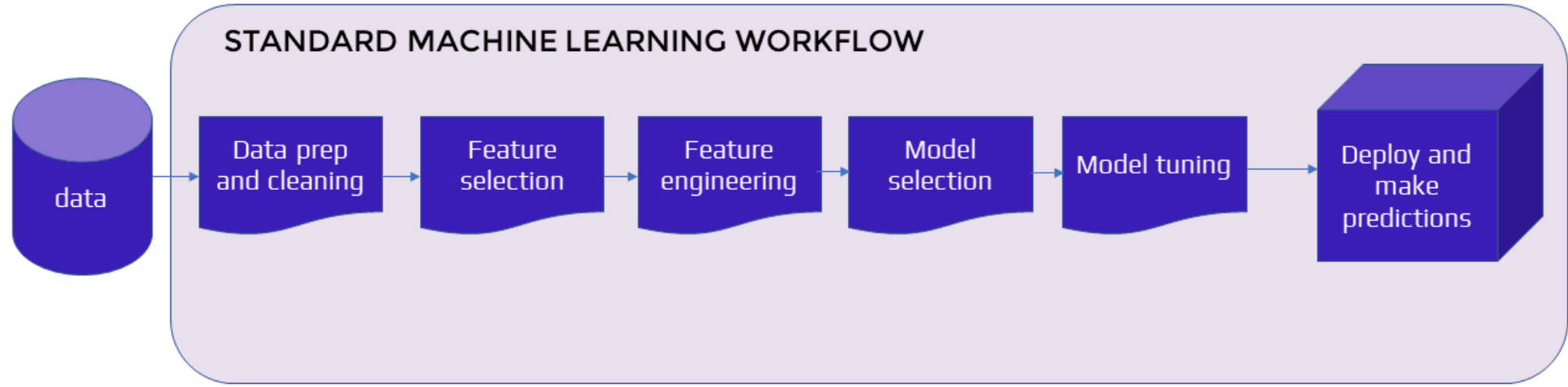


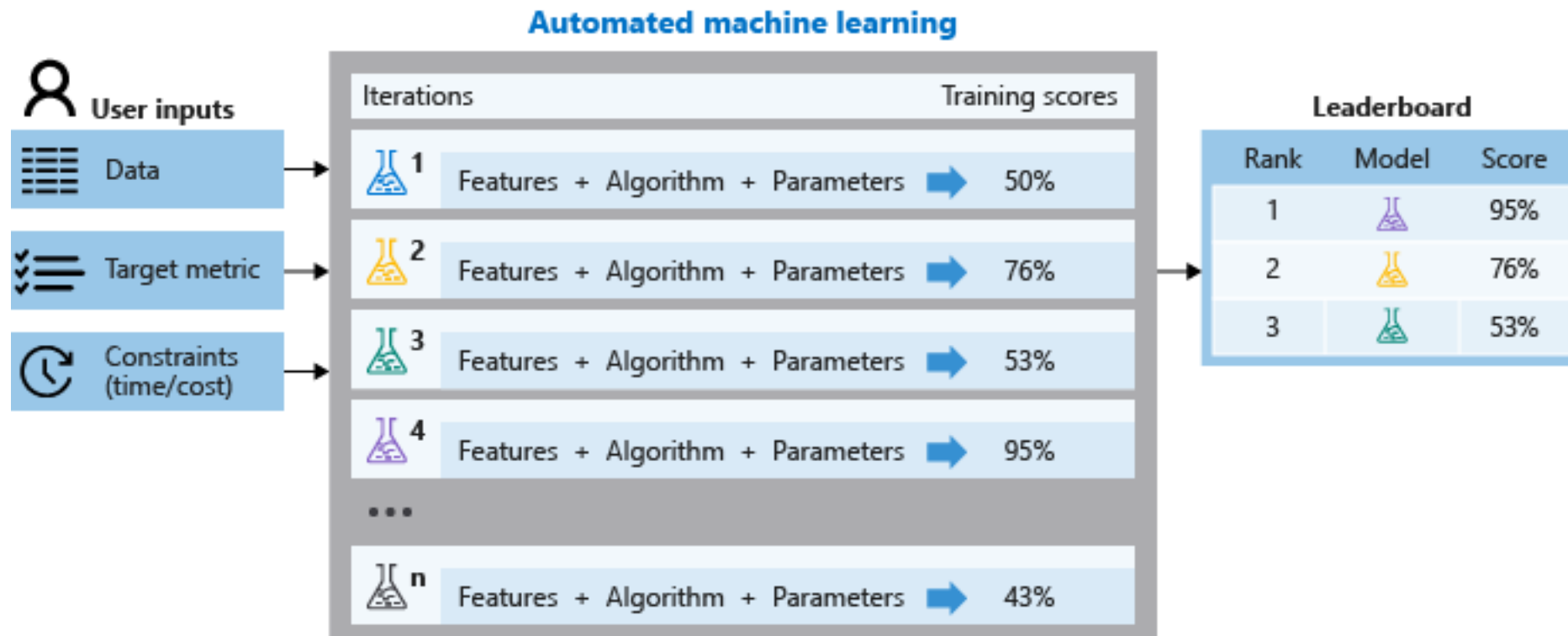
Traditional ML training workflow



AutoML workflow









Custom vs Off-The-Shelf AI

- Introduction & examples
- Approaches to AI Solutions
 1. Hiring Data Scientists
 2. AutoML
 3. Off-the-Shelf AI
- What are you buying with off-the-shelf?
- Conclusions
- Exercise

Custom AI

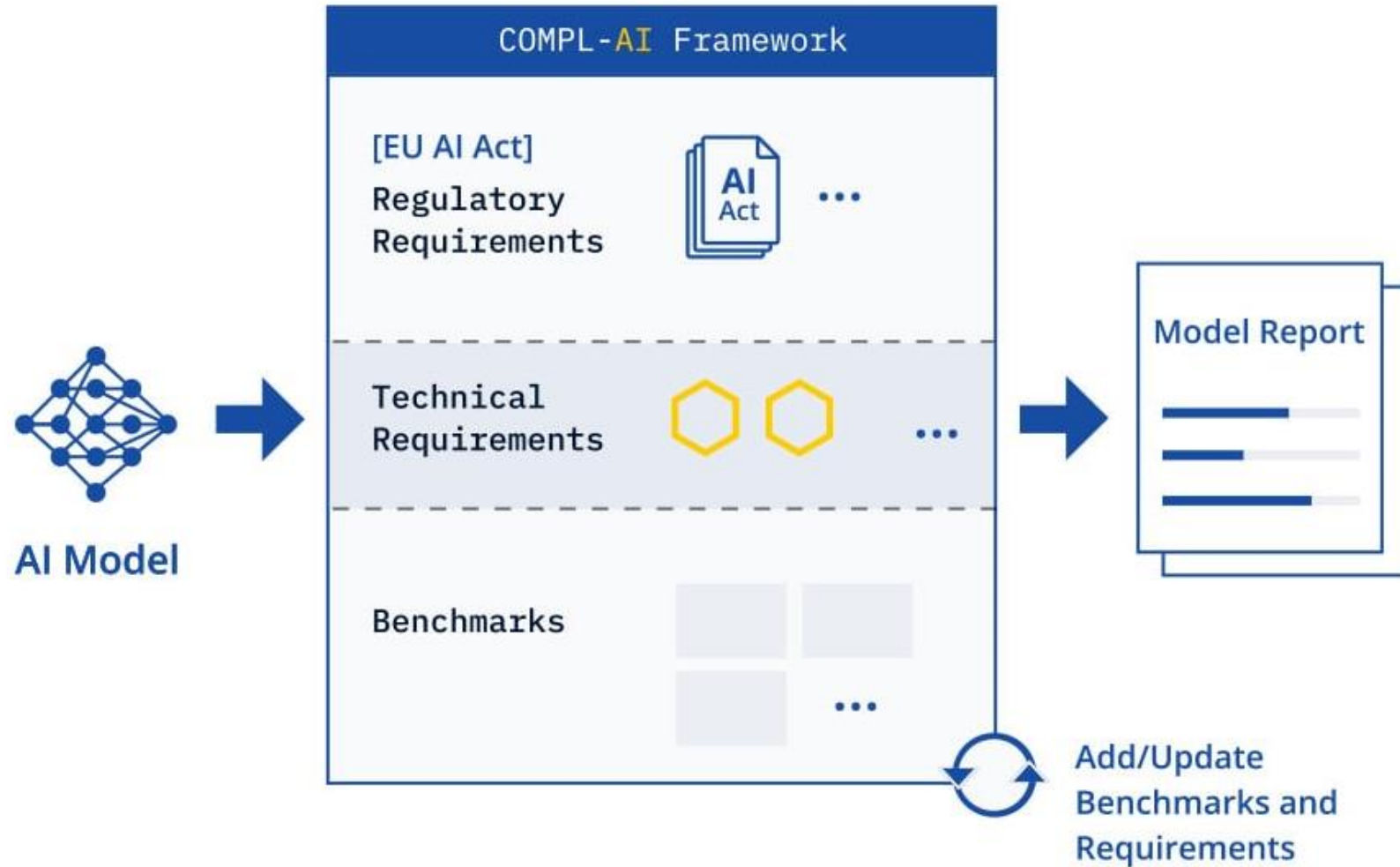


Custom AI \approx Software Building

- It's not *“Just 3 Data Scientists build a model, disappear and everything keeps running”*
- You need an organization:
 - Building the initial models
 - Building industrialization pipelines
 - Testing
 - Monitoring ML models in production
 - Maintaining & updating models
- Requires a decent AI maturity level



EU AI Act: Model Creator *Responsibilities*





EU AI Act: Model Creator *Responsibilities*

1

2

3

4

Category	Keyword	Requirement (summarized)	Section
Data	Data sources	Describe data sources used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Data governance	Use data that is subject to data governance measures (suitability, bias, and appropriate mitigation) to train the foundation model.	Amendment 399, Article 28b, page 200
	Copyrighted data	Summarize copyrighted data used to train the foundation model.	Amendment 399, Article 28b, page 200
Compute	Compute	Disclose compute (model size, computer power, training time) used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Energy	Measure energy consumption and take steps to reduce energy use in training the foundation model.	Amendment 399, Article 28b, page 200
Model	Capabilities/limitations	Describe capabilities and limitations of the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Risks/mitigations	Describe foreseeable risks, associated mitigations, and justify any non-mitigated risks of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Evaluations	Benchmark the foundation model on public/industry standard benchmarks.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Testing	Report the results of internal and external testing of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
Deployment	Machine-generated content	Disclose content from a generative foundation model is machine-generated and not human-generated.	Amendment 101, Recital 60g, page 76
	Member states	Disclose EU member states where the foundation model is on the market.	Amendment 771, Annex VIII, Section C, page 348
	Downstream documentation	Provide sufficient technical compliance for downstream compliance with the EU AI Act.	Amendment 101, Recital 60g, page 76 and Amendment 399, Article 28b, page 200

Information about trained models need to be provided:

- 1. Data** - information about the model training data
- 2. Compute** - information about the computing inputs used to train models
- 3. Model** - information about the model performance and risks
- 4. Deployment** - operational details about model use in production

[How Ready are Leading LLMs for the EU AI Act?](#)



Buy vs Build

- **Advantages:**

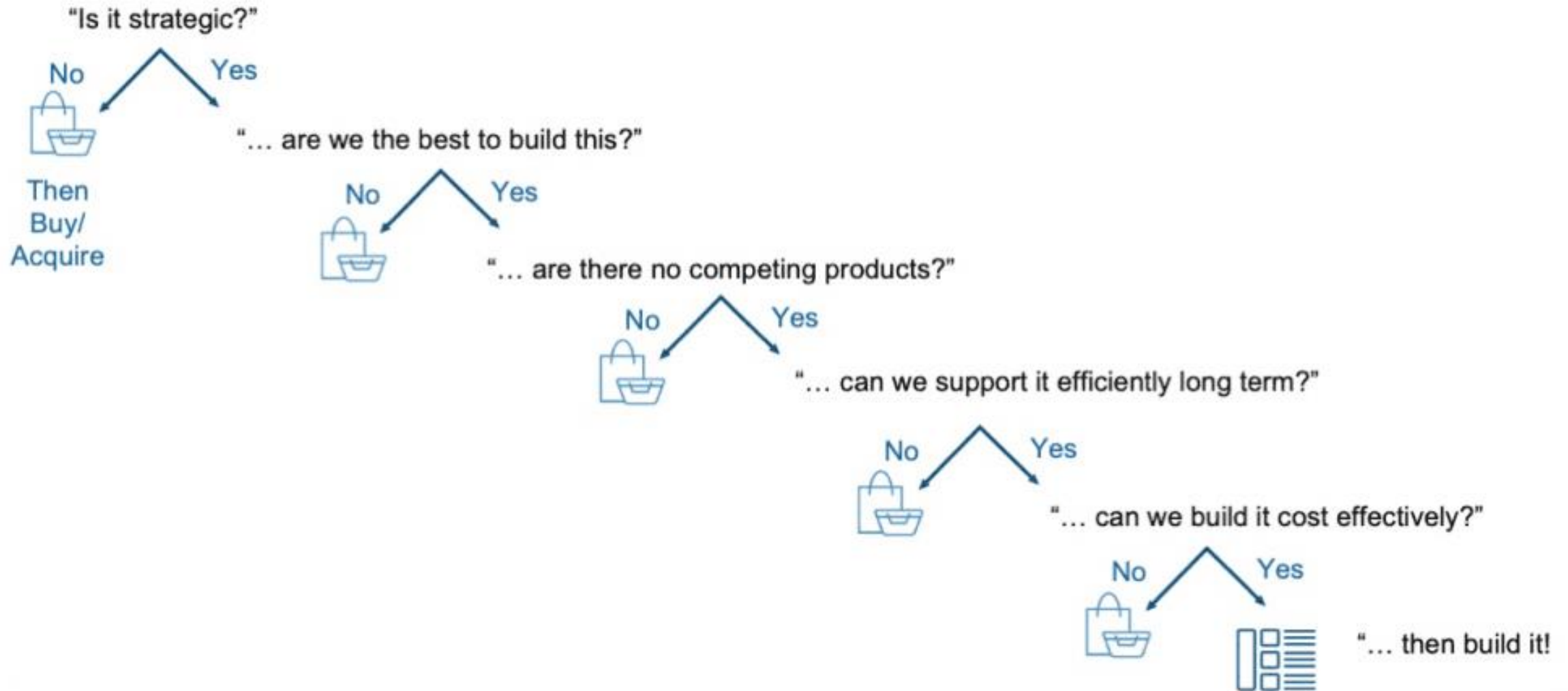
- Full control about the AI models and the underlying data
- Development of customized solutions optimized for a certain business context
- Innovation and competitive advantage

- **Disadvantages:**

- Don't underestimate the time to market
- High development and maintenance costs
- Requires specialized knowledge and capabilities



Buy vs Build



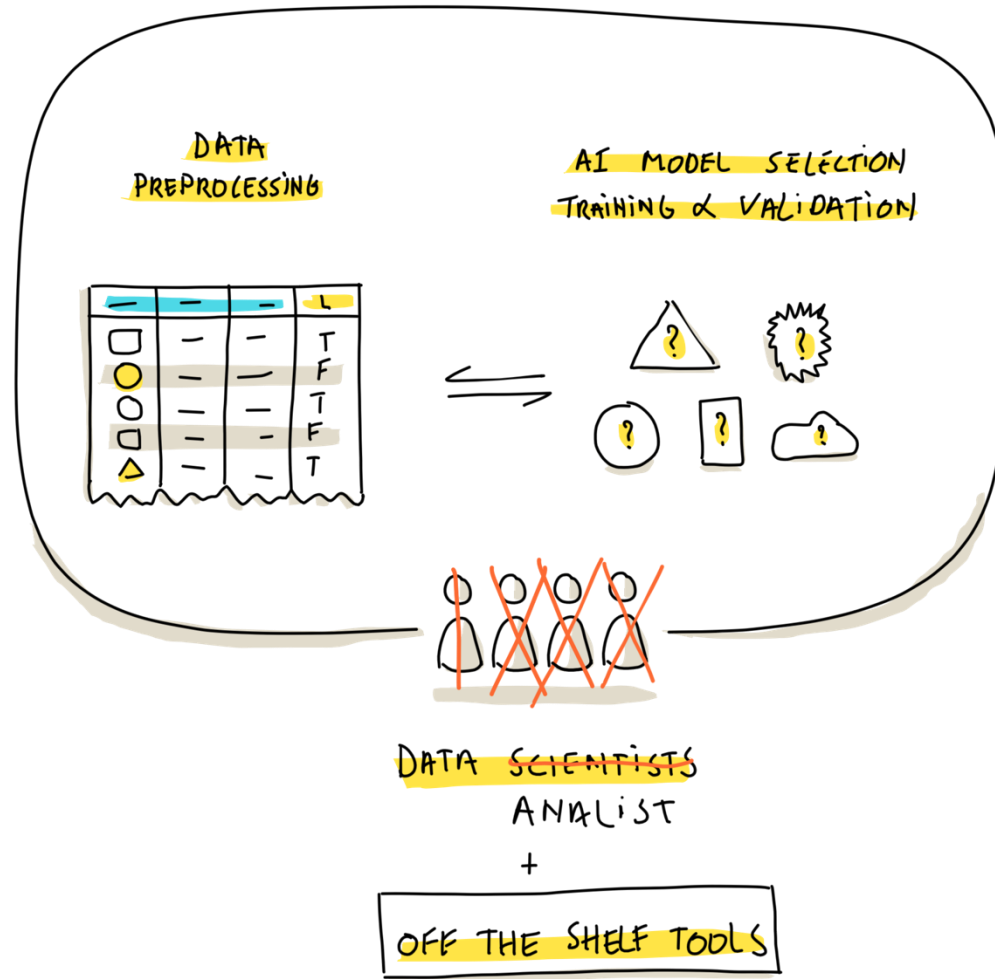


Custom vs Off-The-Shelf AI

- Introduction & examples
 - Approaches to AI Solutions
 1. Hiring Data Scientists
 2. AutoML
 - 3. Off-the-Shelf AI**
- } Custom AI
- What are you buying with off-the-shelf?
 - Trends & thoughts
 - Conclusions
 - Exercise

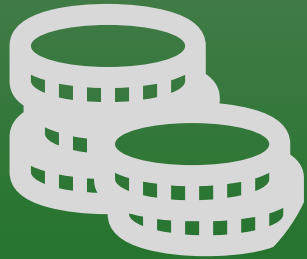


Off-the-Shelf AI





Benefits of choosing off-the-shelf solutions



Cost-effective

- From pre-investment in development to acquisition cost model
- Benefits of scale through multitenancy



Speed of deployment

- AI evolves fast
- Saves months, if not years!



Risk management

- Less financial risk (no pre-investment)
- Less operational risks



Support and maintenance

- Focus on creating value instead of Keeping-It-Running
- Less reliance on data scientists



Scalability & extensibility

- Grows with your project/product
- Fast & low cost growth





Benefits of choosing off-the-shelf solutions

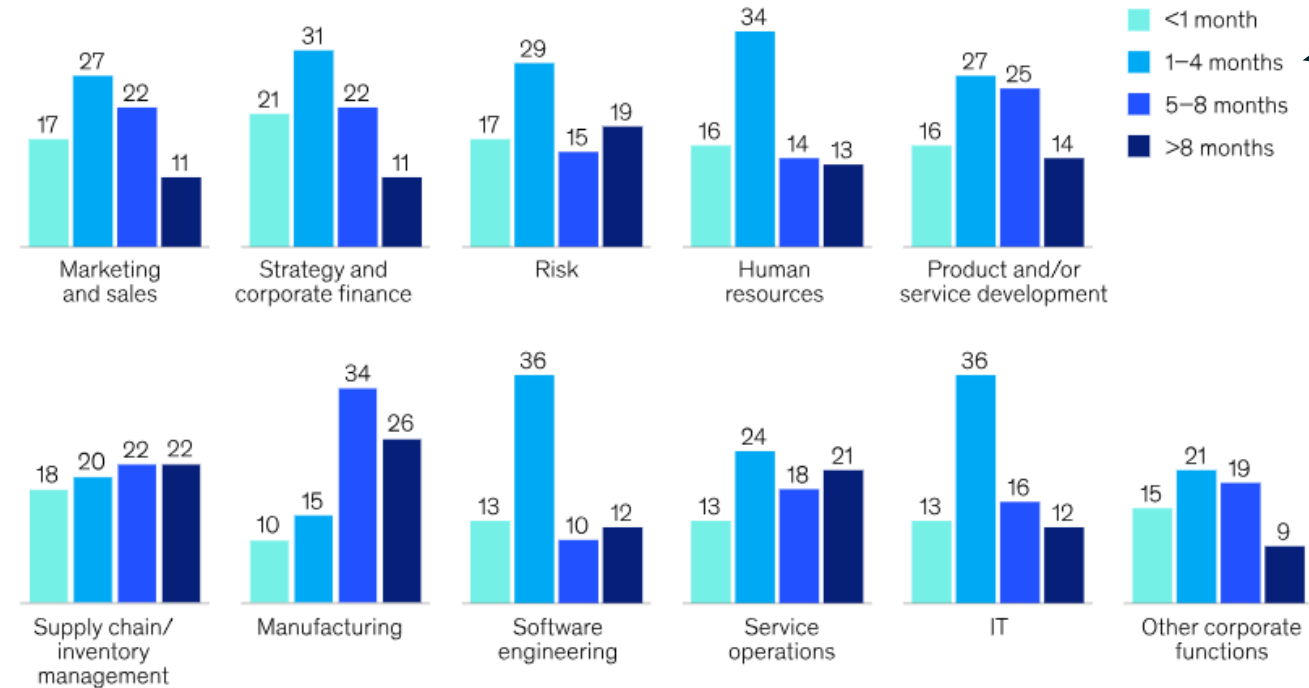


Speed of deployment

- AI evolves fast
- Saves months, if not years!

Business functions are most often able to put their generative AI capabilities to use within one to four months.

Time for organization to put generative AI capabilities to use, from project launch,¹ % of respondents



From project to value in 1 to 4 months

¹Question was asked only of respondents who said their organizations regularly use generative AI in the given business function. Respondents who said "don't know/not applicable" are not shown.
Source: McKinsey Global Survey on AI, 1,363 participants at all levels of the organization, Feb 22–Mar 5, 2024



Caveats



Limited customization

- 80% fit sufficient?
- Customization = premium price



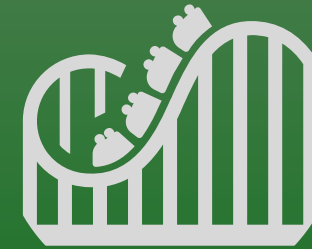
Dependency on 3rd party

- Can you share your data?
- Do you want to share your data?



Data privacy & security

- Supplier guarantees?
- Can be better than DIY, or worse



No control over product evolution

- Pivot of startups
- Pricing changes
- Strategic control over model required?
- Partnerships evolve (OpenAI – MSFT?)





Example: no control over product evolution

- OpenAI and Microsoft have an internal definition for AGI, per The Information.
- The two companies agreed to define AGI as a system that can generate \$100 billion in profits.
- OpenAI says on its website that AGI refers to AI systems that are smarter than humans.

OpenAI and Microsoft have a definition for artificial general intelligence, and it hinges on the money the emerging technology can bring in.

The two companies signed an agreement in 2023 that defined AGI as a system that can generate \$100 billion in profits. The Information reported on Thursday, citing documents it had obtained.

OpenAI has, however, publicly defined AGI on its website as "a highly autonomous system that outperforms humans at most economically valuable work."

The ChatGPT maker added that its nonprofit board would decide whether AGI has been achieved.

→ "Such a system is excluded from IP licenses and other commercial terms with Microsoft, which only apply to pre-AGI technology," the company wrote on its website.

INTELLIGENT COG IN THE WHEEL

Sam Altman says "we are now confident we know how to build AGI"

The race to replace human workers continues in Big Tech, but not everyone is convinced it will happen so soon.

BENJ EDWARDS - 6 JAN 2025 18:18 | 374

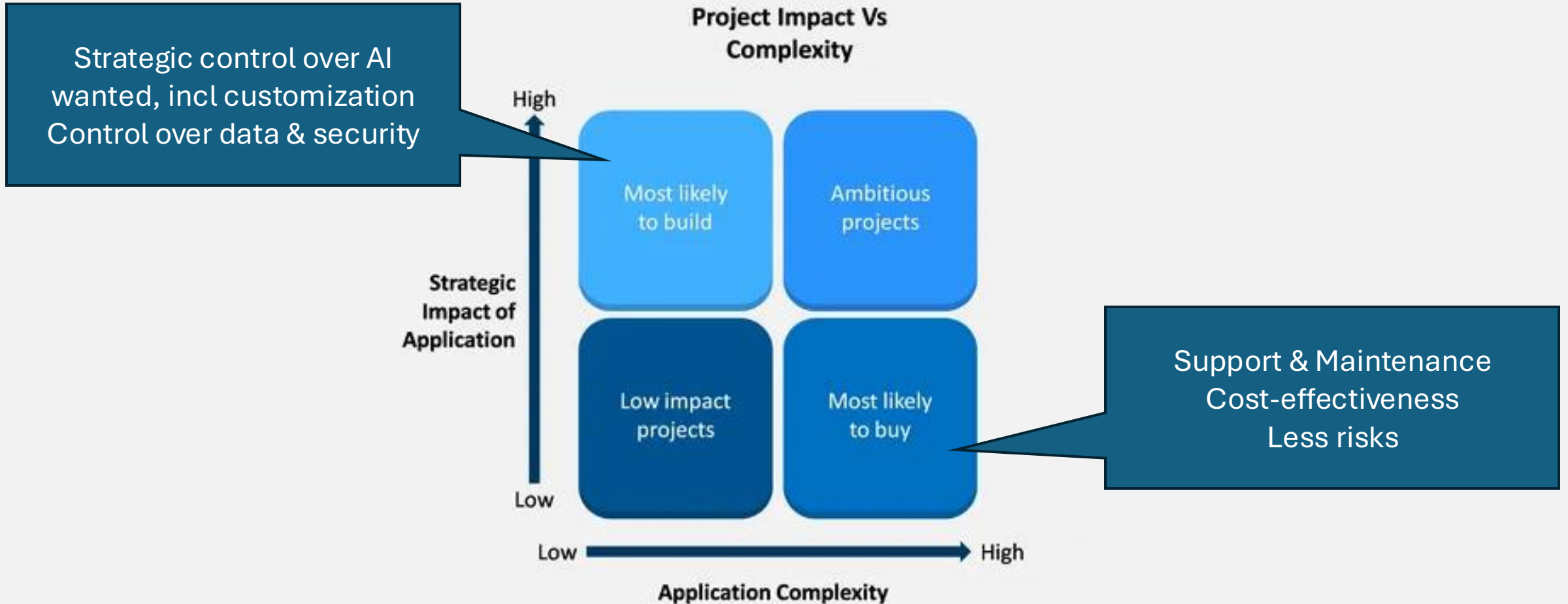


❖ Sam Altman speaks onstage during The New York Times Dealbook Summit 2024 at Jazz at Lincoln Center on December 04, 2024 in New York City. Credit: Eugene Gologursky via Getty Images

On Sunday, OpenAI CEO Sam Altman offered two eye-catching predictions about the near-future of artificial intelligence. In a post titled "Reflections" on his personal blog, Altman wrote, "We are now confident we know how to build AGI as we have traditionally understood it." He added, "We believe that, in 2025, we may see the first AI agents 'join the workforce' and materially change the output of companies."



Buy vs Build (revisited)

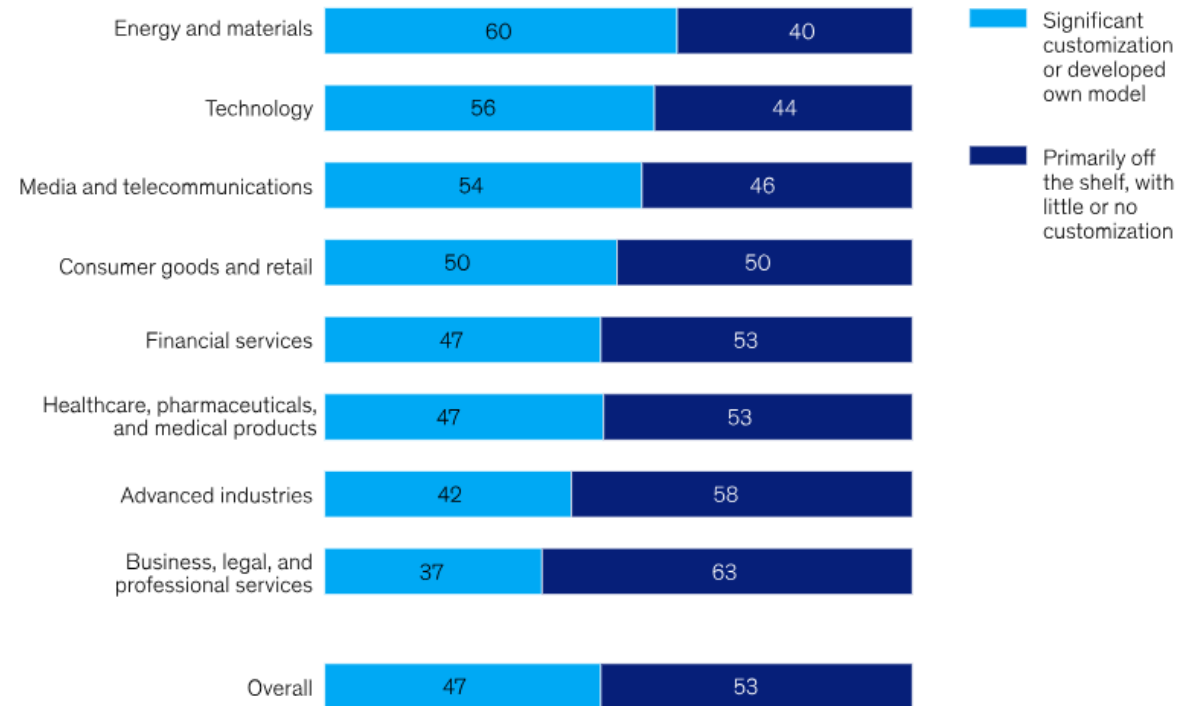




Different discussion for GenAI?

Organizations are pursuing a mix of off-the-shelf generative AI capabilities and also significantly customizing models or developing their own.

Strategy for developing generative AI (gen AI) capabilities, % of reported instances of gen AI use¹



The answer isn't "do nothing"

Build versus buy is often framed as an either-or, and some organizations are considering a third option: do nothing.

Some companies come to the conclusion that they should take a wait-and-see approach, rather than make an investment. In the fast-moving world of AI, that might feel tempting. It's also a mistake. You'll miss out on months (or years) of short-term ROI gains and long-term differentiation from your competitors.

¹Question was asked only of respondents who said their organizations regularly use generative AI in at least 1 business function. Figures were calculated after removing respondents who said "don't know."
Source: McKinsey Global Survey on AI, 1,363 participants at all levels of the organization, Feb 22–Mar 5, 2024



Different discussion for GenAI?

- Generative AI – or foundational models in general – support multiple use cases, but they might be too generic.

Generative AI Solutions: Build vs Buy Comparison

Criteria	Build	Buy
Cost	High upfront investment and ongoing costs	Potential lower upfront costs, mainly for subscription or licensing fees
Implementation Time	Longer due to the development and testing phases	Shorter, as the solution is ready-made
Customization	Highly customizable to specific needs	Limited customization options
Use Cases	Tailored to unique and specific use cases	General use cases
Scalability	Designed for high scalability	Depends on the provider's infrastructure

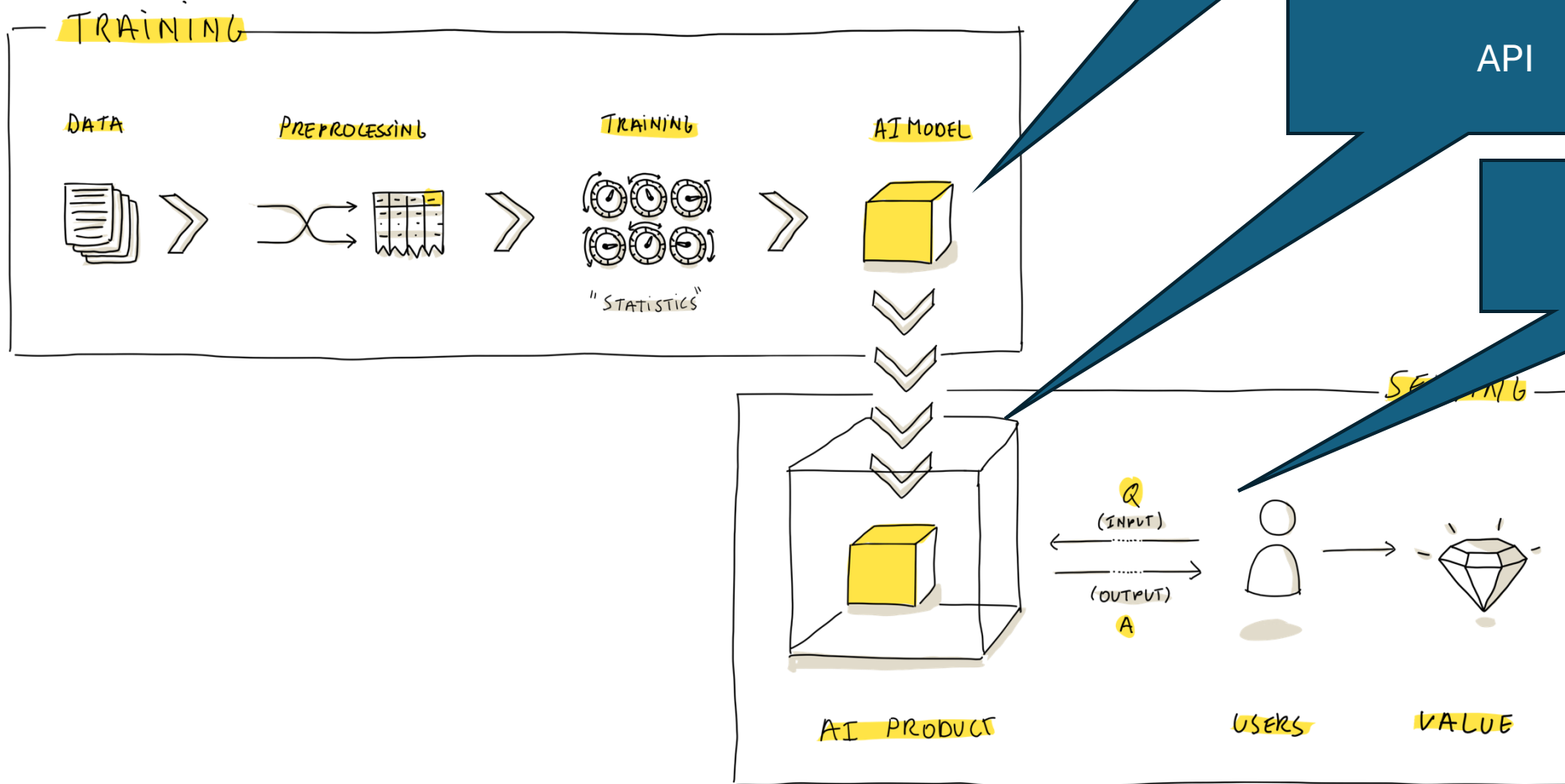


Custom vs Off-The-Shelf AI

- Introduction & examples
- Approaches to AI Solutions
- **What are you buying with off-the-shelf?**
- Conclusions
- Exercise



Recap: AI Product





Generative AI Deployment Approaches

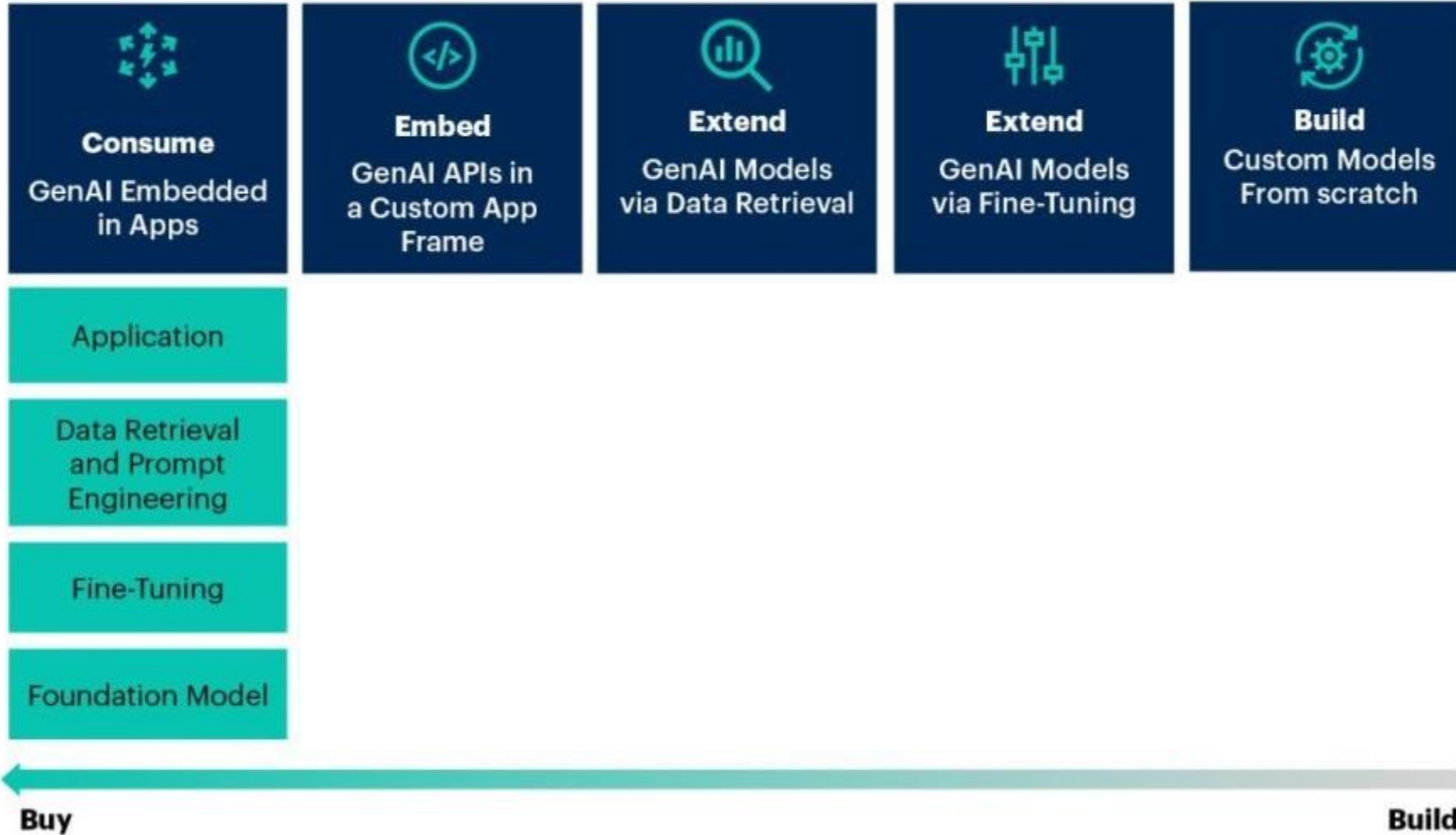
■ Provider-Managed ■ Self-Managed





Generative AI Deployment Approaches

■ Provider-Managed ■ Self-Managed





GenAI deployment approach: consume

VIVA
CITY

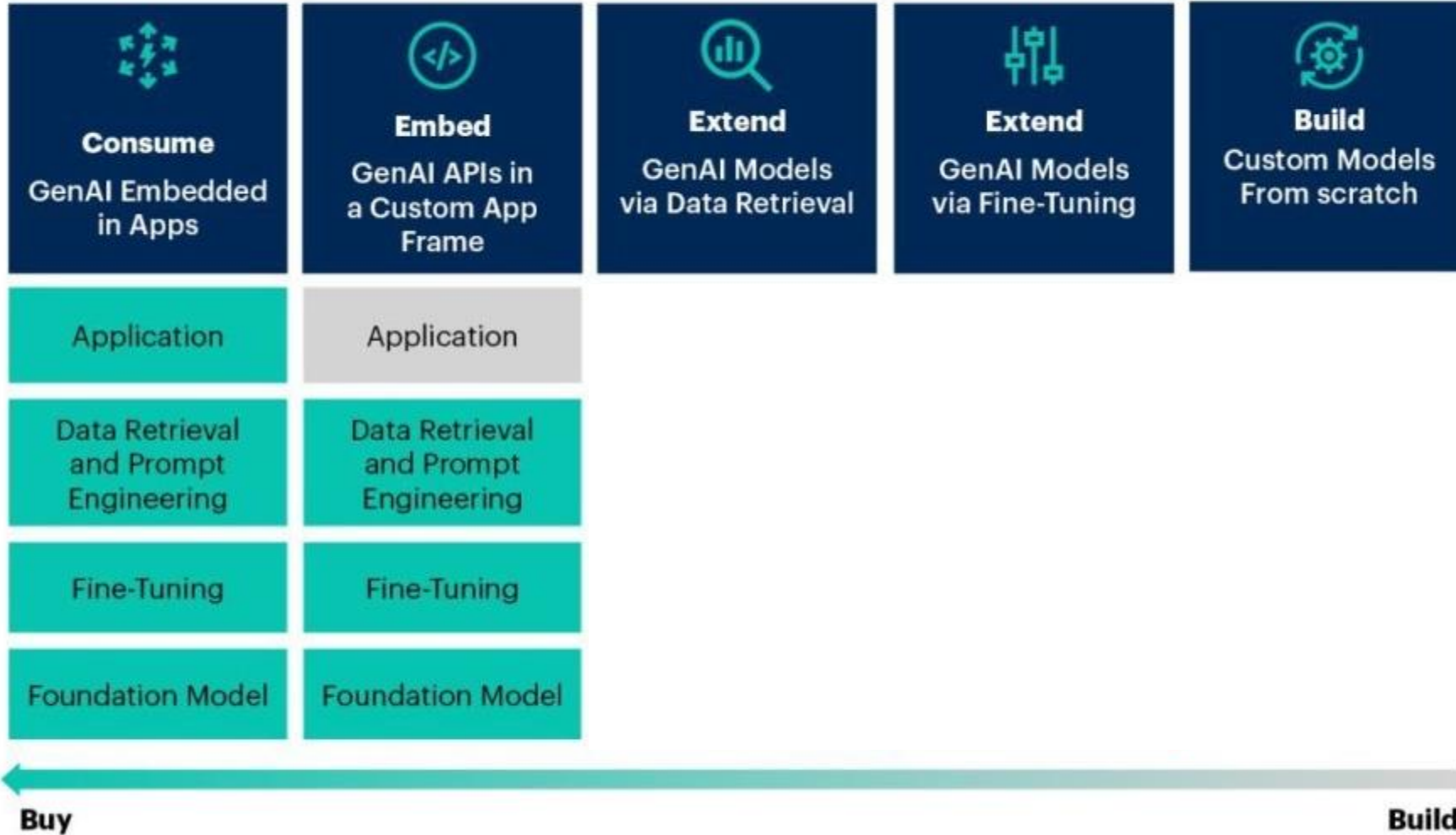
SMARTER
SAFER
SUSTAINABLE
CITIES





Generative AI Deployment Approaches

■ Provider-Managed ■ Self-Managed





GenAI deployment approach: embed

The Vectara End to End Platform

Vectara is a trusted platform for quickly building AI assistants and agents grounded in your own data, offering the trust and control enterprises need.

[Get started](#) [Book a demo >](#)

Builder → **End user**

Vectara

- 1 Extract
- 2 Encode
- 3 Index
- 4 Retrieve
- 5 Rerank
- 6 Generate
- 7 Evaluate

CARBON [Features](#) [Documentation](#) [Customers](#) [Changelog](#) [Self-Host](#) [Book Demo](#)

Platform Resources For Businesses Developers Pricing [Book a demo](#) [Start Free Trial](#)

Nuclia, the #1 all-in-one RAG as a service platform.

Nuclia automatically indexes files and documents, from internal and external sources, to fuel diverse company use cases with LLMs assuring RAG quality.

[Get Started FREE](#)

[LlamaIndex](#) [BOOK A DEMO](#) [GET STARTED](#)

LlamaCloud: Accurate and secure knowledge management for AI Agents

LlamaCloud is the most accurate, secure, and seamless knowledge management layer to make any unstructured data LLM-ready (PDFs, PPTX, XSLX, and more).
Productionize knowledge assistants in hours vs. weeks.

[GET STARTED](#)

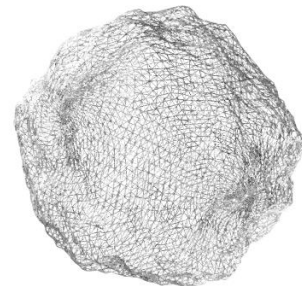
Why Enterprises in Finance, Professional Services, Manufacturing and More Choose LlamaCloud

- 10x Increase in Dev Velocity**
Reduced setup time for data pipeline from ~5 weeks to 3 hours (Cemex)
- 3x Use Cases Delivered**
Top 5 professional services firm used to deliver 1 use case every 3 months. After adopting LlamaCloud, same team delivers a use case in 3-4 weeks!
- Massive Performance Boost**
The *only* RAG solution that can handle complex documents with embedded tables, charts, images, and more

Connect external data to LLMs, no matter the source.

Carbon is a universal retrieval engine for LLMs to access unstructured data from any source.

[Get Started](#)

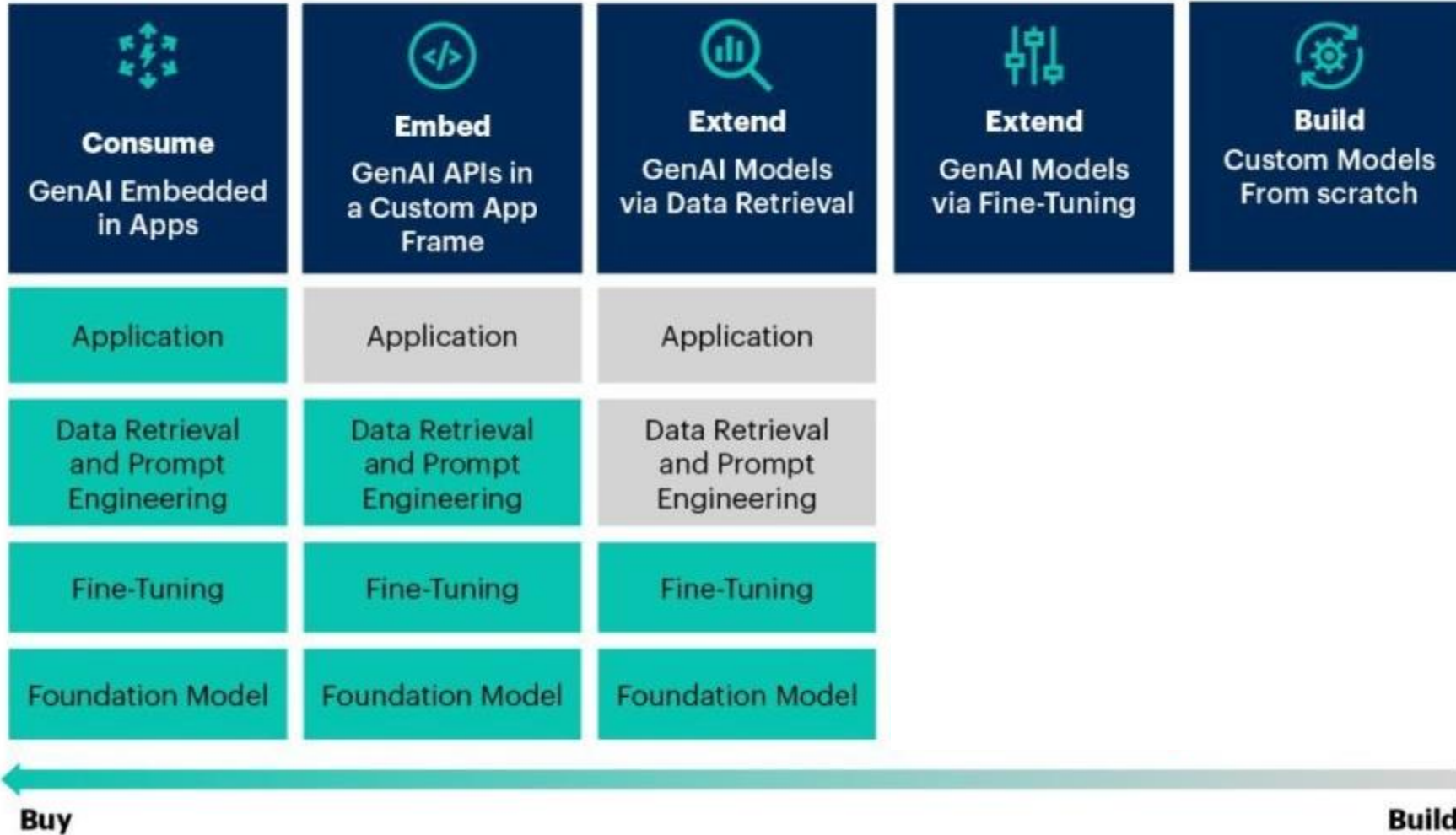


Built with Spline



Generative AI Deployment Approaches

■ Provider-Managed ■ Self-Managed





GenAI deployment approach: extend

OpenAI Research **Products** Safety Company Q

Flagship Models

Our reasoning models

These models spend more time thinking before producing a response, making them ideal for complex, multi-step problems.

o1

Our most powerful reasoning model that supports tools, Structured Outputs, and vision

[Learn more](#)

o1-mini

Our small reasoning model that thinks faster than o1 and is optimized for coding and math

[Learn more](#)

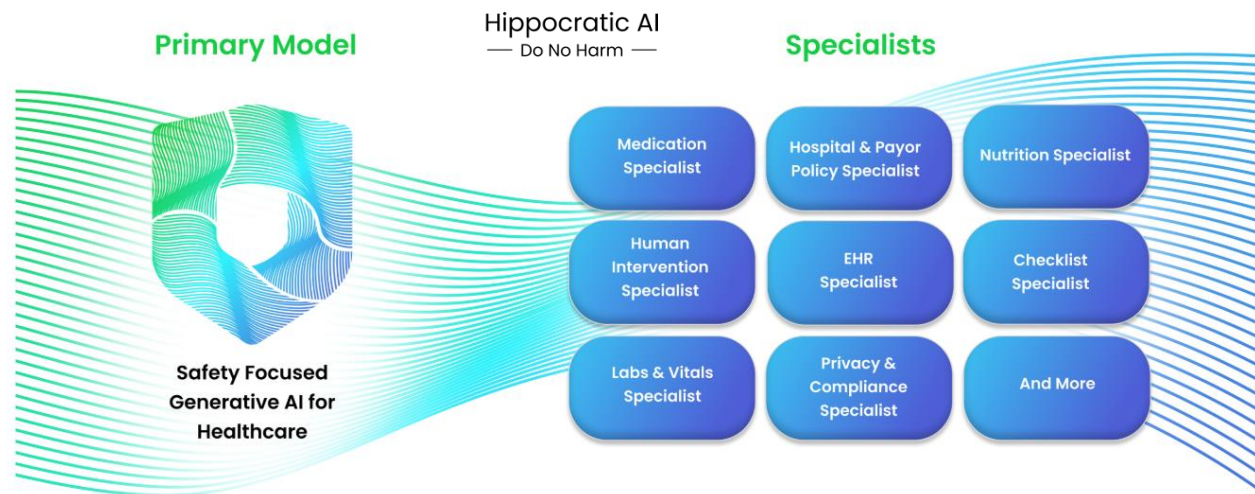
OpenAI NEWS MODELS **PRODUCTS** COM

Classifier

High performance zero-shot and few-shot classifier for multimodal and multilingual data.

- Reader**
Read URLs and search web for better grounding LLMs.
- Embeddings**
World-class multimodal multilingual embeddings.
- Reranker**
World-class neural retriever for maximizing search relevancy.
- Classifier**
Zero-shot and few-shot classification for image and text.
- Segmenter**
Cut long text into chunks and do tokenization.

[<> API](#) [\\$ Pricing](#)





GenAI deployment approach: extend

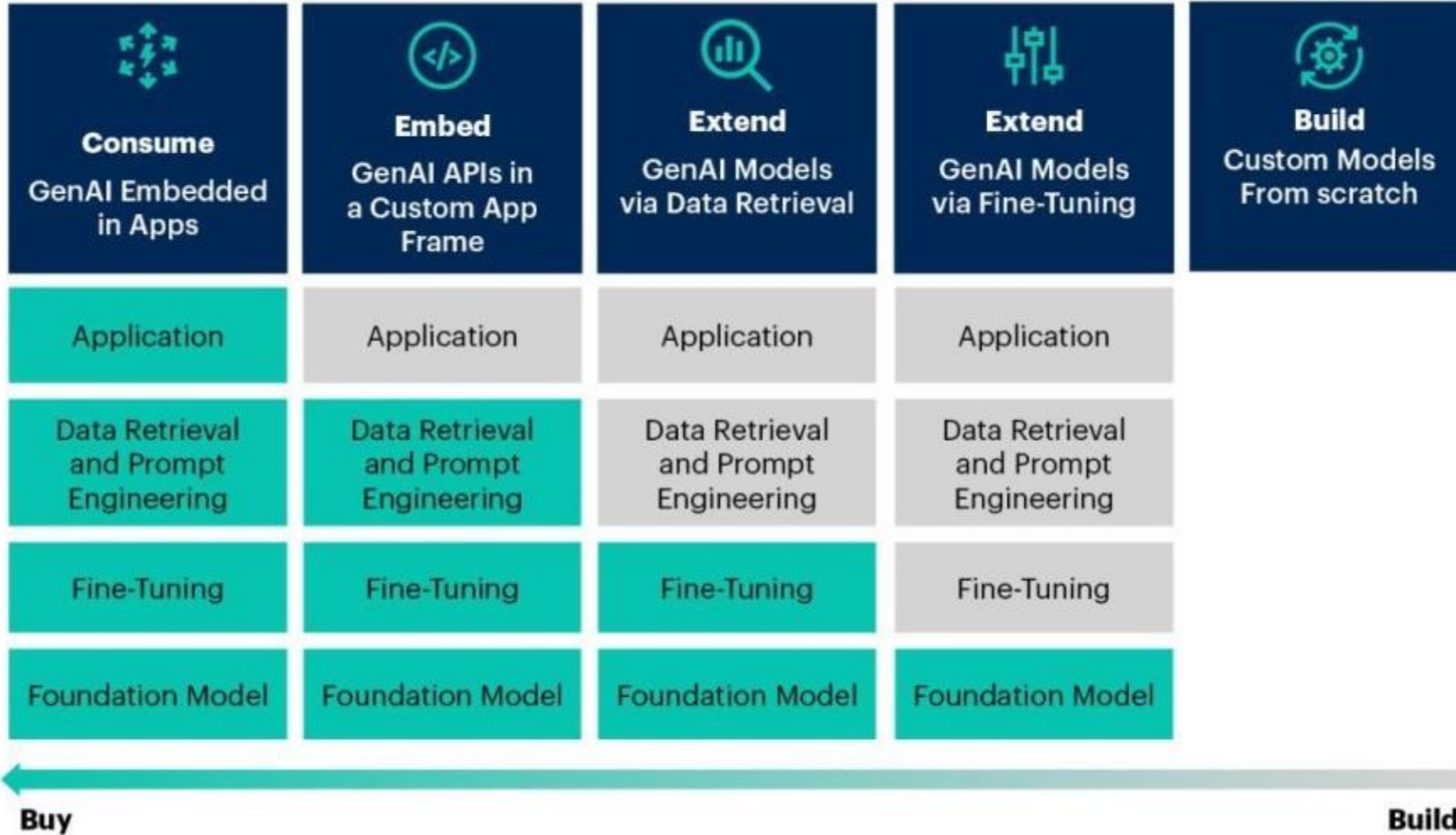
- Sidetrack – **BUILD**: once you take control over data retrieval / vector stores, prompt engineering, model interfacing yourself, leverage abstraction frameworks to have less dependency on the underlying genAI hyperscaler (OpenAI, Google, Meta, ...):





Generative AI Deployment Approaches

■ Provider-Managed ■ Self-Managed





GenAI deployment approach: extend (finetune)



Which models can be fine-tuned?

Fine-tuning is currently available for the following models:

- `gpt-4o-2024-08-06`
- `gpt-4o-mini-2024-07-18`
- `gpt-4-0613`
- `gpt-3.5-turbo-0125`
- `gpt-3.5-turbo-1106`
- `gpt-3.5-turbo-0613`

Llama

The open-source AI models you can fine-tune, distill and deploy anywhere. Choose from our collection of models: Llama 3.1, Llama 3.2, Llama 3.3.

[Download models](#) [Try Llama on Meta AI](#)



[Falcon Home](#) [Our Research](#) [Datasets](#) [Falcon Model](#) [FAQs](#)

Falcon Models

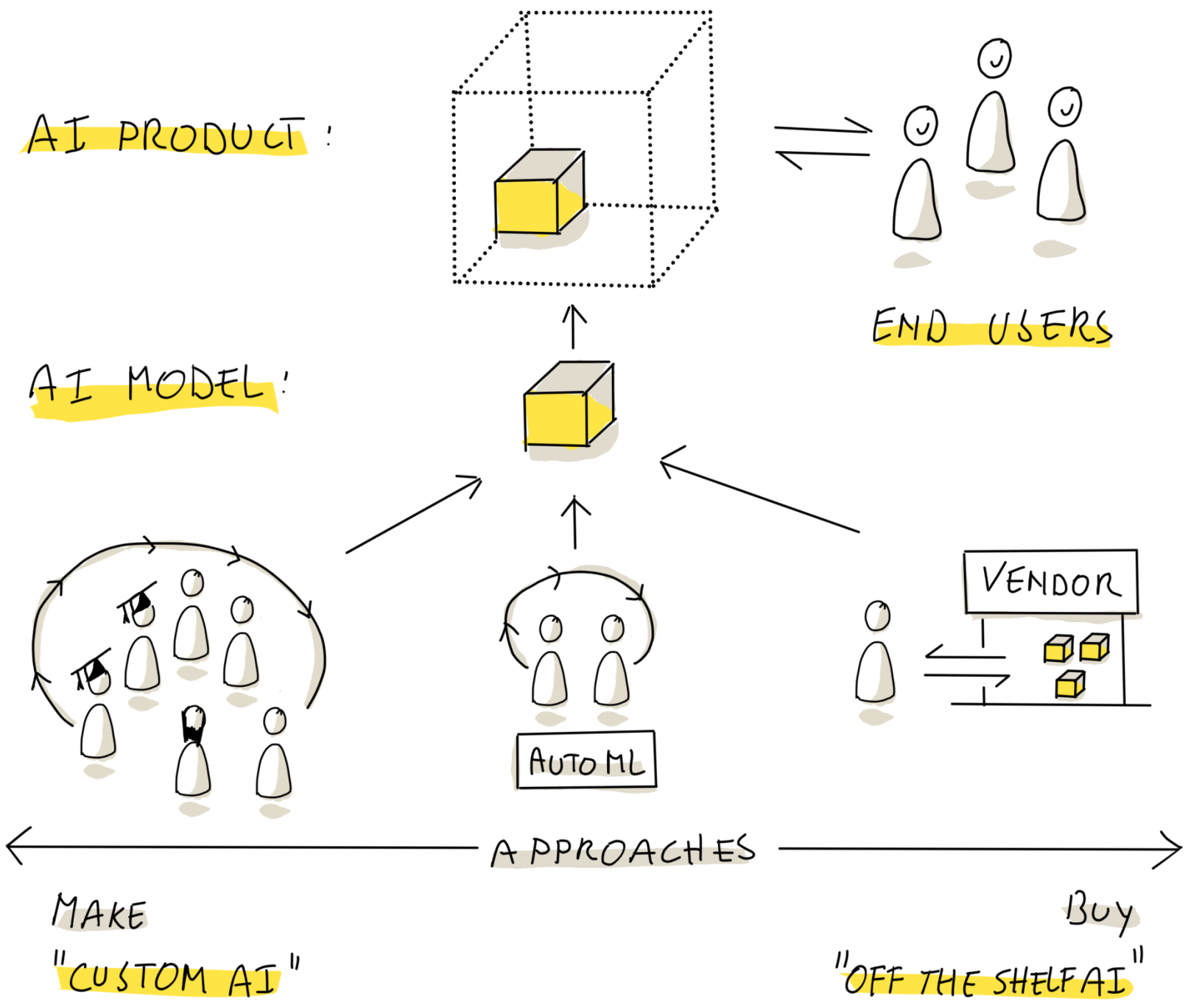
Falcon LLM is a generative large language model (LLM) that helps advance applications and use cases to future-proof our world. Today the Falcon 3, Falcon Mamba 7B, Falcon 2, 180B, 40B, 7.5B, 1.3B parameter AI models, as well as our high-quality REFINEDWEB dataset, form a suite of offerings.





Custom vs Off-The-Shelf AI

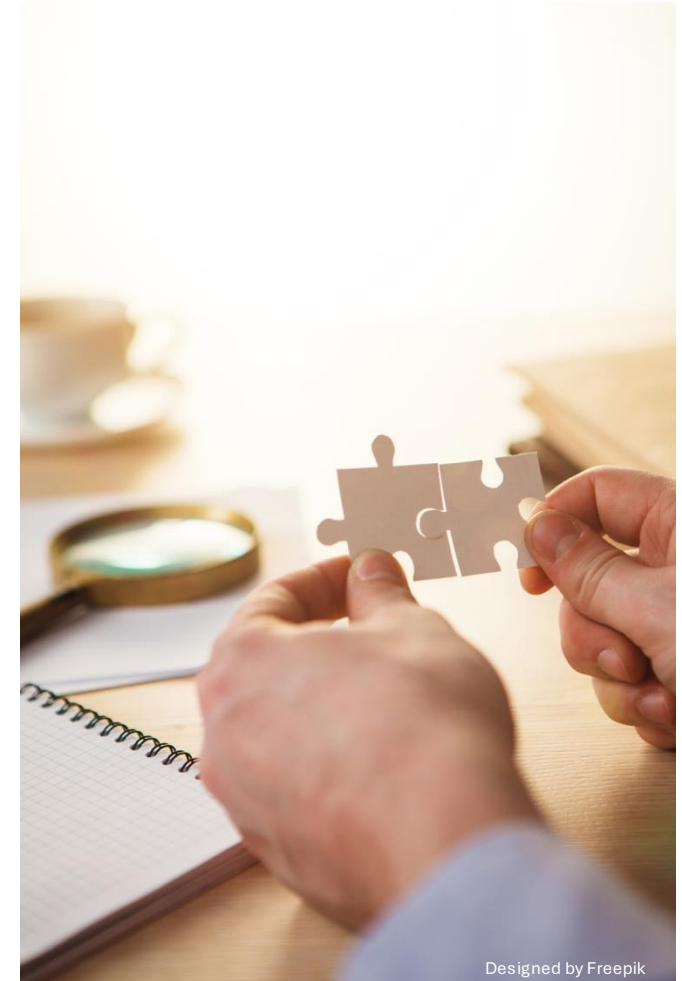
- Introduction & examples
- Approaches to AI Solutions
- What are you buying with off-the-shelf?
- Trends & thoughts
- **Conclusions**
- Exercise





Conclusions

- **Think upfront** what strategy to choose
 - Build is not per se the default option!
GenAI (>Nov '22) is completely reshaping how we look at AI projects
 - McKinsey: value within 1-4 months (!)
- Custom AI or off-the-shelf
 - Economic: make or buy decision?
 - Time: value fast vs project
 - Strategic importance of having control?
- There is **no right or wrong**
 - Make sure you understand the ownership you take and get in each decision & the work that comes with it
 - Take it into account when devising your project plan





Custom vs Off-The-Shelf AI

- Introduction & examples
- Approaches to AI Solutions
- What are you buying with off-the-shelf?
- Conclusions
- **Exercise**

OEFENING: Welke aanpak is relevant voor uw use case?

- Go aan de slag voor jouw case:
 - Wat is voor jouw case relevant?
 - Custom AI?
 - Off-The-Shelf AI?
 - Waarom?

