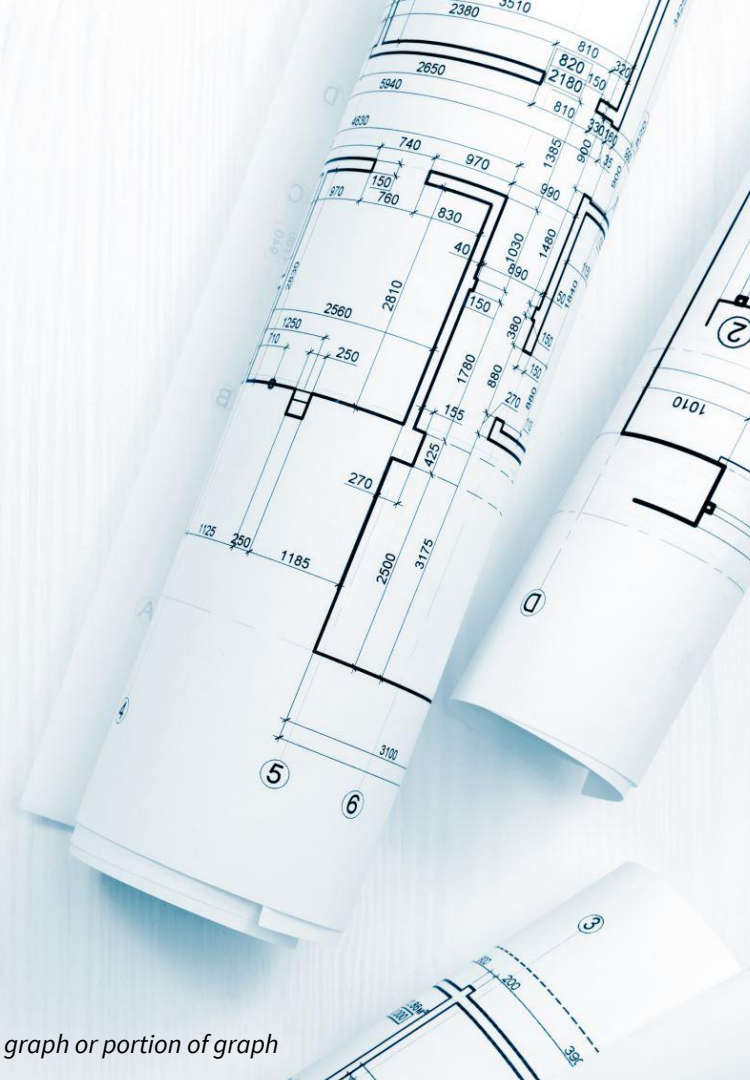
Planning *the* AI Architecture, Infrastructure & Operations

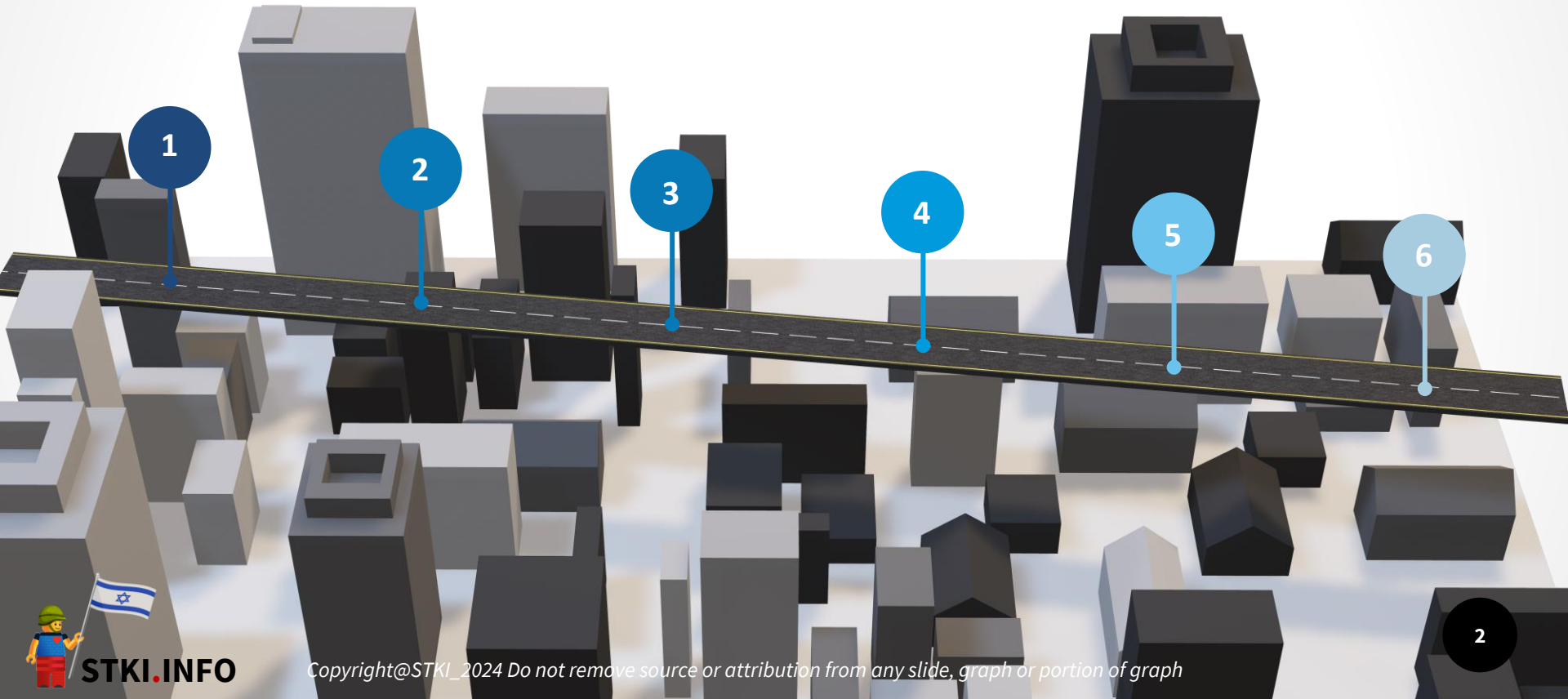
Pini Cohen
Einat Shimoni



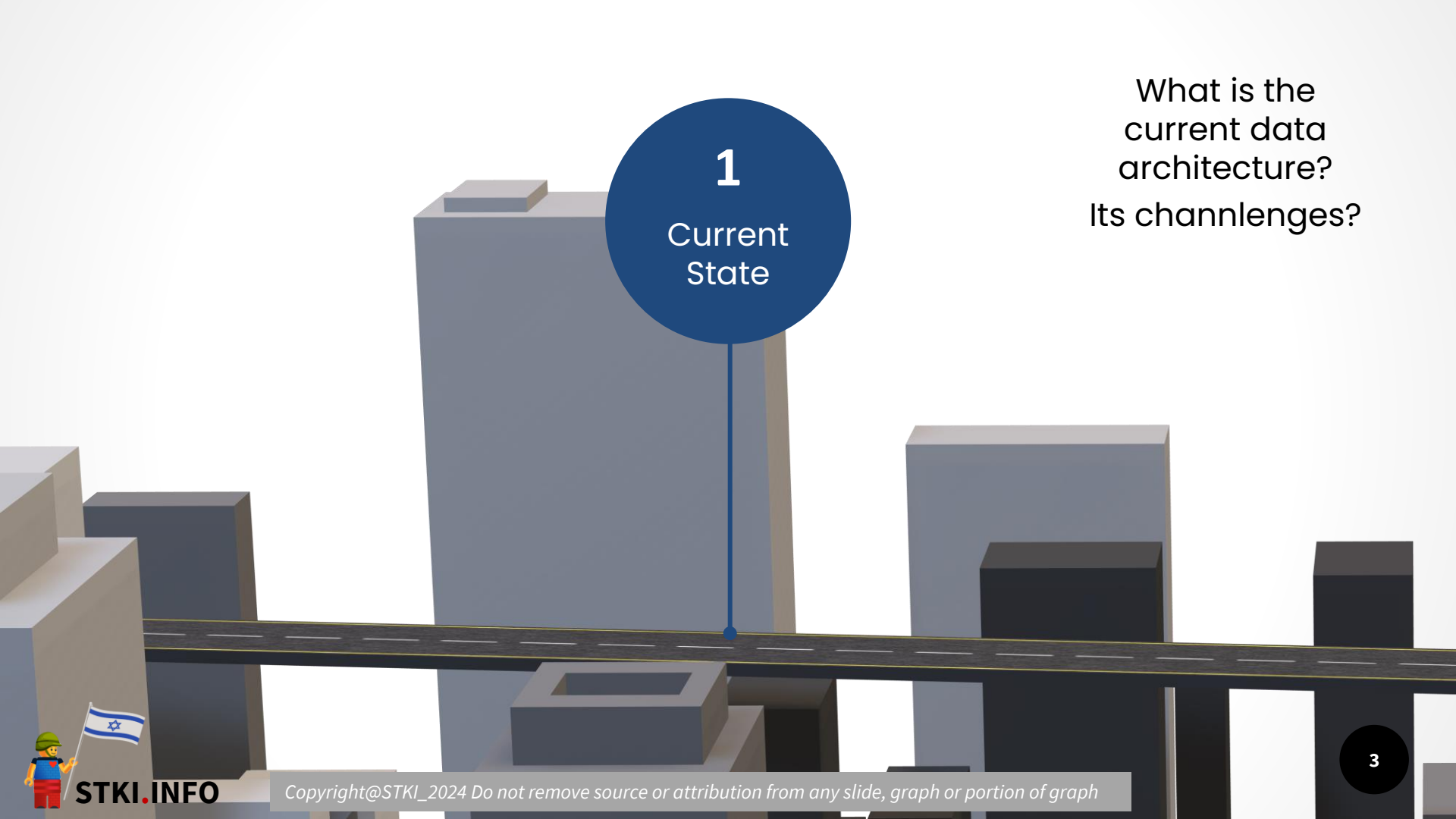
STKI.INFO

This presentation will take you on a journey

Towards planning your organizations' AI architecture, infrastructure and operations



STKI.INFO



What is the current data architecture?

Its channlenges?

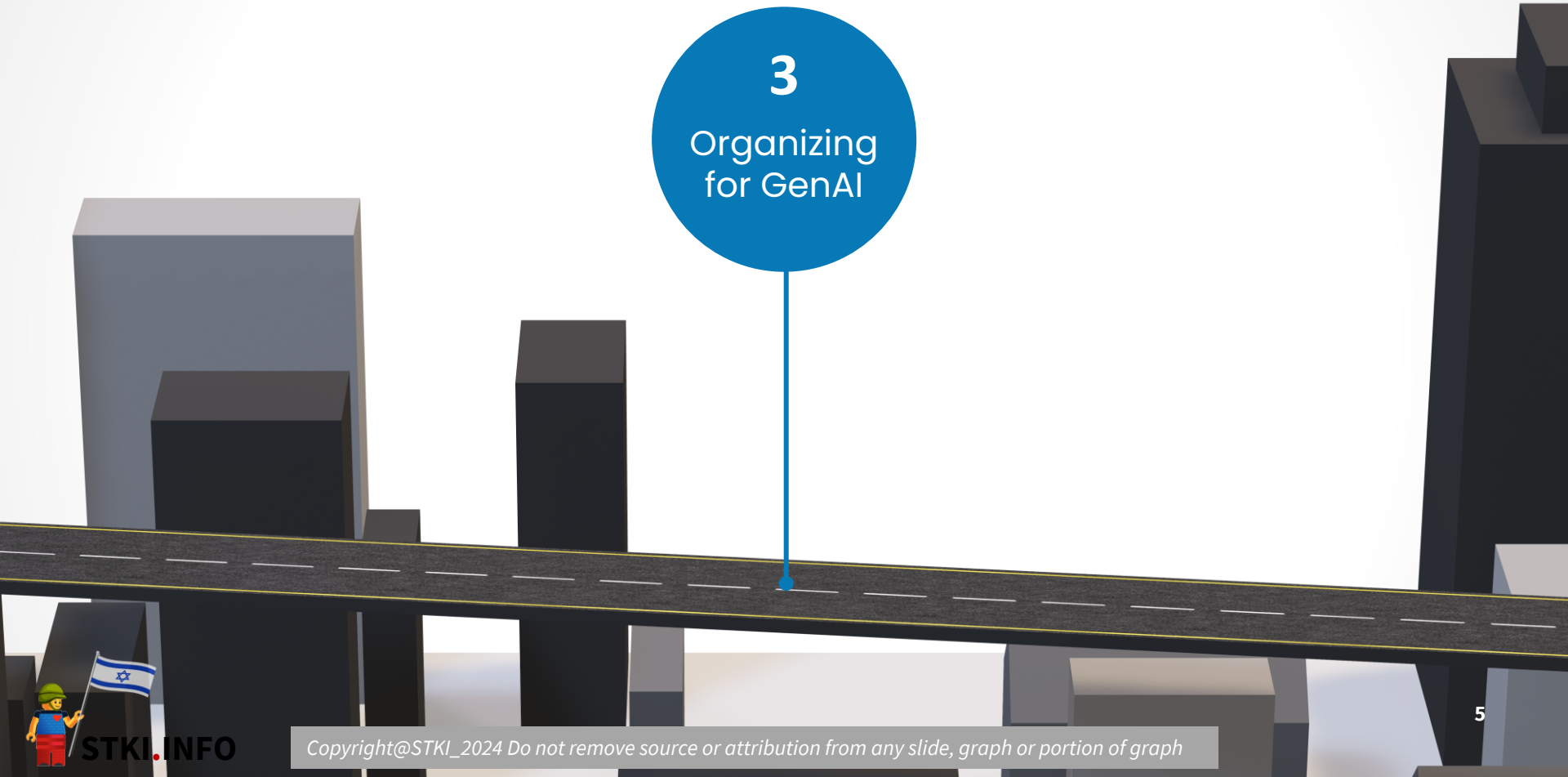
What are the needed architecture, operations and infrastructure to effectively deploy AI?

2
Future Architecture & Tools

What are the different tools categories?

STKI.INFO

How should companies organize to leverage GenerativeAI models?



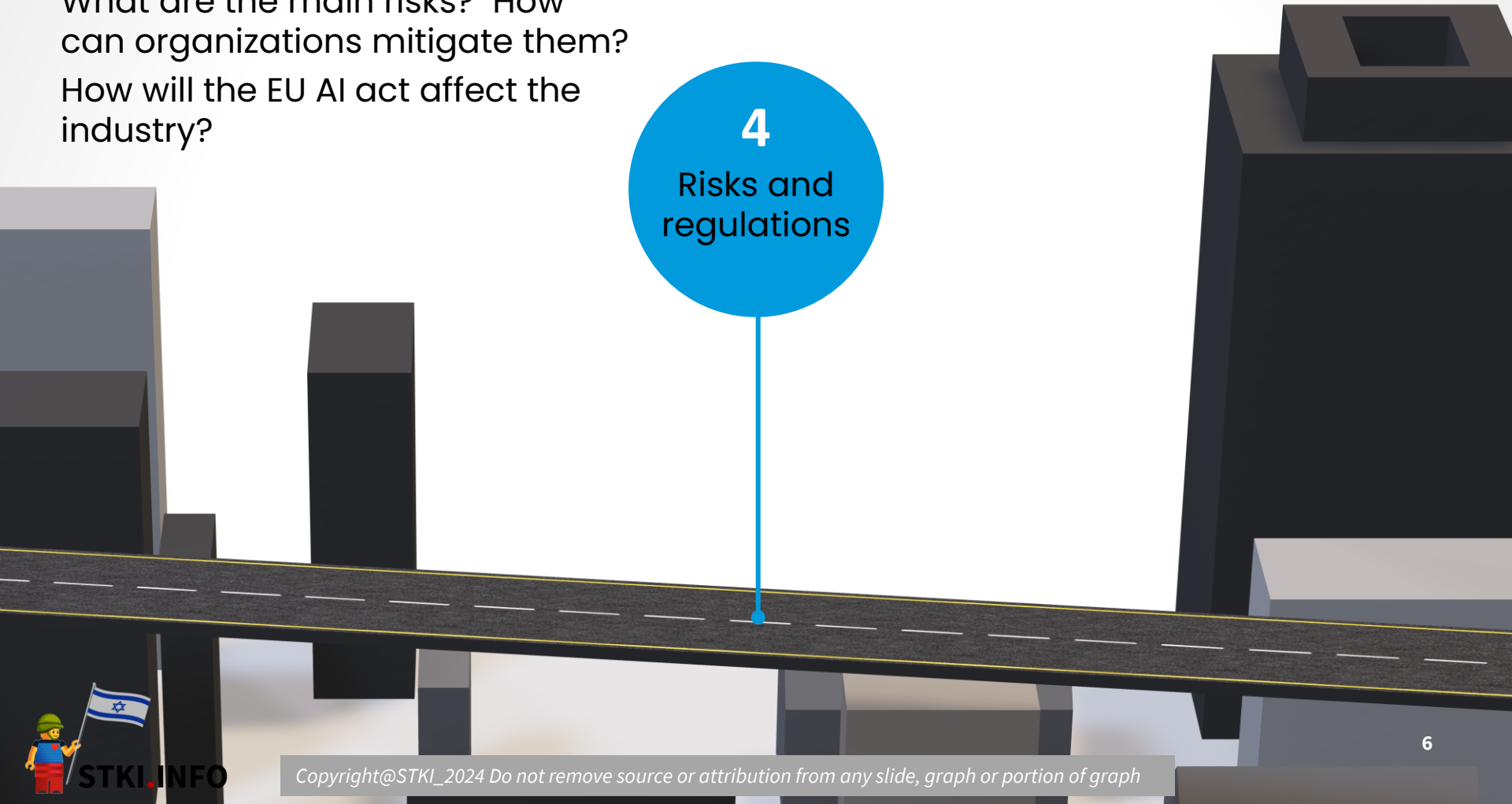
STKI.INFO

What are the main risks? How can organizations mitigate them?

How will the EU AI act affect the industry?

4

Risks and regulations

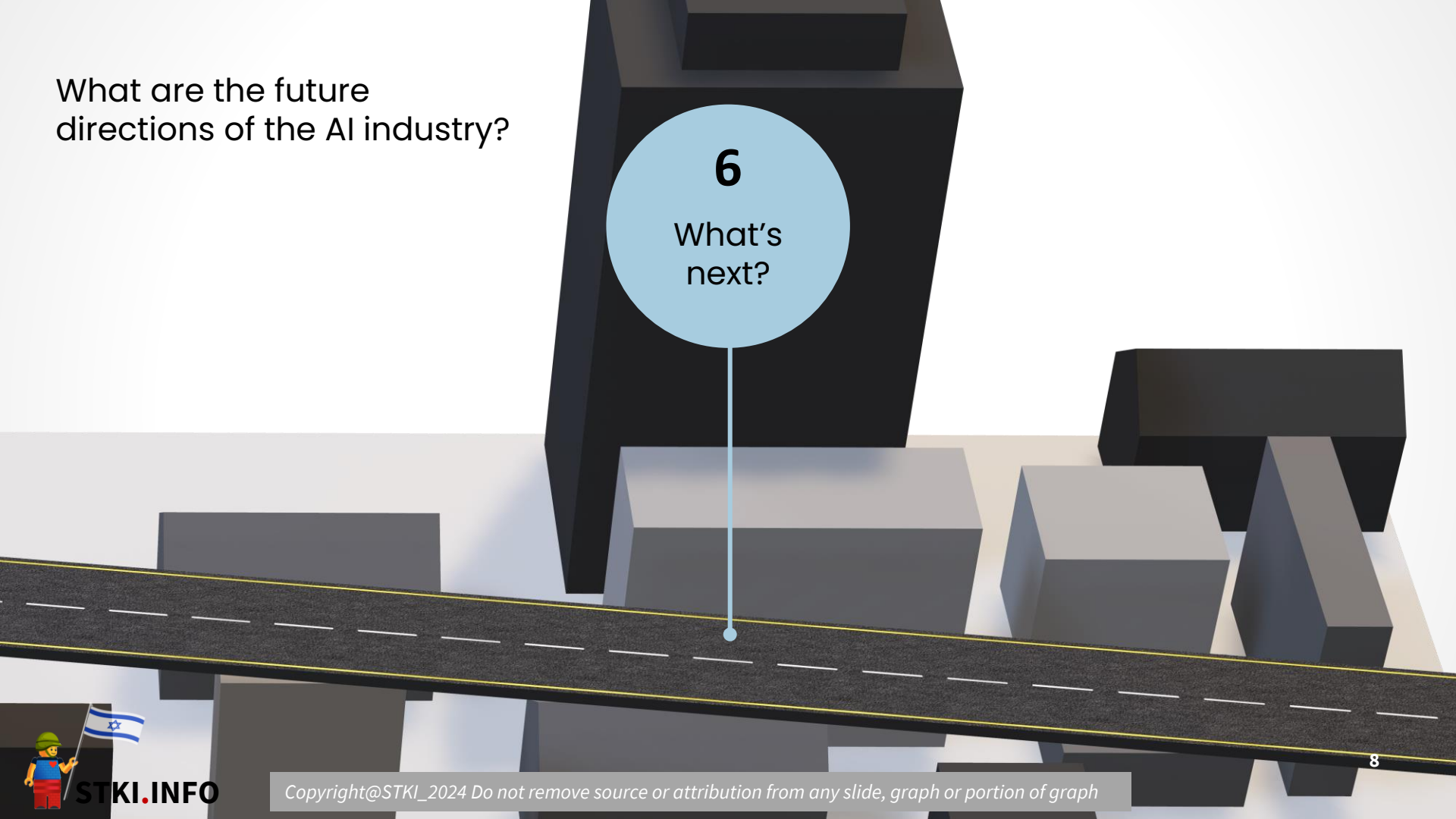


Are there solutions to mitigate risks?
What should organizations do?

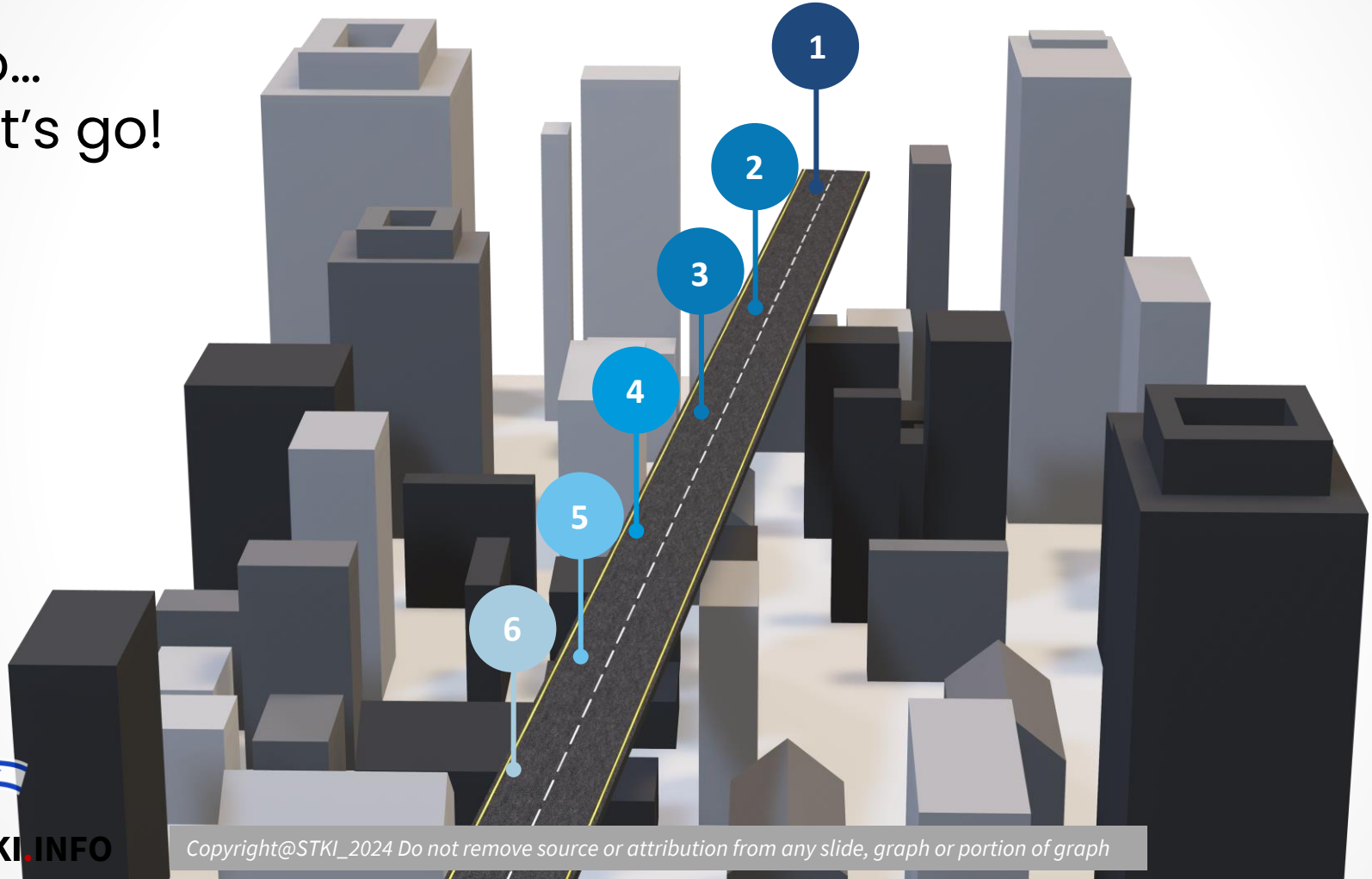
5
Solutions & Recommendations

STKI.INFO

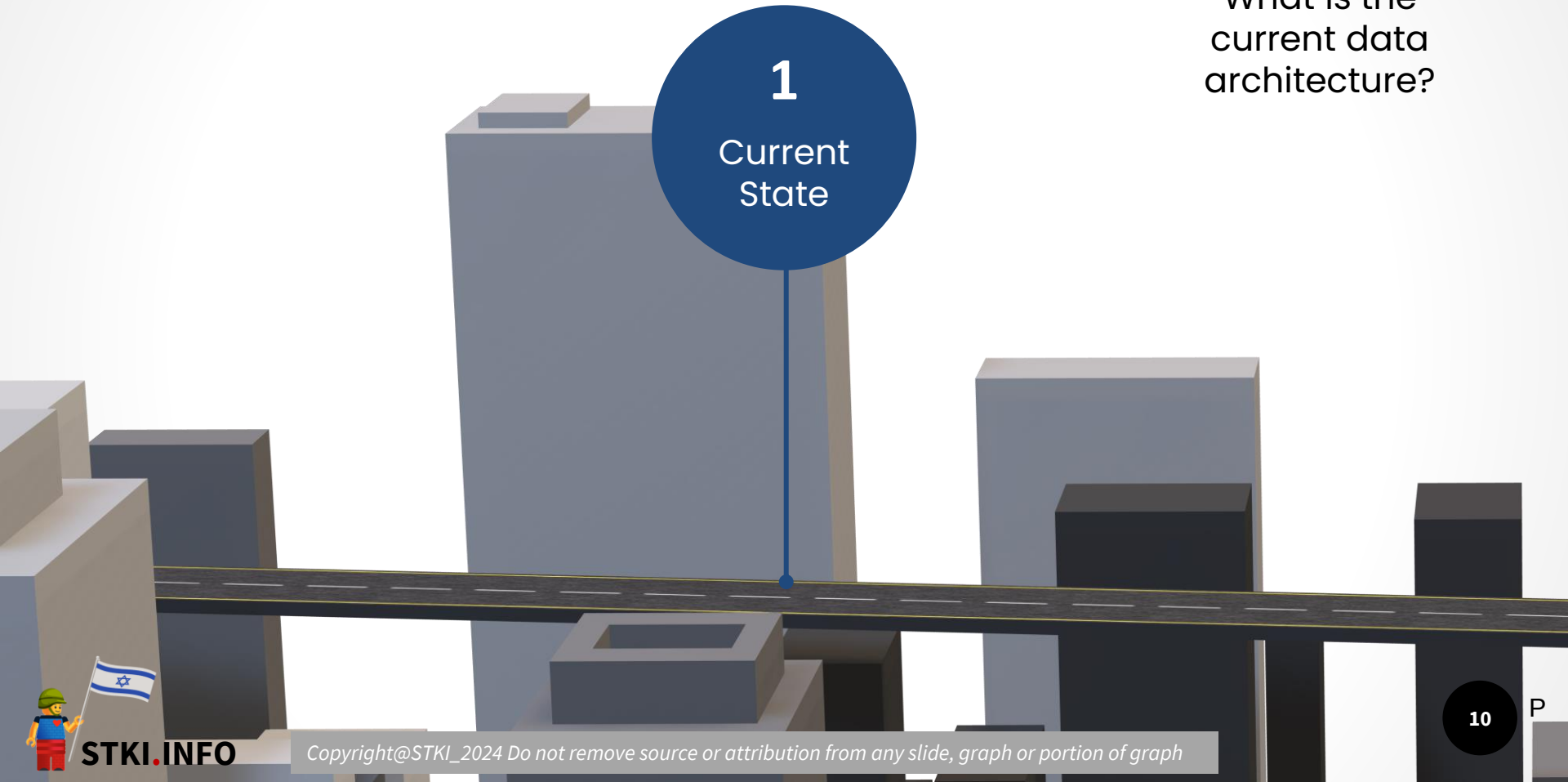
What are the future directions of the AI industry?



So…
Let’s go!

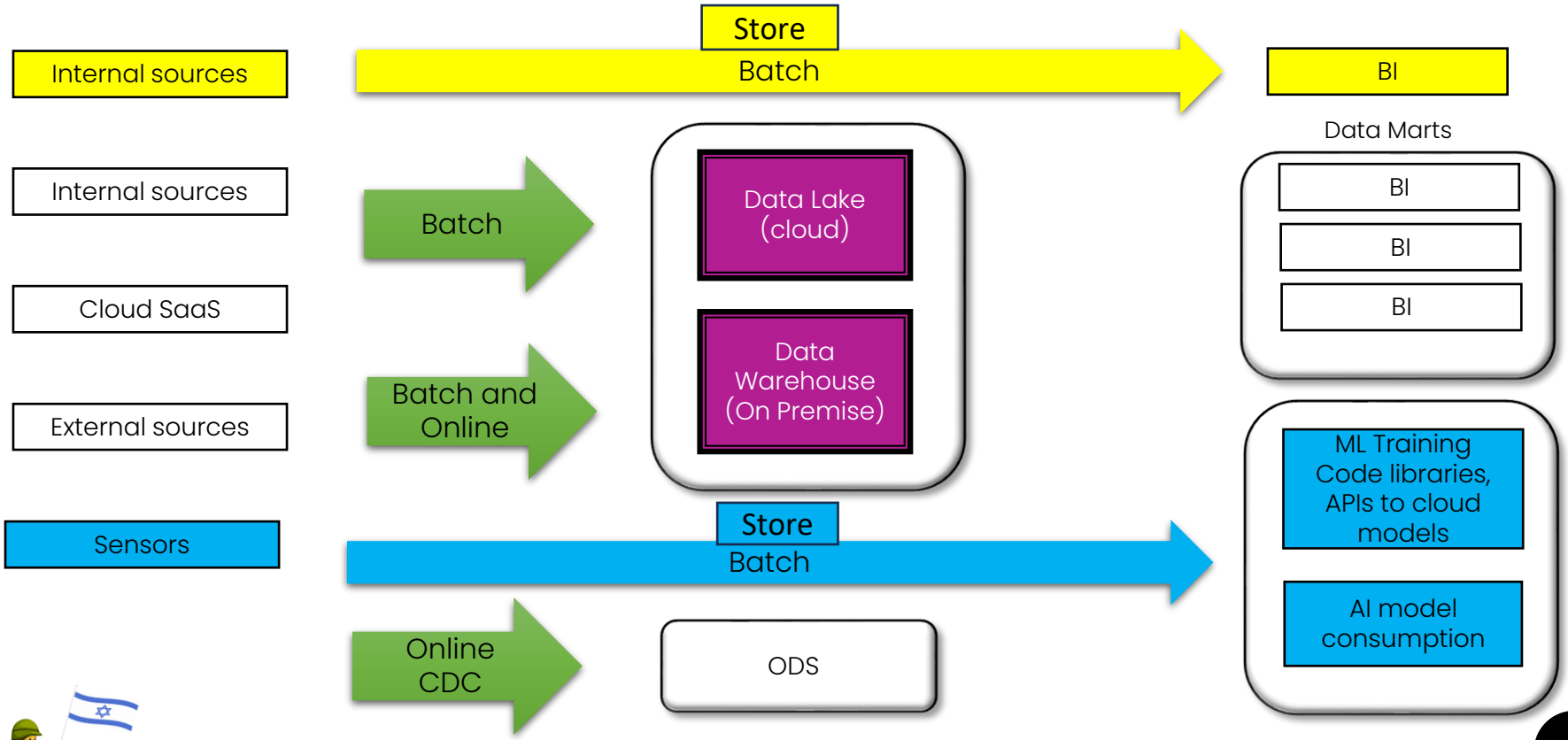


STKI.INFO



What is the current data architecture?

Current Data Analytics Architecture



STKI.INFO

Current Data Analytics Architecture Issues

Data Silos

Inflexibility

Poor Data Quality

Inconsistency and Repetitive Security & Privacy operations-

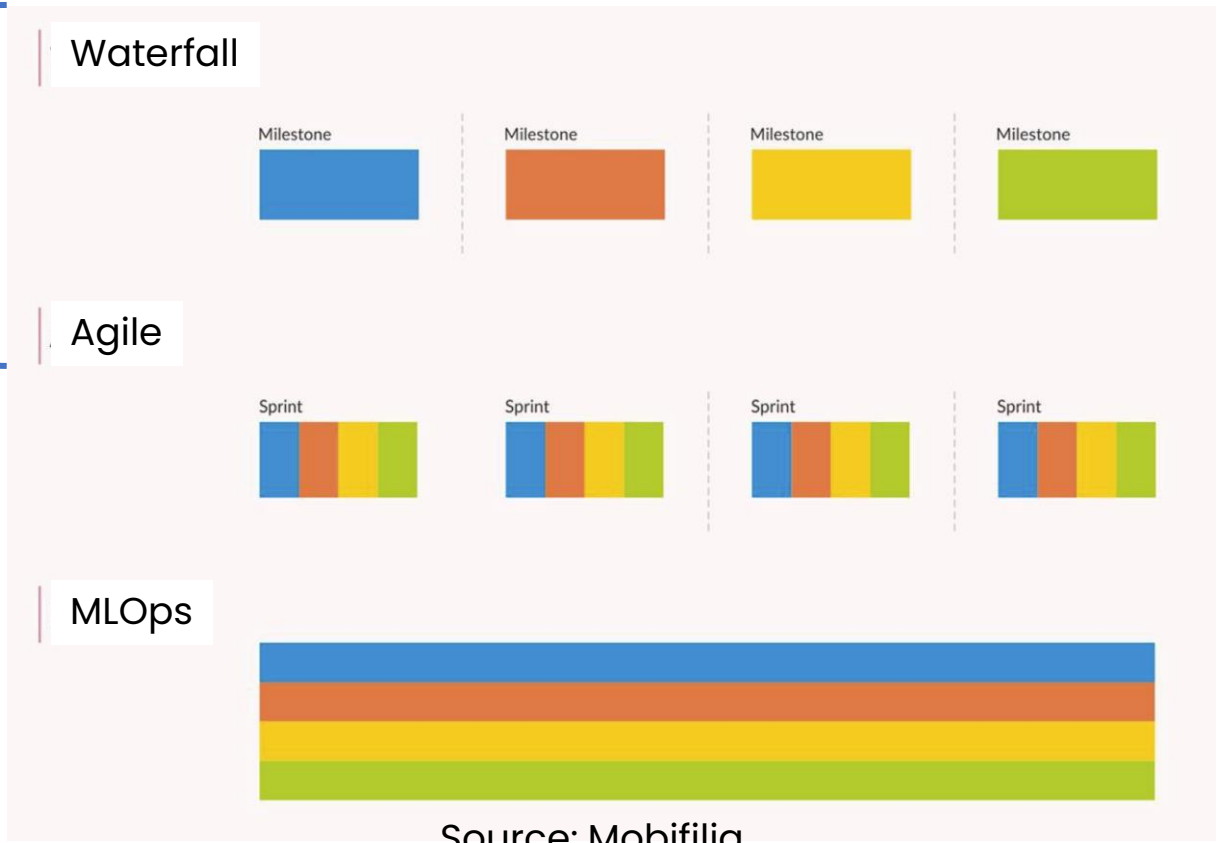
Scaling & Performance

High Costs

STKI.INFO

Which operational model do companies use **today**?

DW and BI



ML models →

Design

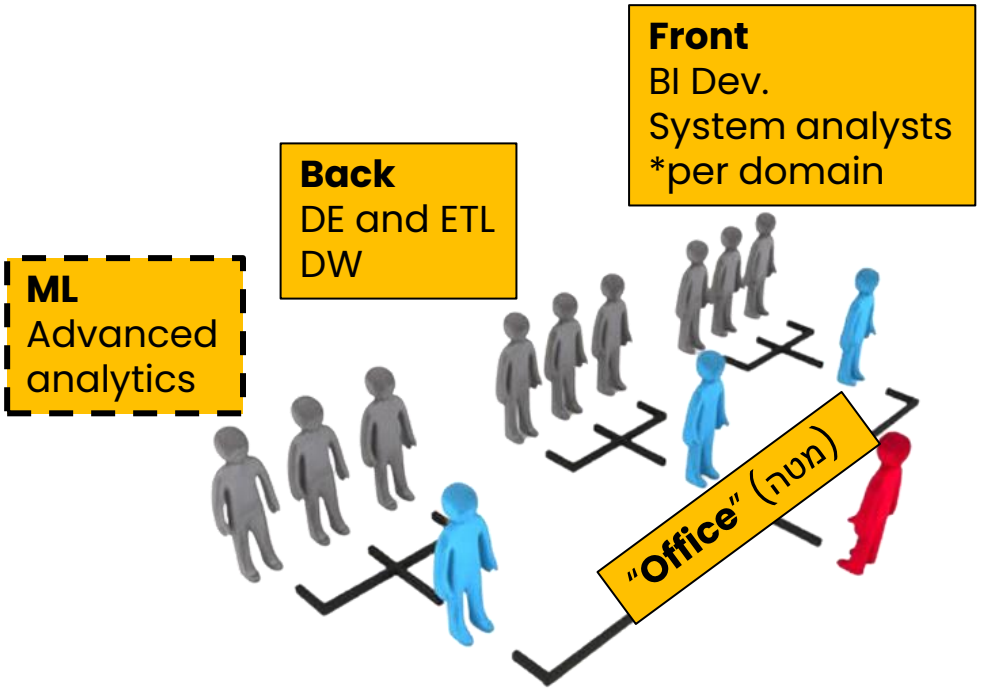
Build

Test

Deploy

Source: Mobifilia

Existing data organizations look like this:



Issues:

- Slow TTM
- Dependency
- Focus on delivery (not discovery)

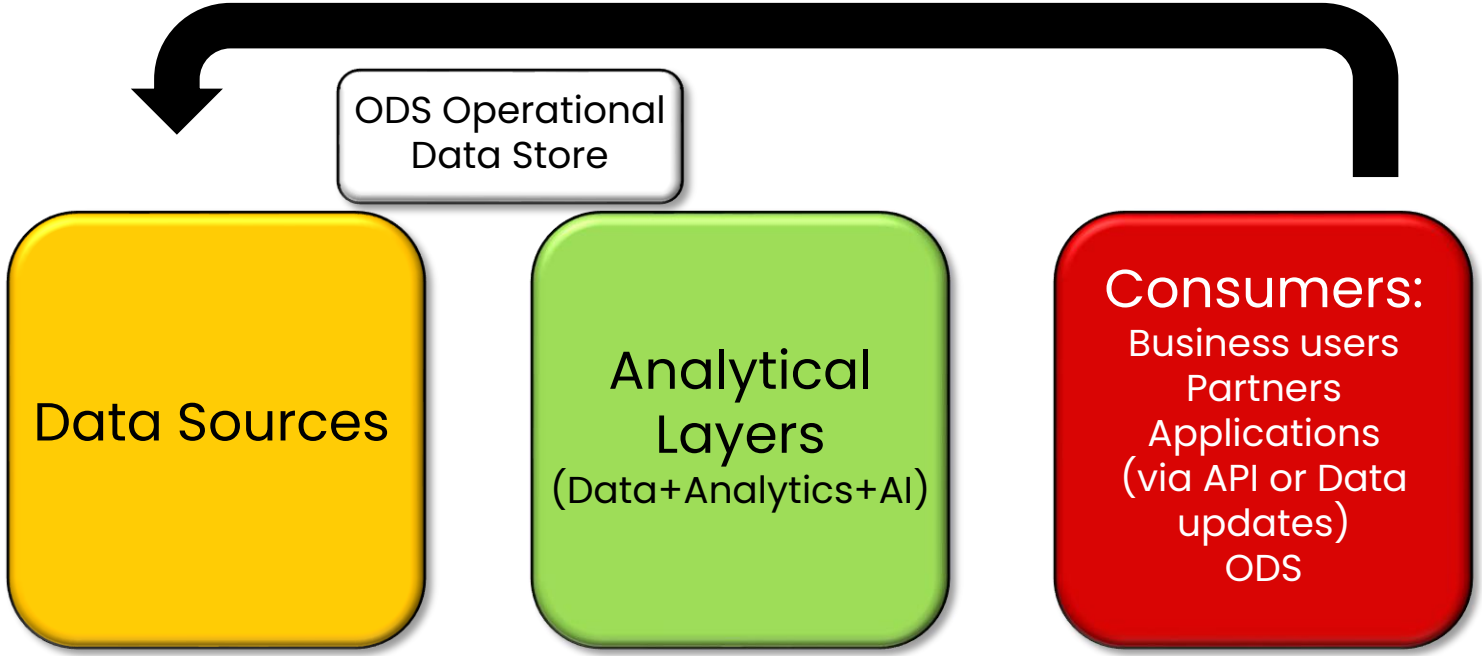
STKI.INFO

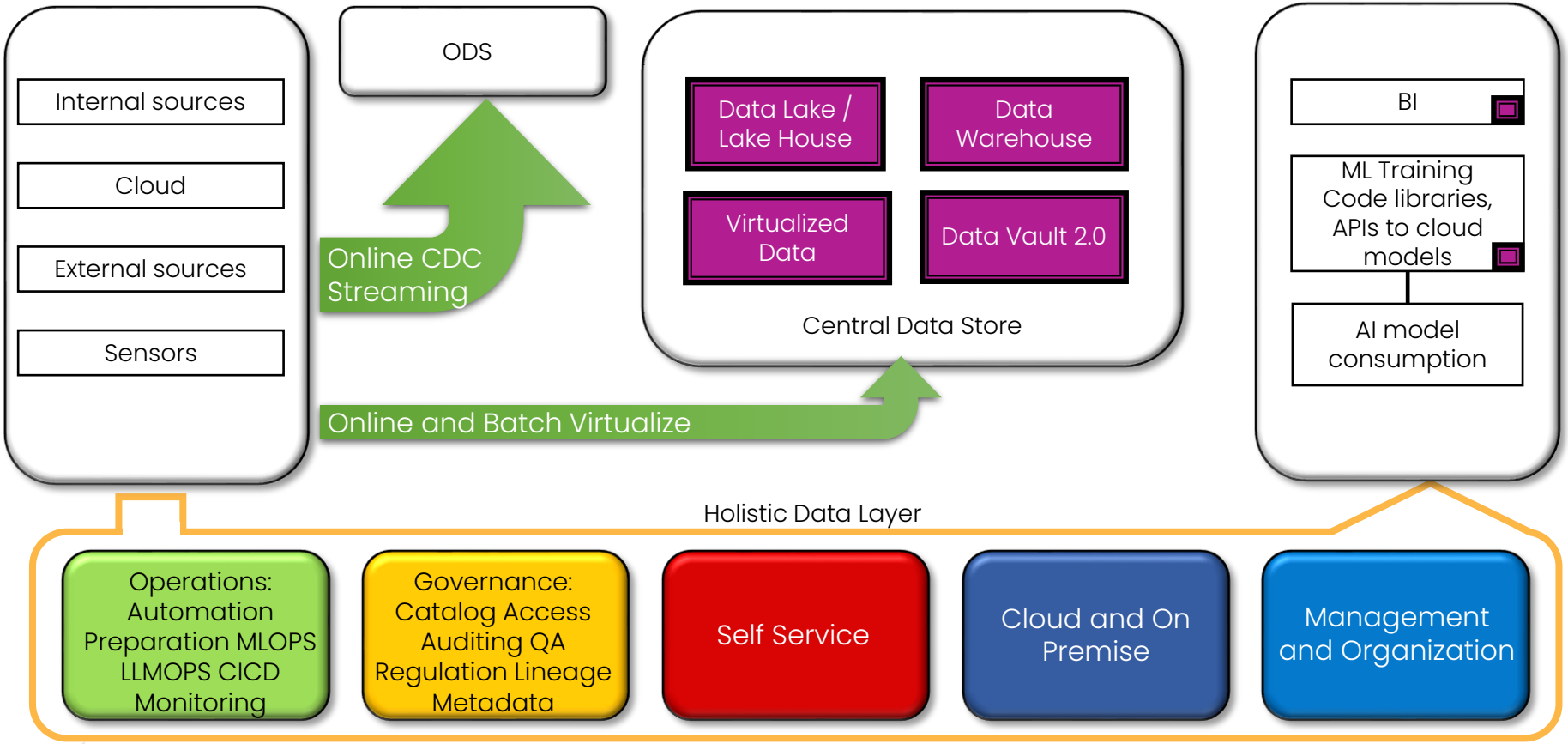
What are the needed architecture, operations and infrastructure to effectively deploy AI?

2
Future Architecture & Tools

What are the different tools categories?

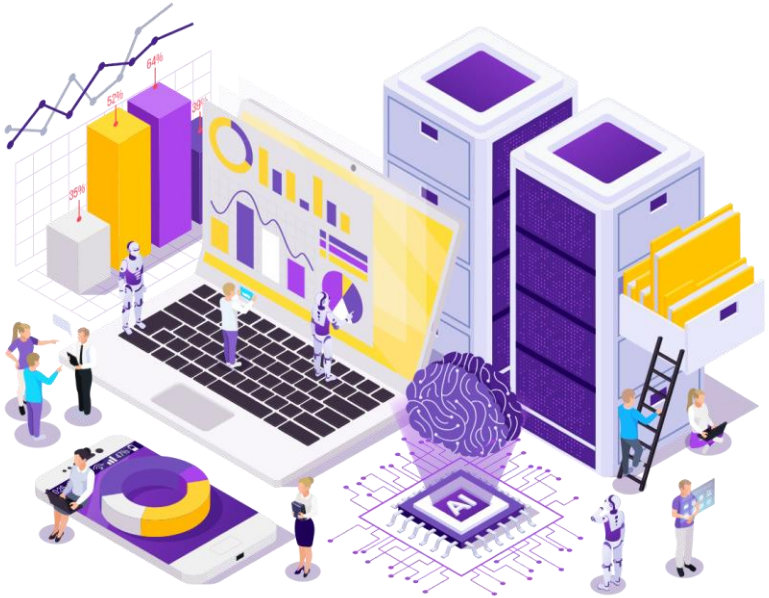
STKI.INFO



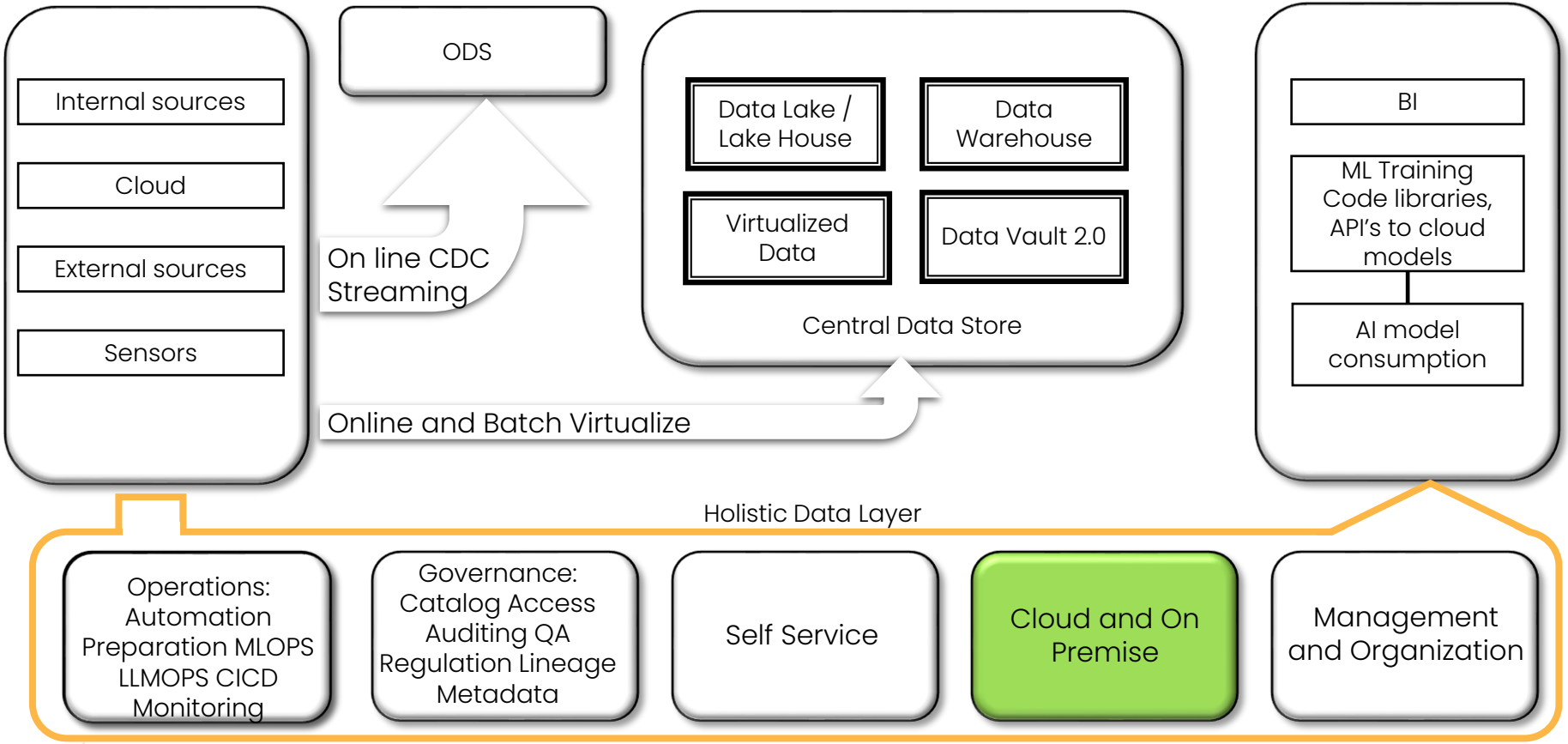


Which tools are needed for your modern data factory?

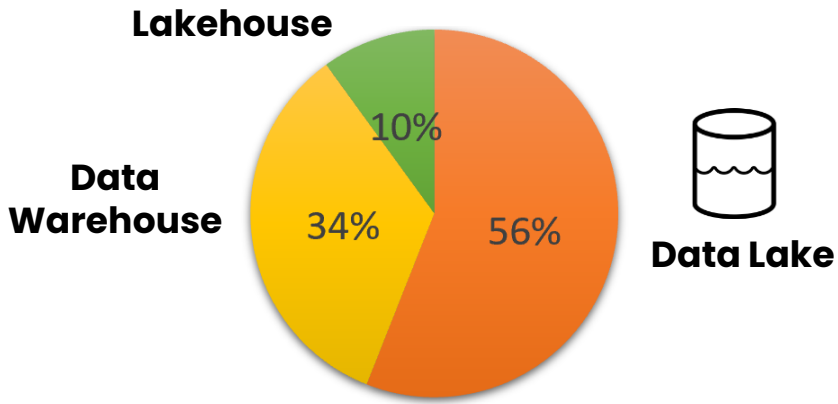
- Data Ingestion and processing-preparation : ETL, Streaming, CDC, code (python)
- Data stores Cloud\On premise (Vector Databases)
- Data virtualization tools (Logical Data)
- Data governance tools : QA, Security, compliance, lineage, data catalog, metadata management



- Self service BI tools – visualization tools
- ML and AI tools
- Data pipeline & orchestration Data Ops. MLOps LLMOPS
- IDP Internal Development Platforms – how will developers access analytics components



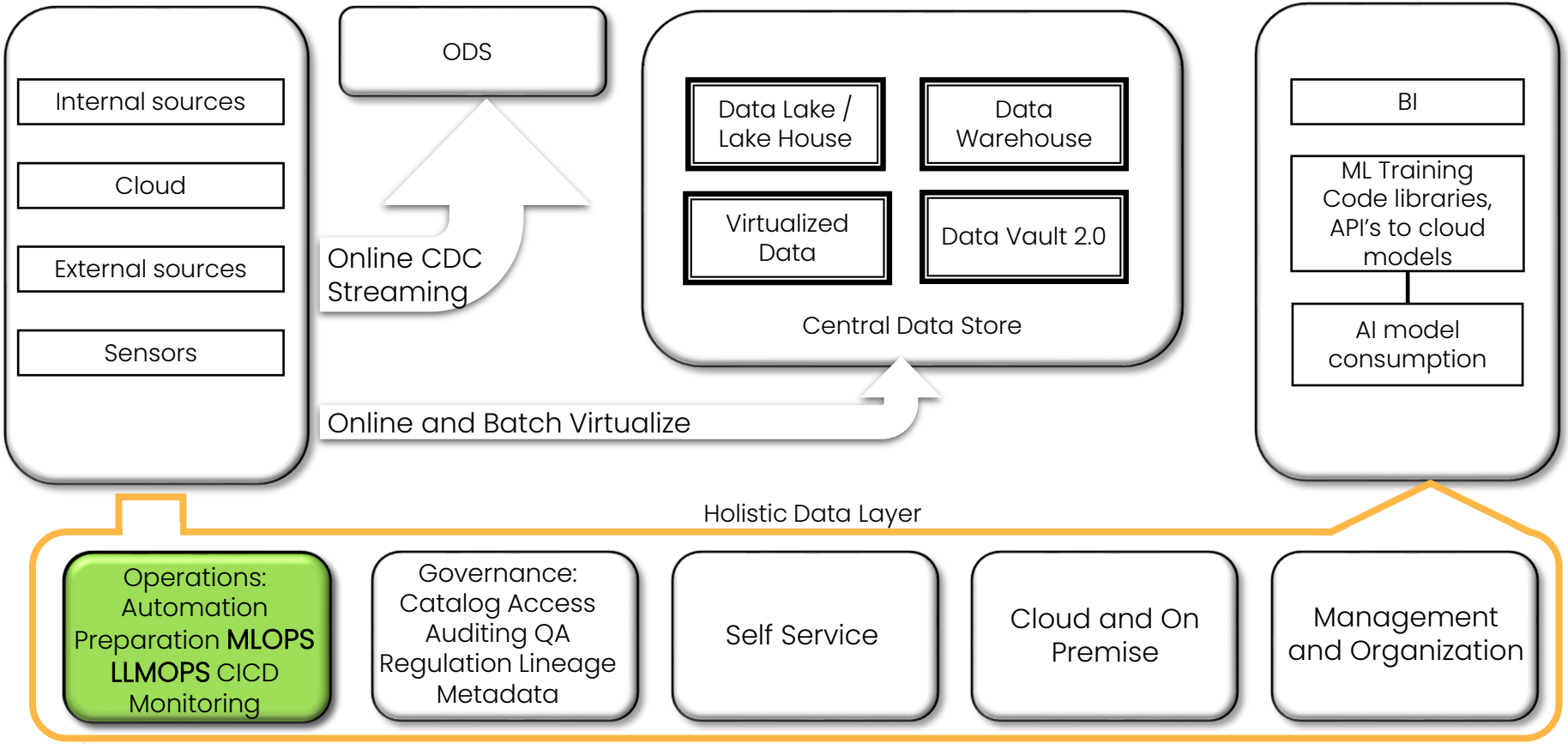
Cloud's leading use case: Data Lake



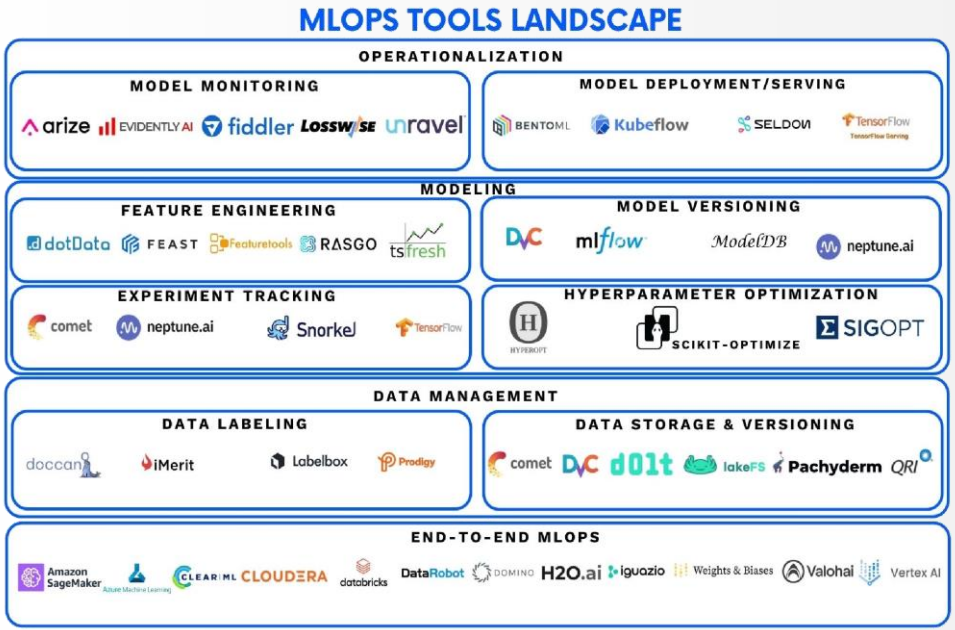
DW remains mostly on premise

34%
עובדים כיום

45%
מתכננים להיכנס לעננים אנליטיים בקרוב



Automation tools - MLOPS main features



https://research.aimultiple.com/mlops-tools/

- Model building automation - model selection, data preparation, feature engineering, model training
- Model registry / versioning
- Automation testing (& AB testing) and experiment management
- Model monitoring logging
- Model governance (risk, compliance)
- Model optimization (GPU / Cloud options)
- Continuous integration and deployment

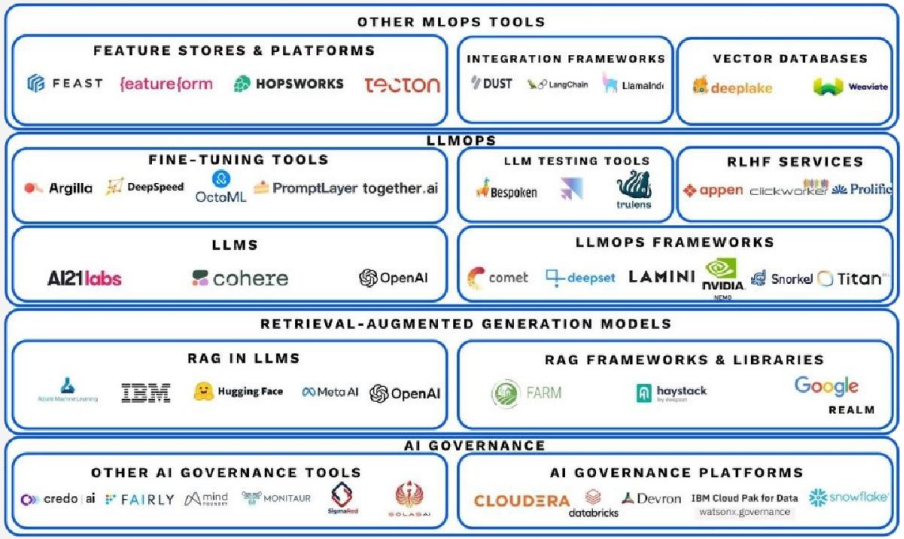
STKI.INFO

Automation tools – LLM Ops

The Operational management of LLMs in production environments. It encompasses the practices, techniques, and tools used to effectively deploy, monitor, and maintain LLMs:

- Data Collection and Preprocessing
- Model Selection and Training
- Evaluation
- Prompt Engineering
- Explanation and Interpretability
- Model Optimization
- Testing and Monitoring
- Deployment and Scaling
- Versioning and Updates
- Ethics and Governance

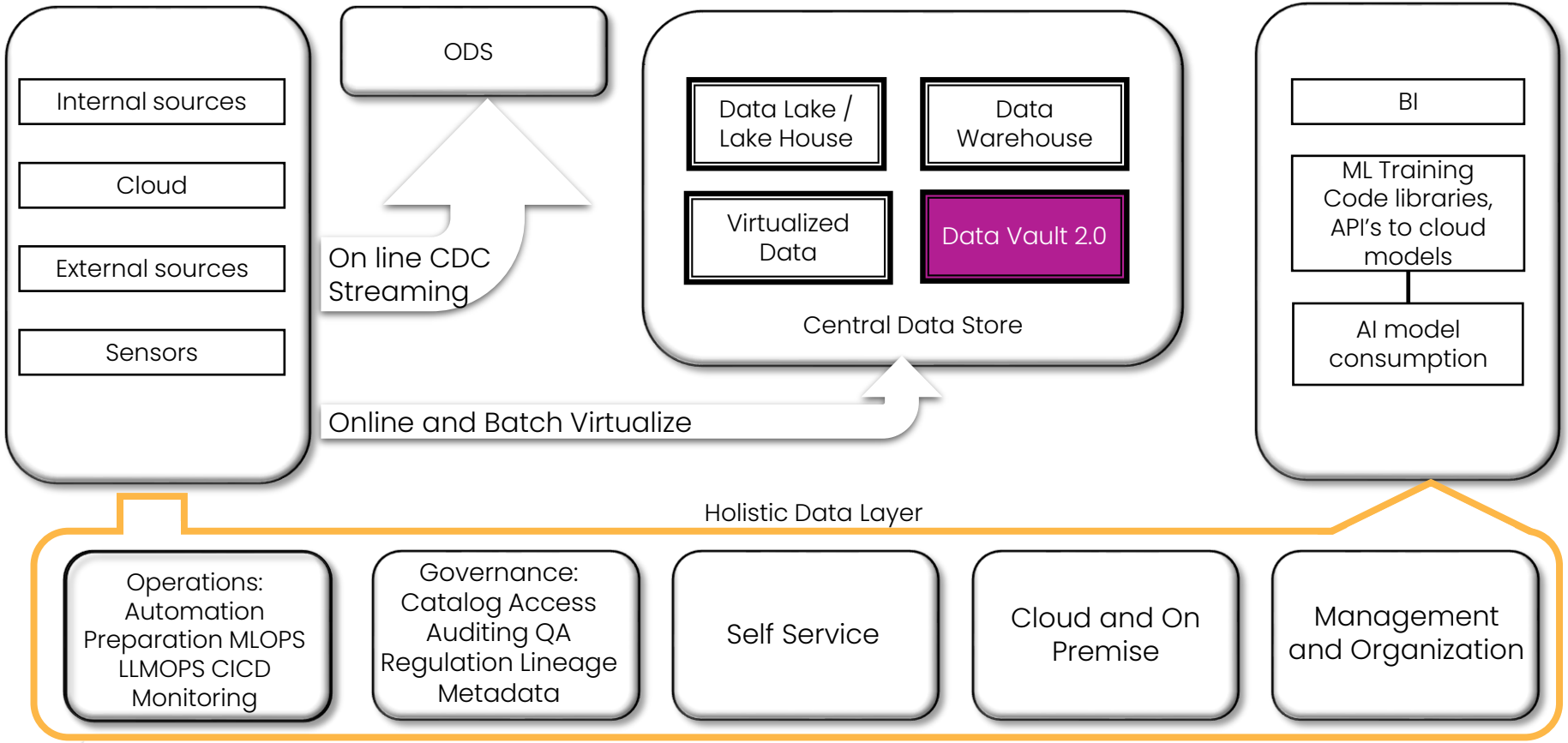
The LLMOPS part from the landscape



https://research.aimultiple.com/mlops-tools/

Differences between MLOPS and LLMOPS

	MLOPS	LLMOP
Maturity of tools and field	Mature	Less Mature
Data Management	Smaller, label-based training data	Massive text corpuses requiring petabyte storage
Training / customizing	Training from scratch or use transfer learning	Use pretrained Foundation Models with lots of customization options
Prompt engineering	N/A	Crafting prompts
Resource requirements	Varies	Extremely large models with high computational needs
Cost optimization	Sometimes important	Crucial
Output filtering	N/A	Block generated toxic text
Rate limiting and query triage (analyzes questions to extract complexity signals)	Sometimes important	Very important



Data Vault 2.0 – the middle road

Data Warehouse

Most of DW is based on Star Scheme

- Advantages:
  - Simplicity and Understandability:
  - Query Performance
- Disadvantages:
  - Data Redundancy
  - Lack of Flexibility

Data Vault 2.0

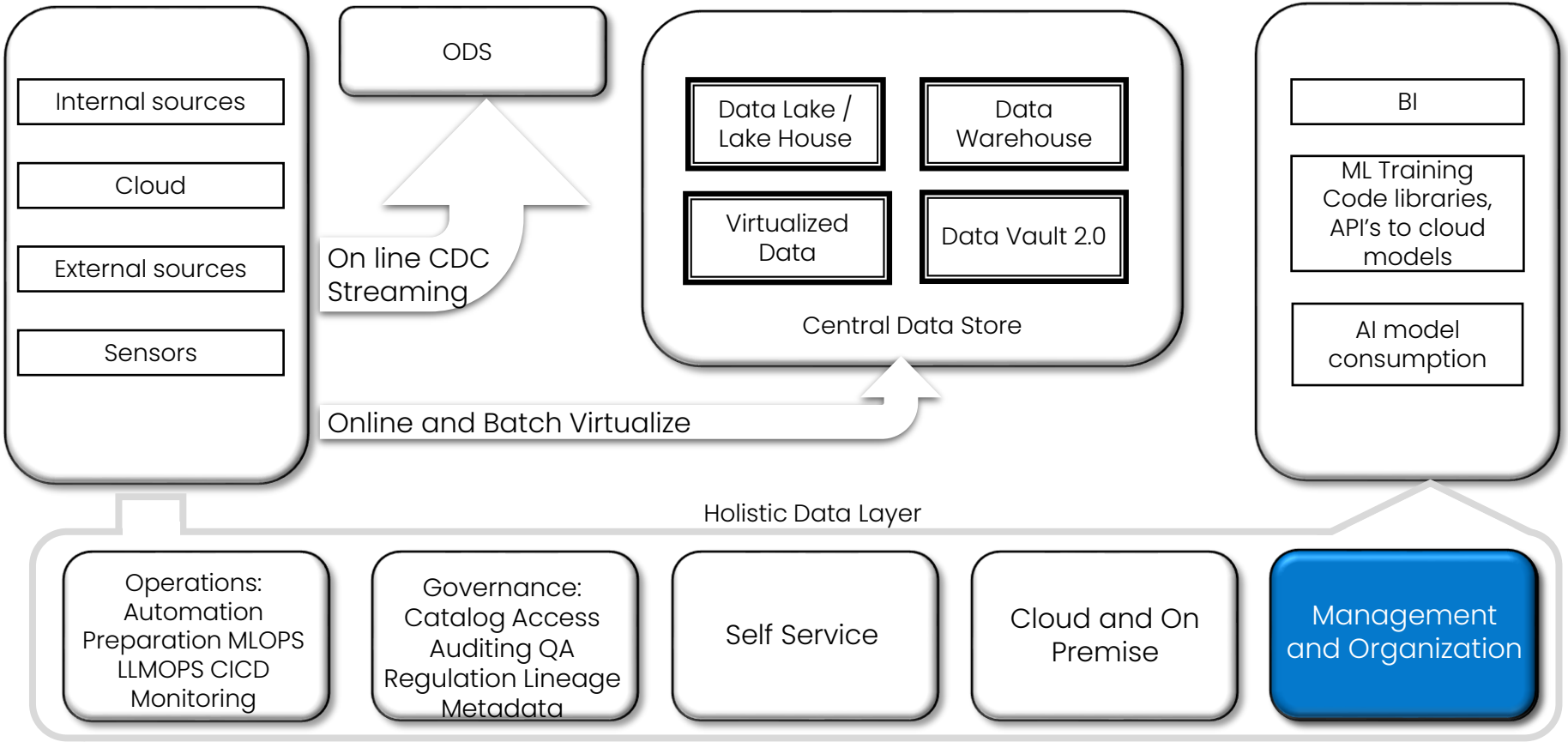
- Enables fast modification to the model
- Every data source is related to the existing model
- Fast uploads
- Primary Technical Keys are generated + source/time of loading

Data Lakes

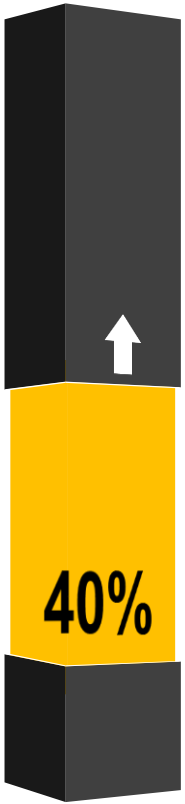
Do it yourself. This makes everything faster. But:

- Data Quality and Governance
- Data Silos and Lack of Integration
- Security and Privacy Concerns
- Data Discoverability
- Performance and Scalability
- Lack of Data Ownership and Accountability
- Becoming “Data Swamp”

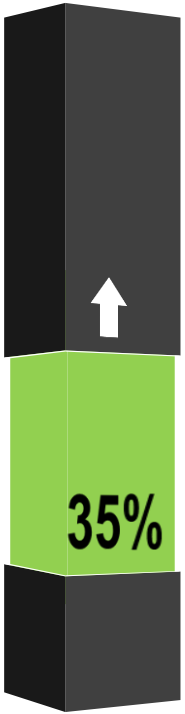
STKI.INFO



Israeli organizations are **just starting** to adopt a product approach for data & analytics



יש התחלה
שימוש בחלק
מהעקרונות



כיום לא
אך מכוונים
לכיוון זה

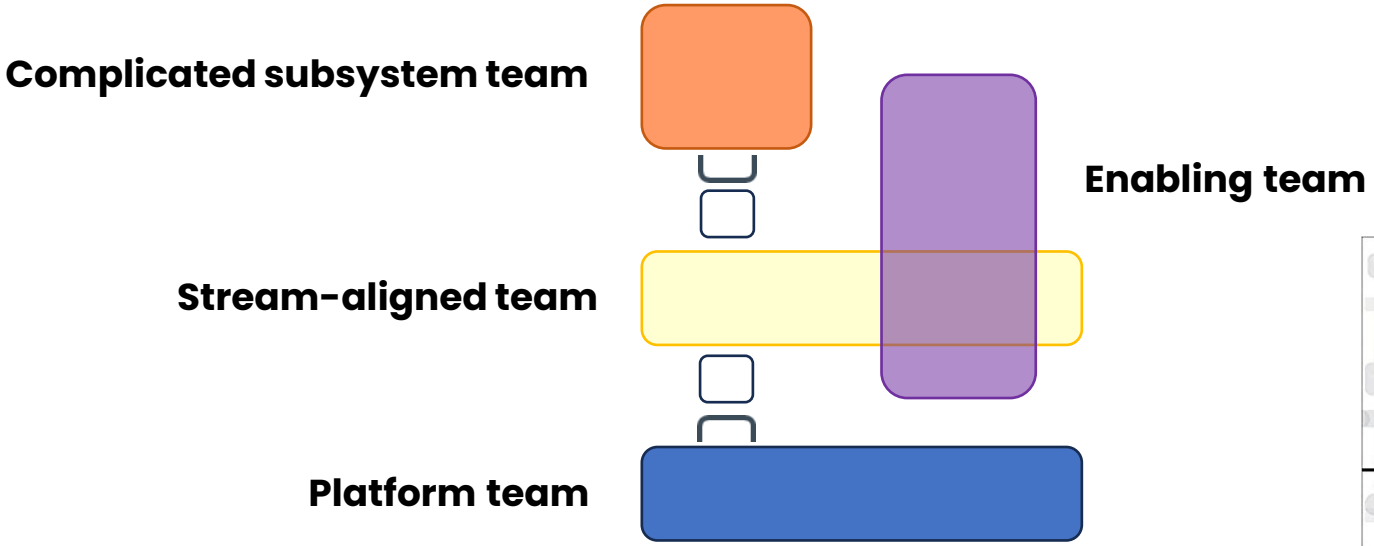
כלל לא
ואין כוונה

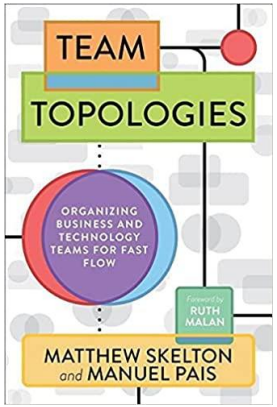
State of Data in Israel
STKI 2023

האם מממשים תפיסת "ניהול מוצר" בעולם הדאטה ואנליטיקה?

STKI.INFO

The Four Fundamental Team Topologies

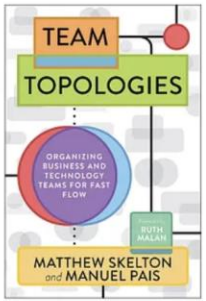




Source: Skelton, Matthew; Pais, Manuel. Team Topologies

Data Product Teams

A Data Product team is a "stream-aligned" team – several other teams including "platform", "enabling" and "complicated subsystem" teams make it easier for Data Product teams to do their work



Team Topologies provides an outstanding description of the various teams involved in delivering software. I use the Team Topologies terminology to describe a Data Mesh / Data Product team structures and roles.

Producer Teams (Stream Aligned Team)
This team produces data that is ingested by the data product team; This team may be a data product team.

Data Product Team
A Data Product team is a Stream Aligned team, supported by enabling, platform, and complicated subsystem teams

Stream Aligned Team
This team is aligned to a flow of work and has end-to-end accountability, and skills needed, for delivery of the work.

Consumer Teams (Stream Aligned Team)
This team consumes data that is managed by the data product team; This team may be team or another data product team.

Platform Teams
These teams provides technical services that accelerate delivery by Stream-aligned teams. This team typically provides "X-as-a-Service" capabilities

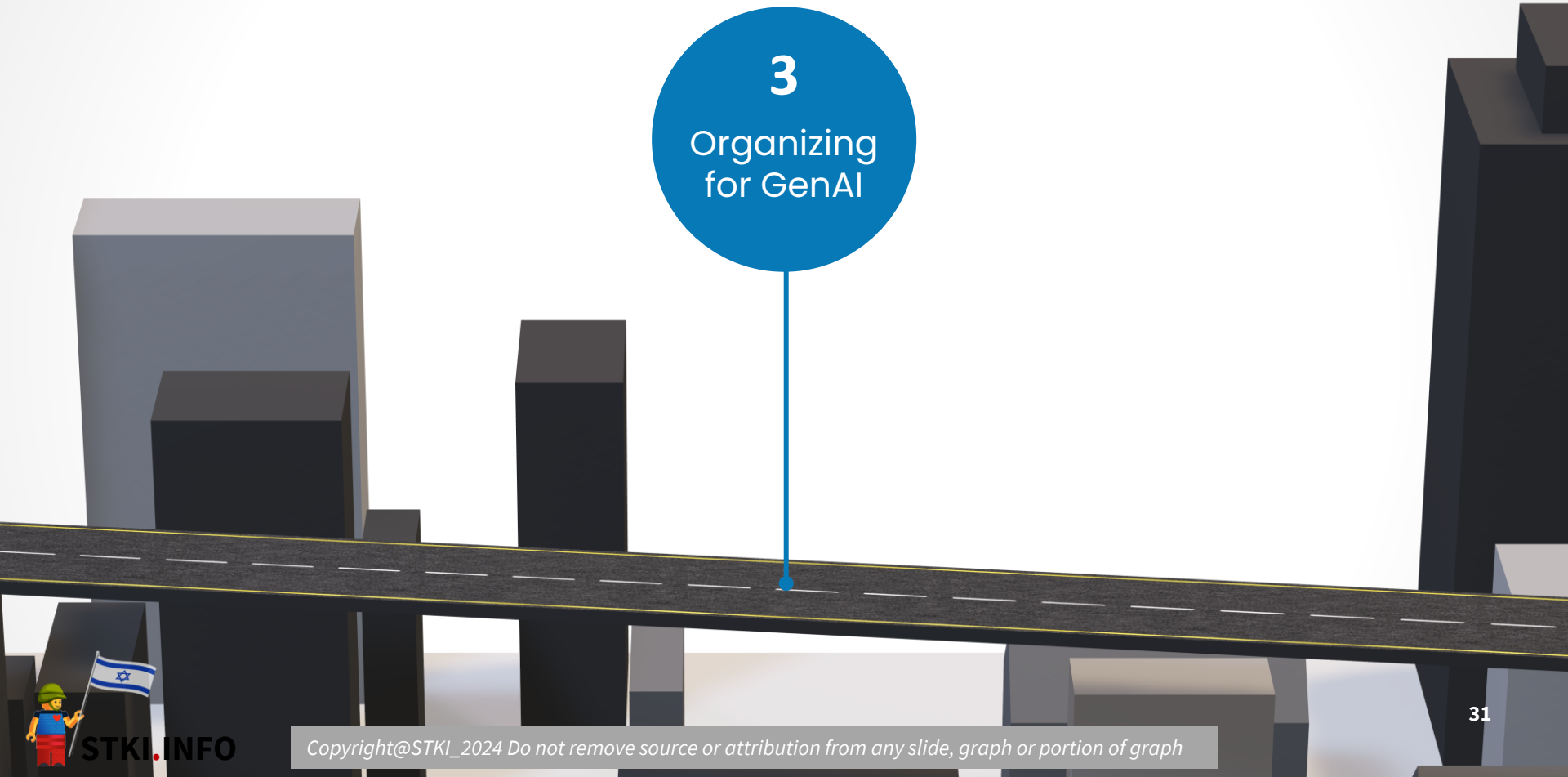
Complicated Subsystems
These team addresses situations where sophisticated capabilities (unique technologies, calculations, or skills) are needed. In many cases, these subsystems are "wrapped" to make them easier for Data Product team consumption

Enabling Teams
These team helps a Stream-aligned team to overcome obstacles. This team acts as a "consulting" team provided short-term services and typically interacts on a collaborative, short-term basis

Figure 1, Data Product Teams

Source: Eric Broda

How should companies organize to leverage GenerativeAI models?



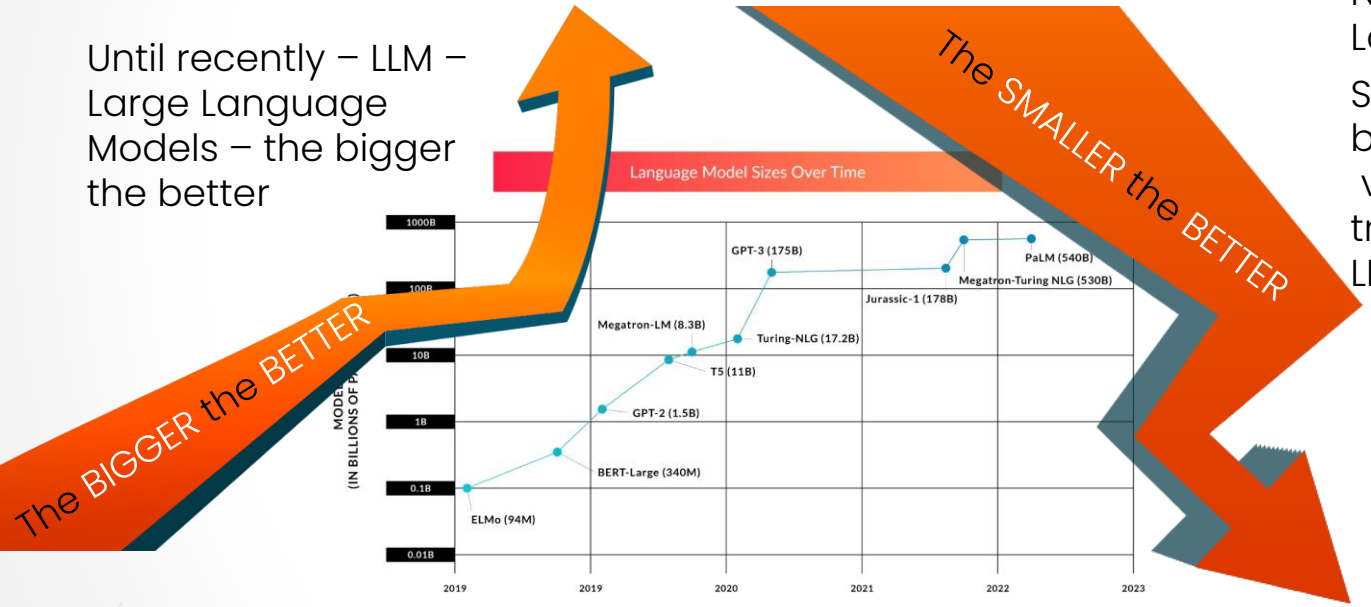

בתאריך 5.2.24 נקיים Webinar
על מונחי בסיס בעולם ה GenAI
Stki.info/events

GenAI is evolving fast – technology is not stable

GenAI model basic characteristics:

- Number of Parameters & number of layers
- Corpus Size the amount of text data the LLM was trained on – measured in tokens

Until recently – LLM – Large Language Models – the bigger the better



Now we see SLM "Small Language Model"

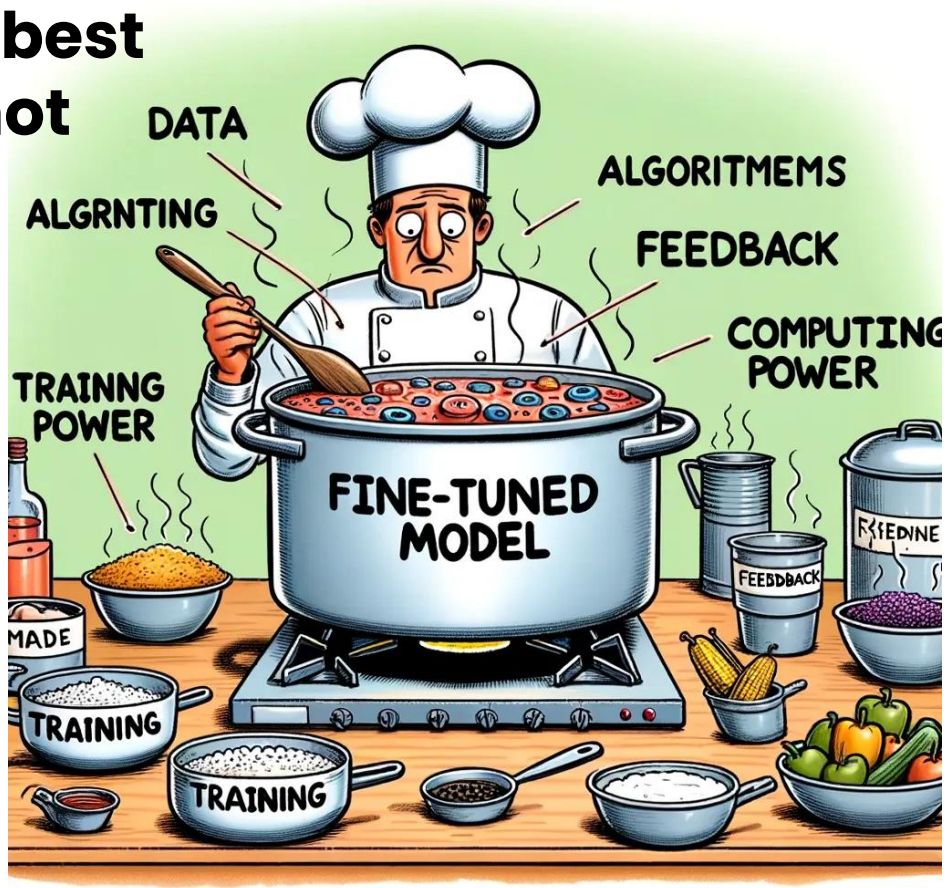
Smaller models with under 1 billion parameters vs. many billions up to trillions of parameters in the LLM's

STKI.INFO

Foundation Model customization best practices are not stable

Prompt Engineering
- Iteratively refining prompts for better results.

Prompt Tuning
(hard and soft tunning) - Optimizing LLM parameters based on specific prompts and outcomes.



Fine Tuning-
Retraining the LLM on a small, task-specific dataset.

Parameter Tuning

RAG - Retrieval Augmented Generation

Retraining-
Training the LLM from scratch on a new, possibly large dataset

GenAI and your spending – It's complicated

Training a model:

- According to unverified information leaks, GPT-4 was trained on about **25,000 Nvidia A100 GPUs for 90–100 days**
- Estimation is about **3,125 servers * 324K$ per server** were needed for total of **1B$**
- Estimation of training costs for GPT-4 was **around $63 million.**

Soon we will need AIFinOps practitioners

Using a model (Inference)

Mistral Open Source model summarizing text

	A10 24GB GPU (1500 input + 100 output tokens)	A100 40GB GPU (1500 input + 100 output tokens)
Best Latency	4.1 sec	2 sec
Best Throughput (Max RPS without significantly hurting latency)	0.9	3.6
Latency (at max RPS) (p50 / p90)	5.8 sec / 5.8 sec	4.7 sec / 4.7 sec
Machine Type	AWS (g5.xlarge)	GCP (a2-highgpu-1g)
Cost per hour (Spot)	$0.30	$1.45
Cost per hour (On-demand)	$1	$3.50

Source: https://www.linkedin.com/pulse/7-lartsim-30b-benchmarks-truefoundry-uj9bc/

STKI.INFO

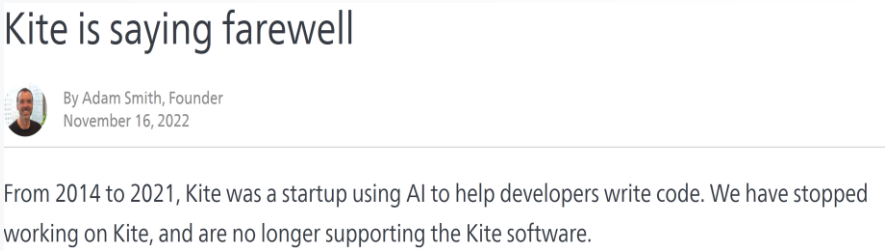
GenAI issues:

- LLM's have tendency to generate plausible-sounding but false or nonsensical statements
- *Hallucinations type*:
  - *Incorrect facts "was born in 2001"*
  - *Fabricated information "use this URL" "this company is doing ..."*
  - *Weird or Creepy Answers "I am in love with New-York"*

https://builtin.com/artificial-intelligence/ai-hallucination

Example of hallucination

Question to LLM : give list of products for code generation?
Answer: …. **Kite:** Provides intelligent code completions, documentation lookup, and on-the-fly code examples for Python, JavaScript, and other languages



Kite is saying farewell

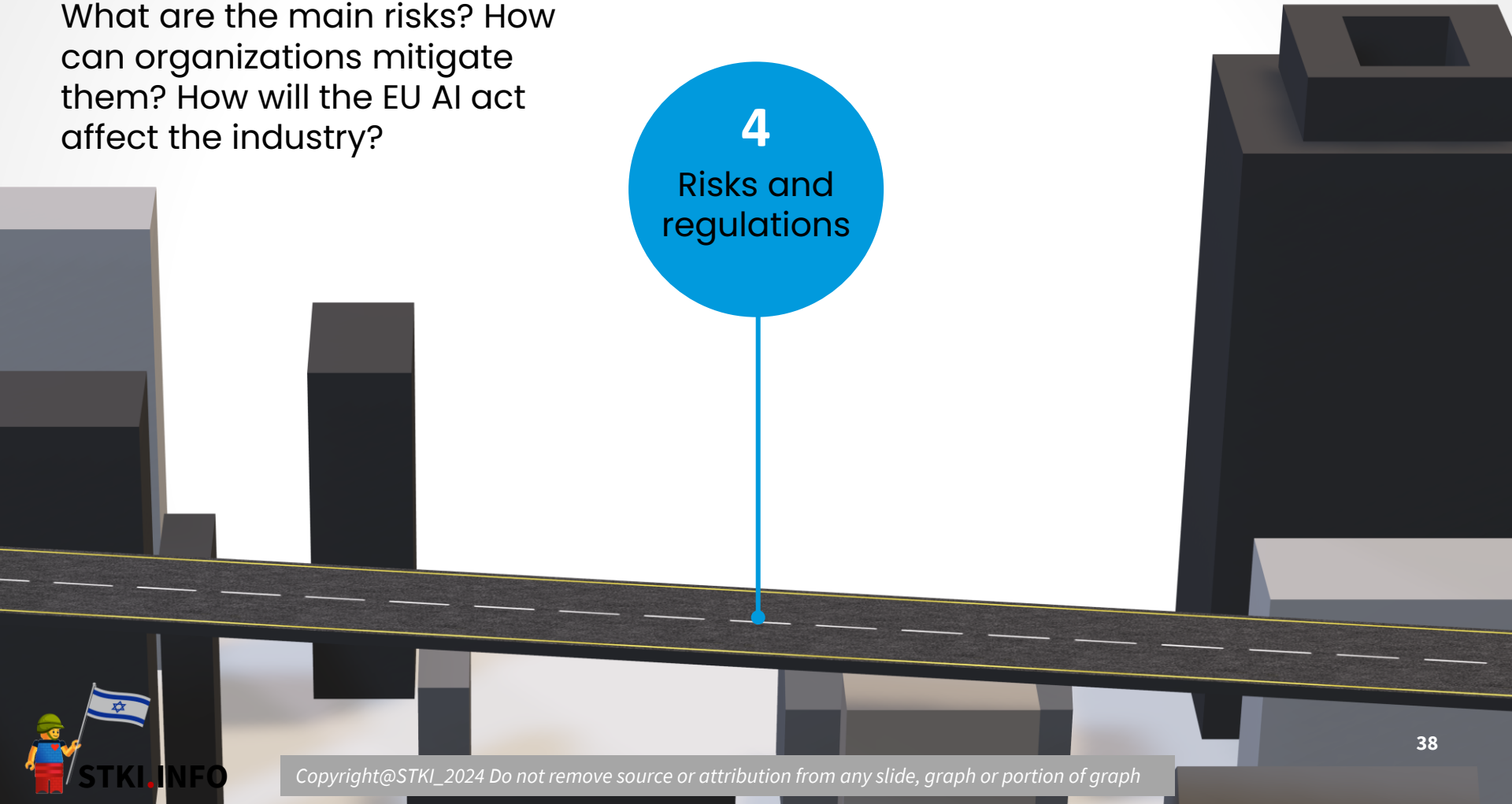
By Adam Smith, Founder
November 16, 2022

From 2014 to 2021, Kite was a startup using AI to help developers write code. We have stopped working on Kite, and are no longer supporting the Kite software.

What are the main risks? How can organizations mitigate them? How will the EU AI act affect the industry?

4

Risks and regulations



STKI.INFO

What is the risk for enterprises using AI?

IP and copyrights violations in the model recommendations , data trained by the model, Who has the IP rights for content created with AI help

Security – technology
Malware inside AI output (especially code)
Information is leaking (credentials, company secrets)

Security – business
Bad/harmful Recommendations /business actions caused by AI hallucinations or by poisoned trained data

Privacy & regulatory
Toxic text generated , Data leakage of PII, Algorithmic discriminations

All the above in your supply chain

EU AI Act

Highest Fines
(up to €40 million
or 7% of global)

High-risk AI obligations
effective: June 2025

Prohibitions effective:
December 2024

Most provisions
effective: Mid-2026

Formal adoption:
Mid-2024

STKI.INFO

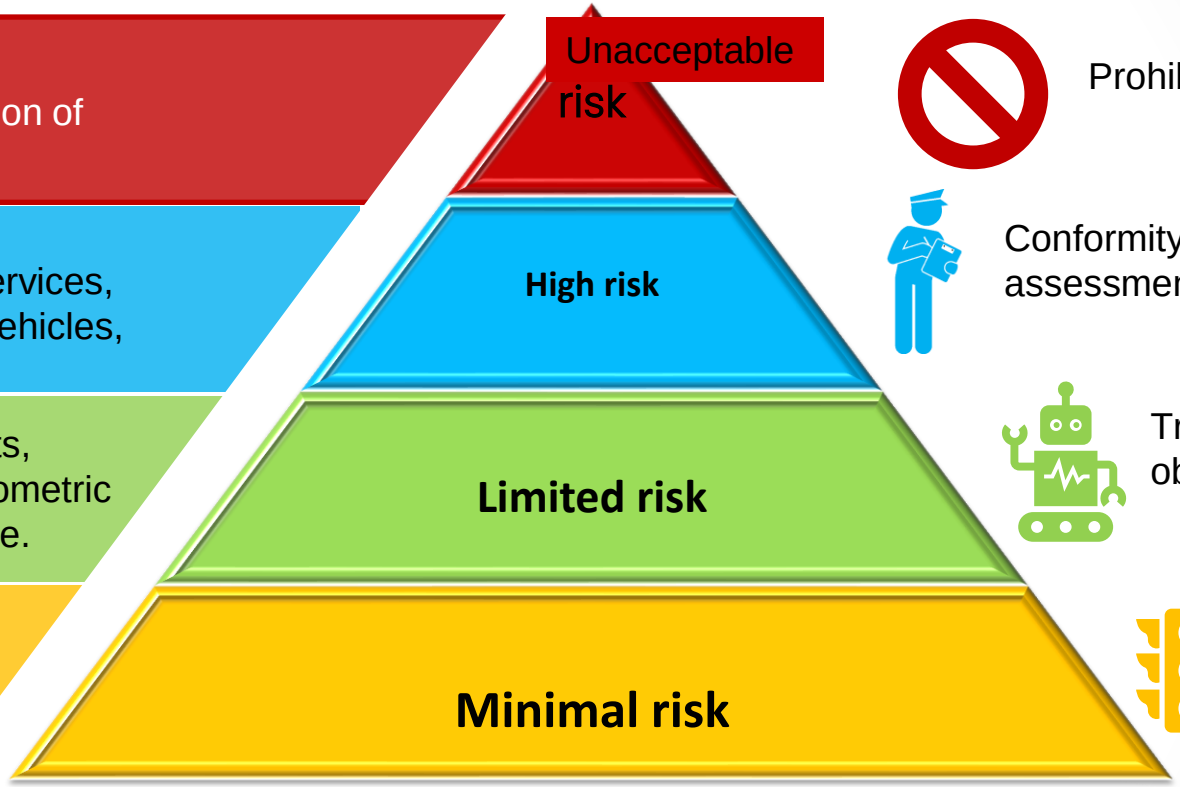
EU Artificial Intelligence Act: Risk levels

Social scoring, mass surveillance, manipulation of behavior causing harm

Access to employment, education and public services, safety components of vehicles, low enforcement, etc.

Impersonation, Chatbots, emotion recognition, biometric categorization deep fake.

Remaining



Prohibited

Conformity assessment

Transparency obligation

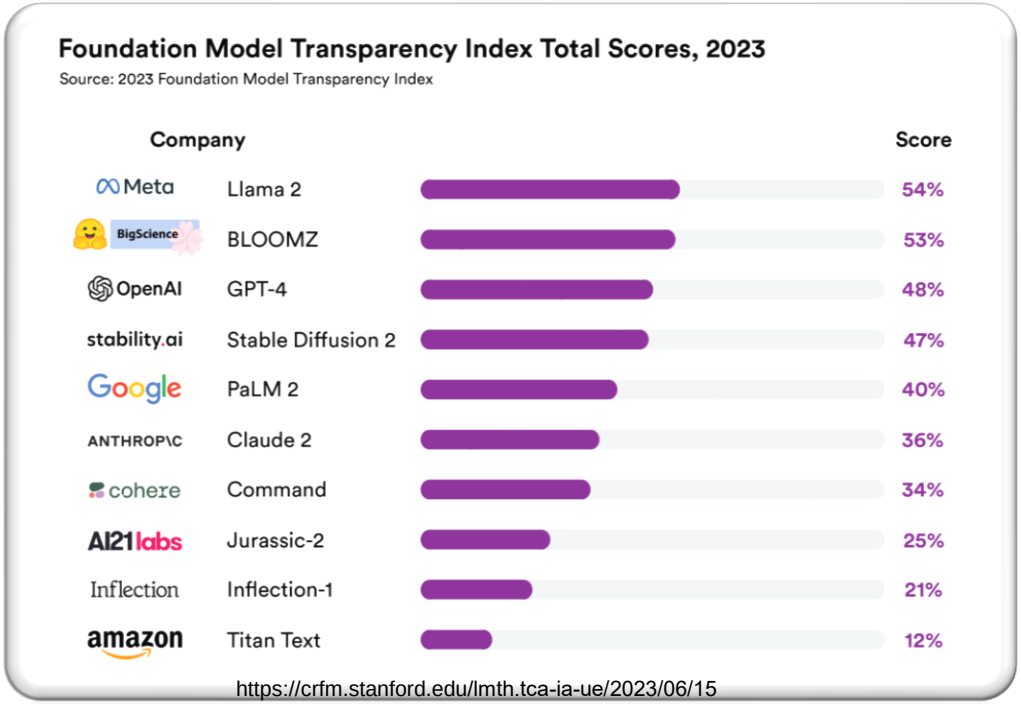
No obligation

Which GenAI models/service currently comply with the AI UI ACT?

No existing (leading) models can claim full legal compliance to the EU AI Act yet*

"*No foundation model provider achieves a perfect score"

Stanford University
Human-Centered
Artificial Intelligence



https://crfm.stanford.edu/lmth.tca-ia-ue/2023/06/15

STKI.INFO

GenAI is helping the big companies to become bigger

Foundation model training is very expensive

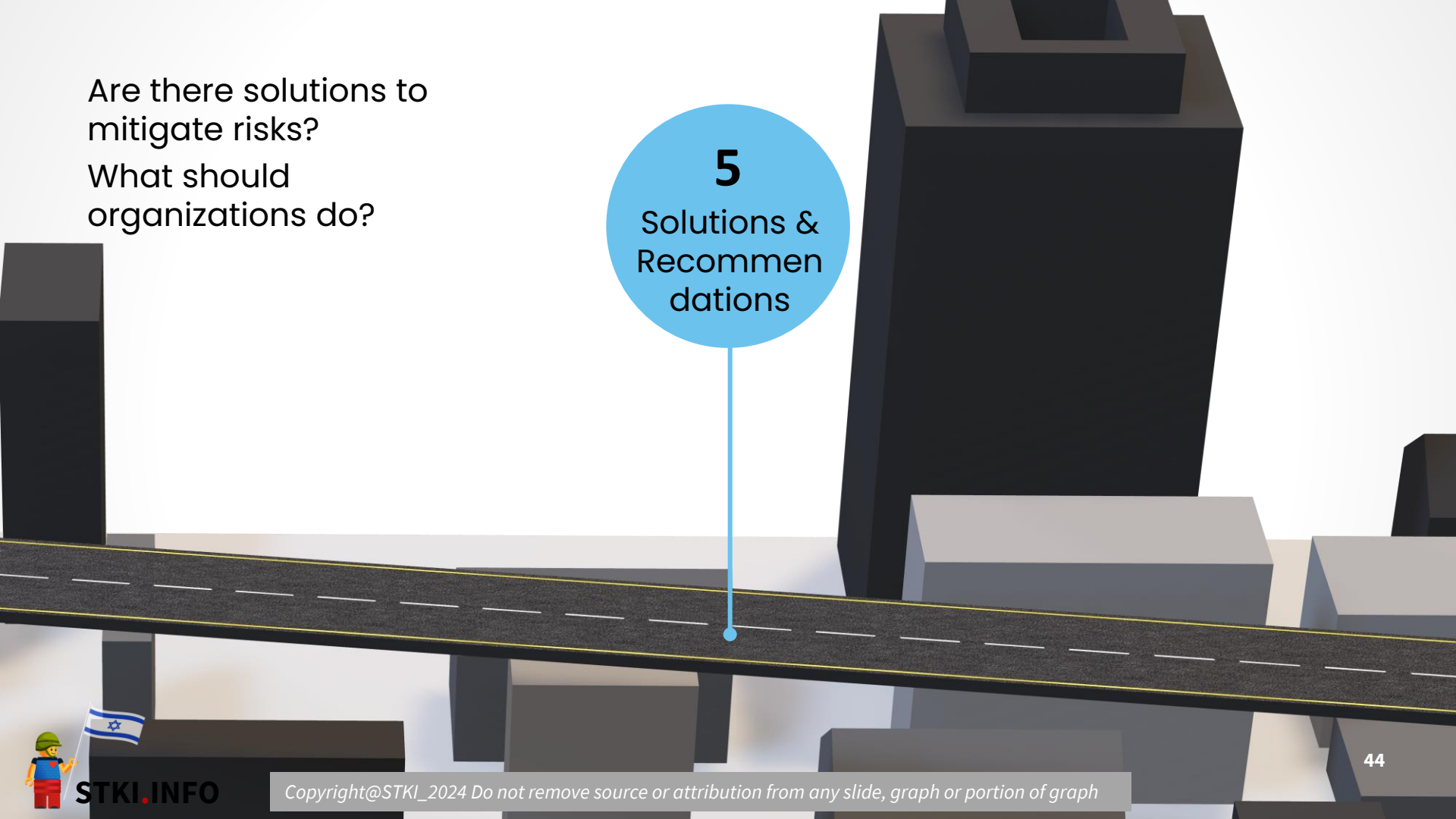
Customization models is better when you have more data

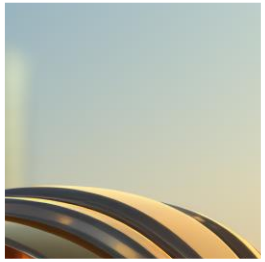
Are there solutions to mitigate risks?

What should organizations do?

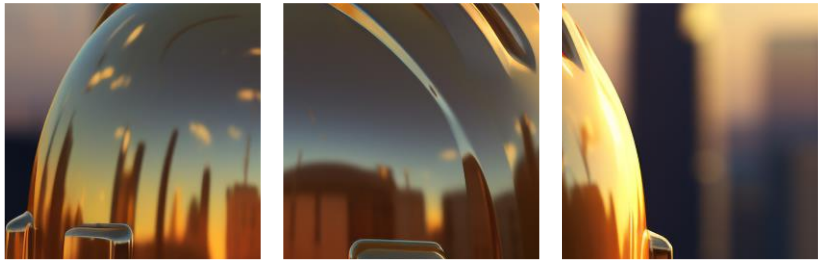
5

Solutions & Recommendations

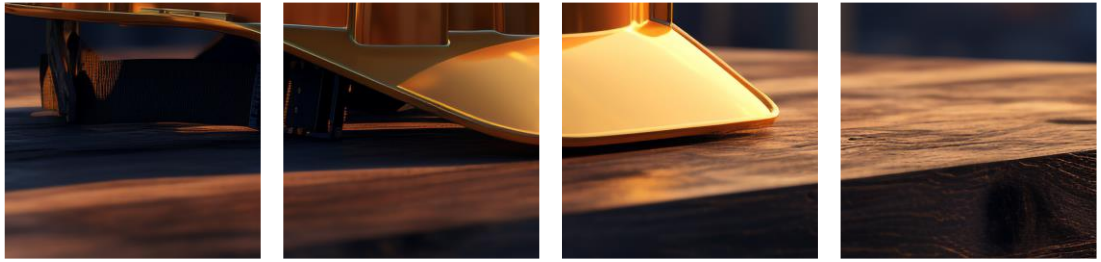




Tools / options to mitigate risks



- LLMOps capabilities (discrimination /bias engines / toxic text blocking, etc.)
- DLP when prompting or when using with API
- Compliance / security management tools for Opensource



- Use AI risk tools when available
- “Enterprise Edition” cloud solutions
- On premise installations

STKI.INFO

Recommendations for enterprises regarding AI risks

Evaluate their risks in the different vectors and set their policy to the different activities & domains

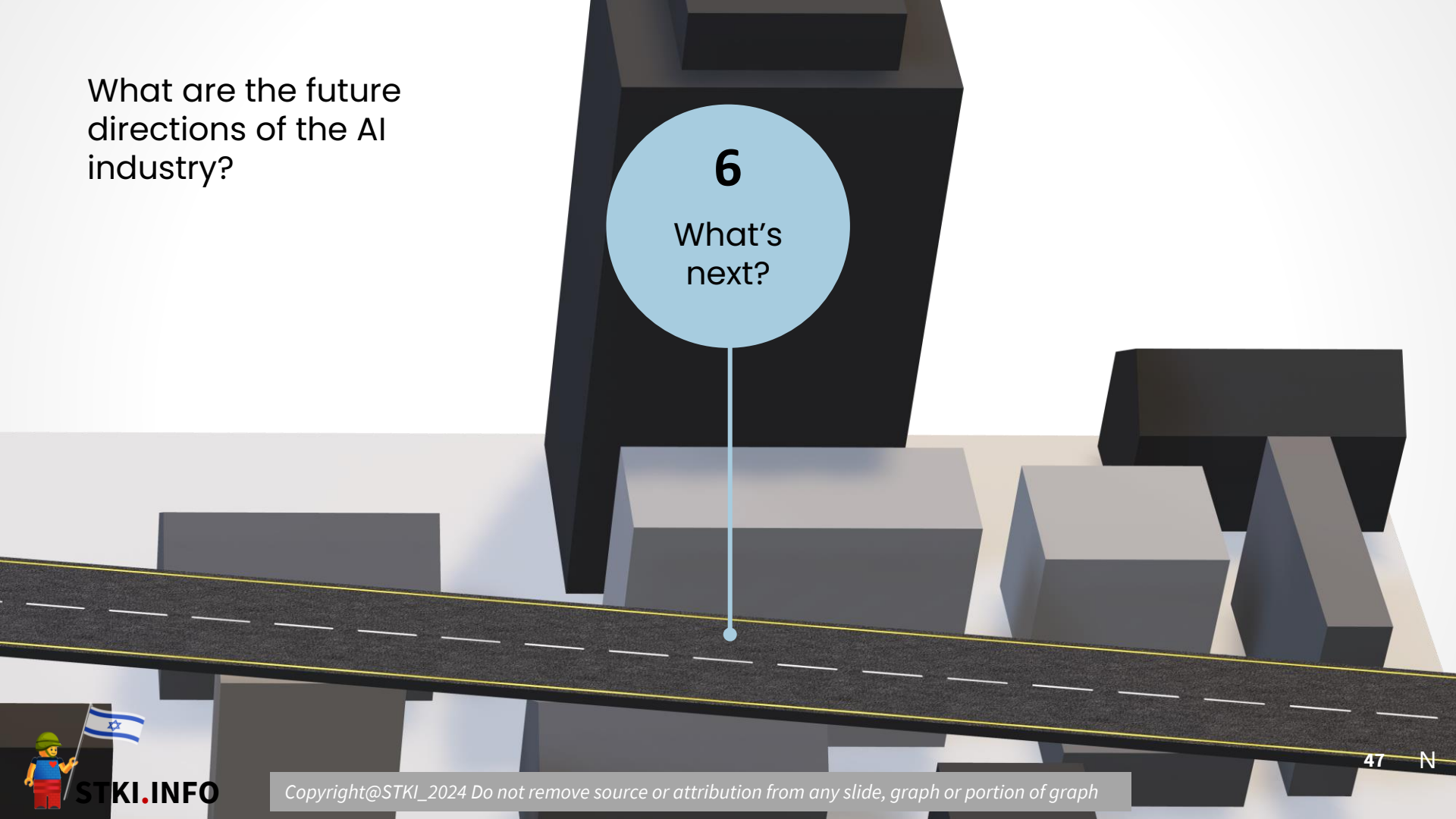
Enhance the security awareness program with AI related topics

Implement AI security tools when available

For most of the enterprises the “enterprise edition” offering might be good enough

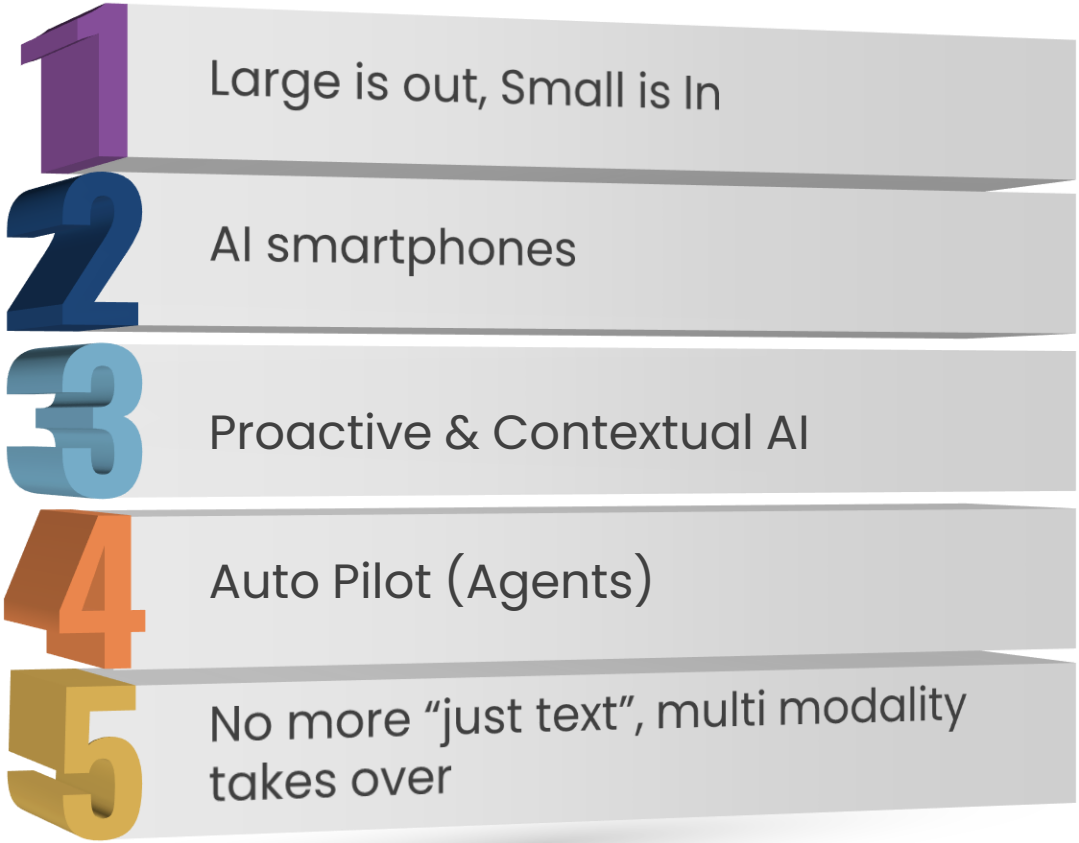
Evaluate AI risks and solutions every 6 months

What are the future directions of the AI industry?



STKI.INFO

A glimpse to the AI future



STKI.INFO

Beware of the **“shiny object syndrome”**

There is no substitute for:

A. Having a clear business case
B. Having trustworthy data
C. Continuous discovery

STKI.INFO

thank you

Good luck with planning your future

AI architecture & operations

STKI.INFO