Navigating the AI Journey:
From Proof-of-Concept to Scale

AI FACTORY AUSTRIA

Speaker: Endri Deliu (AI:AT)

Panelists:
- Laurenz Hinterholzer (Aximote)
- Philipp Heideker (Sleak.ai)

AI:AT Services

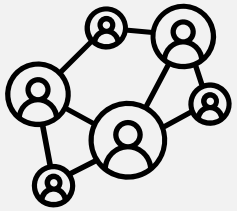
AI:AT Offerings

- Seamless Onboarding & Central Guidance
- Sovereign, High-Performance AI Infrastructure
- End-to-End AI Lifecycle Support
- Integrated, Secure Data & Software Ecosystem
- Startup & Innovation Acceleration
- AI Talent Development & Training
- Ethics, Compliance & Legal Support
- Ecosystem Connectivity & Strategic Alignment

Why do we need AI Factory Austria?

Sovereignty

Ethics and Trustworthiness



Connecting the Ecosystem

Innovation Domains

Core Areas	Biotech	Industry Manufacturing	Public Administration	Physics
Additional Areas	Health	Fintech Lawtech	Environment & Sustainability	Other

AI:AT Network

Users

- Companies
- Public Users
- Projects

AI Factory

- AI Factory Infrastructure
- AI Factory Hub

Partner Network

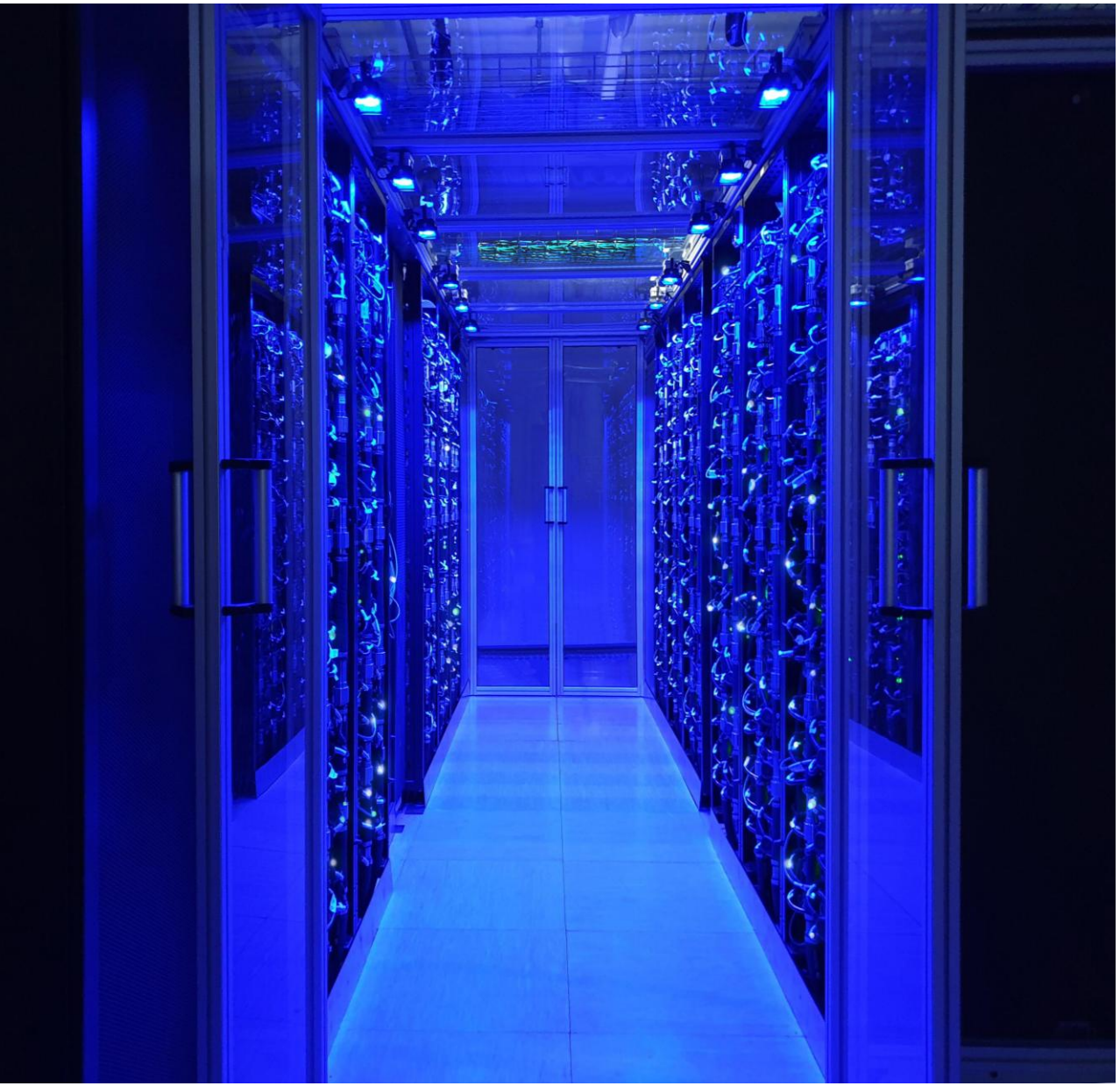
- External Infrastructure
- Collaborations

AI:AT Key Strength

- We bring everything **together in one place**: spaces, programs, resources, and people.
- **Start-ups, Research, Companies, and Public Institutions** work here side by side.
- Technical development, legal security, and entrepreneurial thinking interlock seamlessly.
- We not only assist with implementation – we also support the **search for potential**.
- Included: Access to **Computing power, coaching, Funding Agencies, and Investors**.
- **Goal:** From the **first idea to a scalable product** – all under one roof, without detours.

Physical Hub

- Modern Co-Working Space in Vienna
- Size: 1.500-2.500m²
- for Start-Ups, Companies and Researchers
- Including Training & Seminar Rooms, etc.
- Ready by March 2026



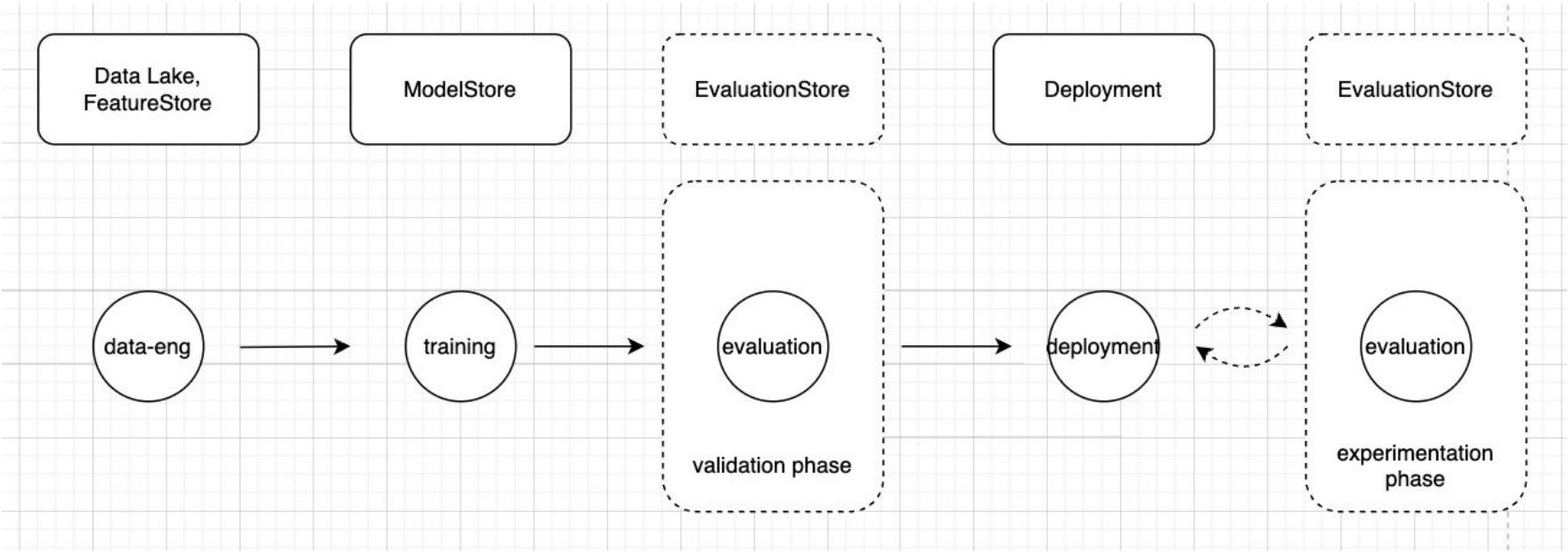
AI-optimized Supercomputer

Metric	Target KPI
GPU Nodes	~168 nodes (672 GPUs)
GPU Connection	NVLINK or similar
Peak Theoretical Performance (FP64)	~120 PFlop/s
Storage Capacity (Fast/Capacity Tier)	5 PB Fast NVMe / 20 PB Capacity HDD
Storage Throughput (Fast Tier)	≥ 3 TB/sw
Cooling Method	Direct water cooling (>95% efficiency)
Power Consumption	≤1.5 MW/h (peak operation)
Power Usage Effectiveness (PUE)	≤1.15
Compliance and Security	ISO/IEC 27001, NIS2 (planned)

From POC to Scale
AI Product, Organisational and Infra Journey

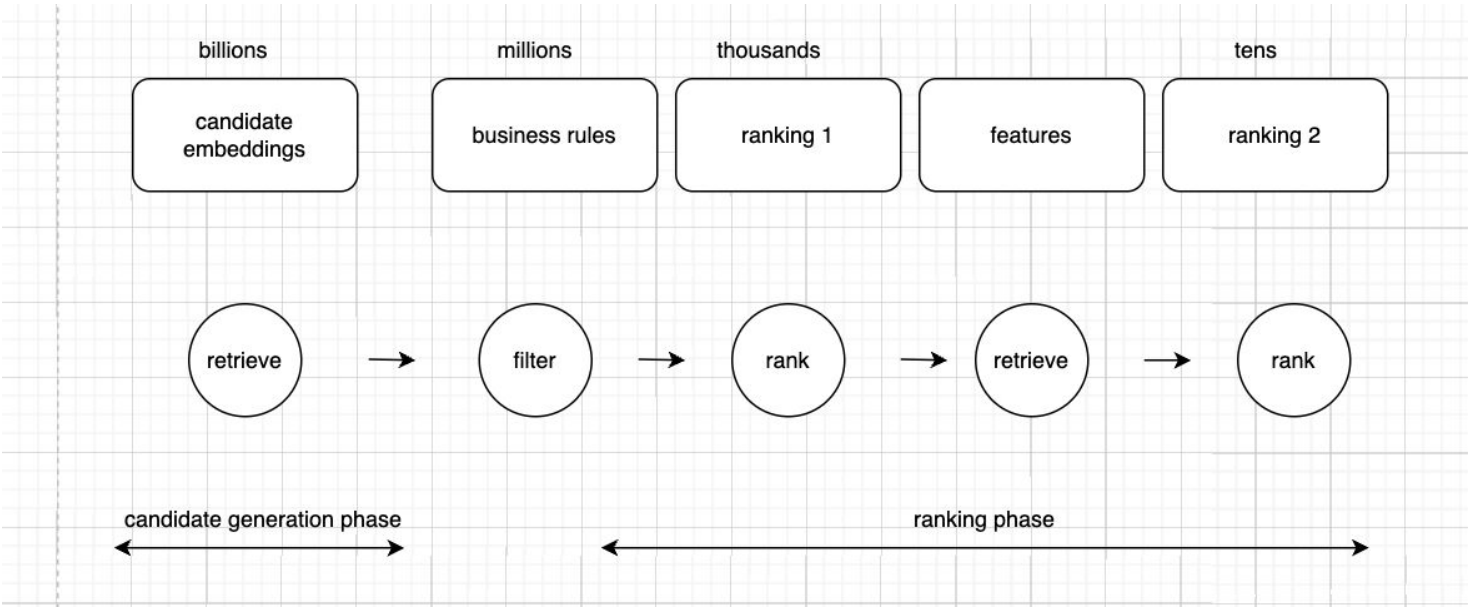
ML Canonical Lifecycle - Simple

- From Data to Deployment and Beyond



Lifecycle - What about Realtime/Online?

- Lifecycle of request…
- Recommender system example
- Realtime pipeline systems (internal to big companies...)

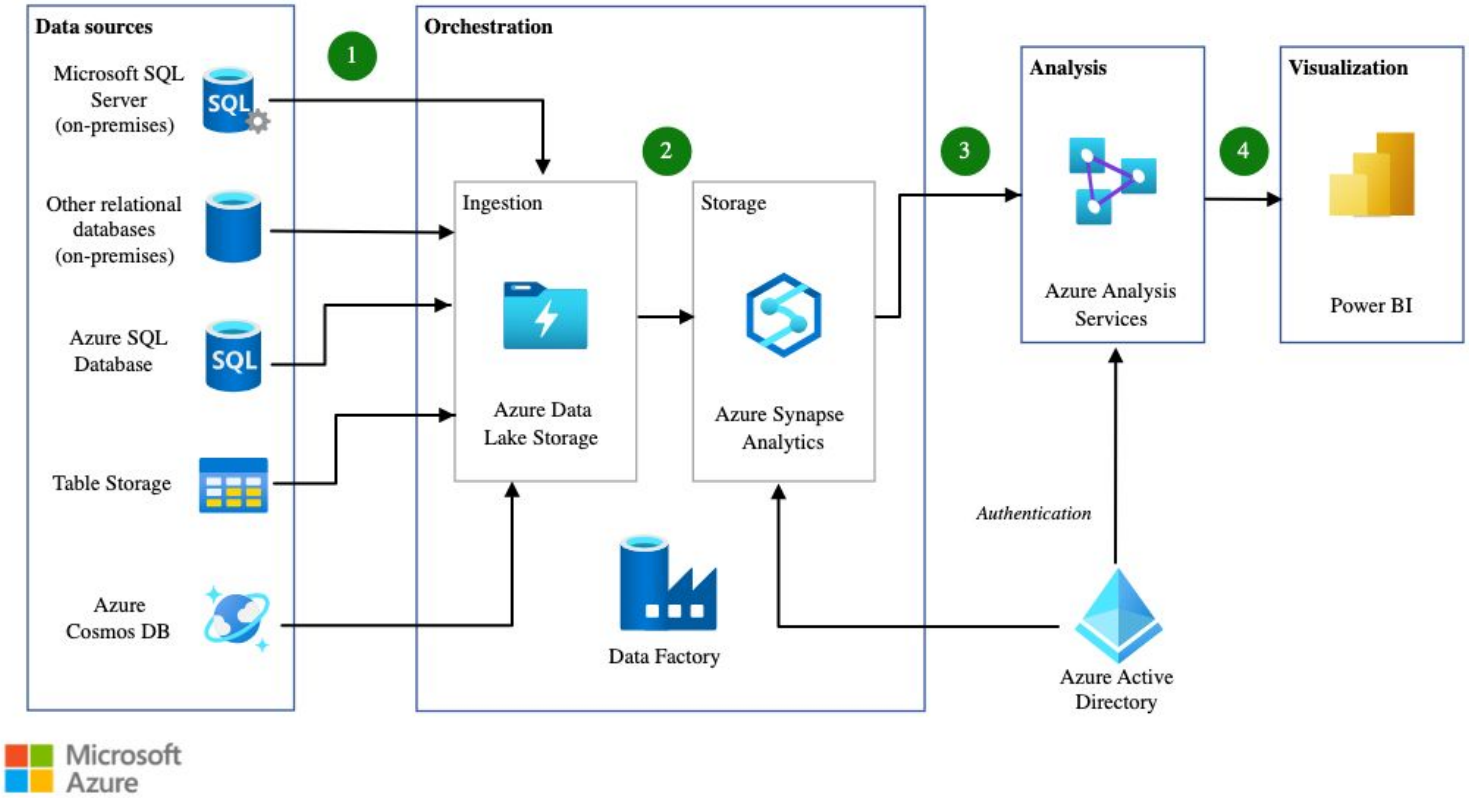


Data Systems in ML Operations

- Discover, Store and Reuse data for high scale ML
- **Data Warehouses**: large tables, curated data: used for analytics and history
  - Requires Query Processing: AWS Redshift + Tableau, Google BigQuery + Looker
- **Data Lakes**:
  - structured/unstructured, large-scale, offline, data of all company, analytics, ML, etc. cheap(ish).
  - simple* metadata systems for schemas, versions and raw data
  - Requires processing (typically **Spark**)
- **Feature Store(s)**:
  - Specialized for ML features, offline **and** online, not cheap, for curated and reusable data
  - Online data in DB, manages offline online skew, realtime features via streaming

Data Systems in ML Operations

- **Data Warehouses**:
- **Data Lakes**:
- **Feature Store(s)**:
- **Build vs Buy**



Model Infra and Systems

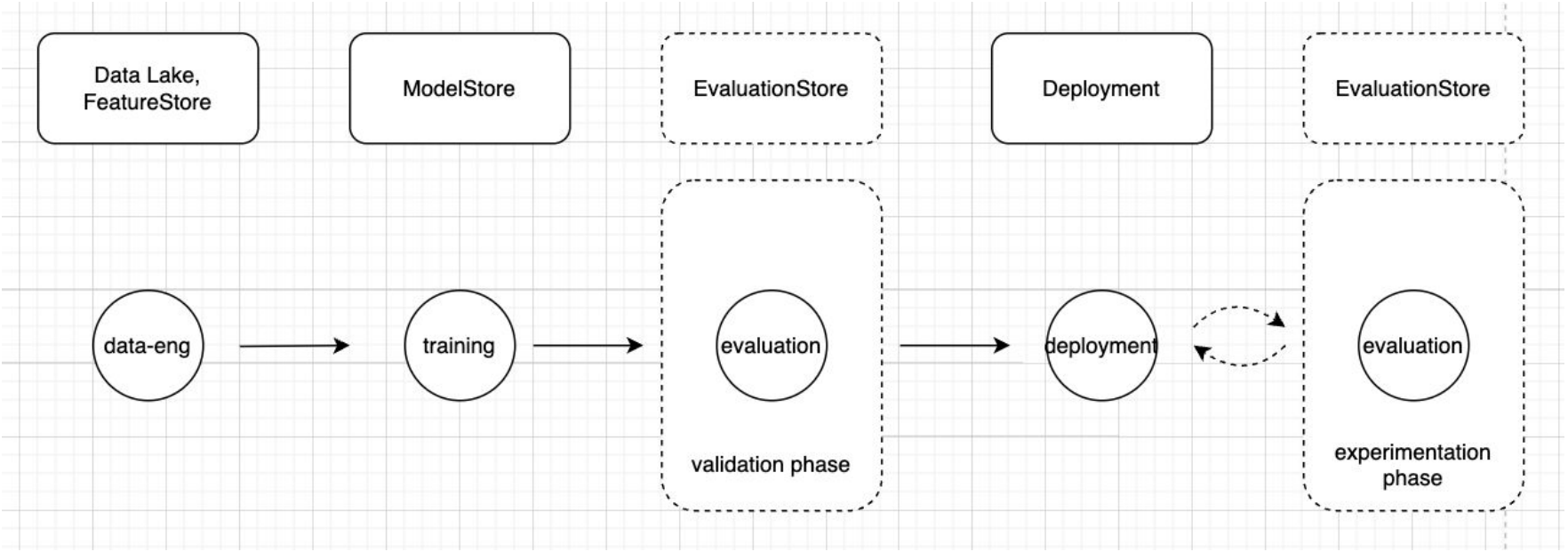
- Model Training Systems/Infrastructure:
  - Large **Spark** Cluster(s), or **K8**, or **Ray, Cloud Vendors** … Support for distributed training (Decision trees, Boosted Models, DNN etc.). Scheduling, Multi-tenancy, **HPC is here**! Rate limiting, Billing, …
- Inference Systems/Infrastructure:
  - Trained model != deployed model i.e. compilation
- ModelStores:
  - Stores models, provides versioning, and model *metadata*,
  - Versions, tracks code/lib dependencies, model lineage, input/output schemas, model cards,... checkpoints, cadence of retraining, App specific tags, …
- Build vs Buy

Testing in AI

- Current Situation in Evaluation Approach
  - Not very principled: manual, ad-hoc, blinders on narrow performance aspects (i.e. accuracy)
  - Metric centric
- Quality in ML/AI Context:
  - Quality is about validating **behavioral scenarios**
  - Clear **pass/fail** outcome, similar to **software eng. testing** (unit, integration…)
  - Metrics are just part of story, they represent *data*
  - Talk about Quality Assurance
  - Shift from Metric Centric to **Test-Centric Paradigm**

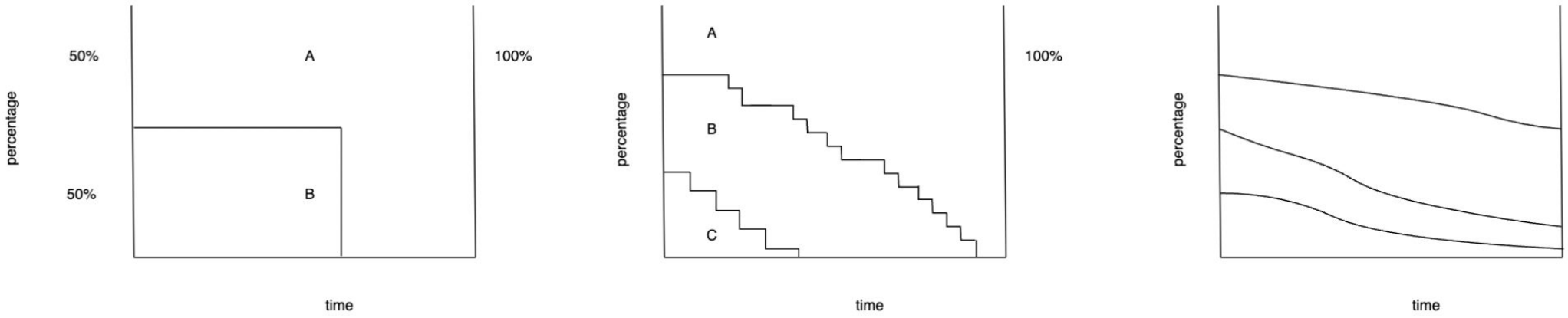
Testing and Deployment - An Interplay

- Lifecycle - see as continuous journey to check/ensure **quality**:
- Deployments - (long) **Processes,** not Events



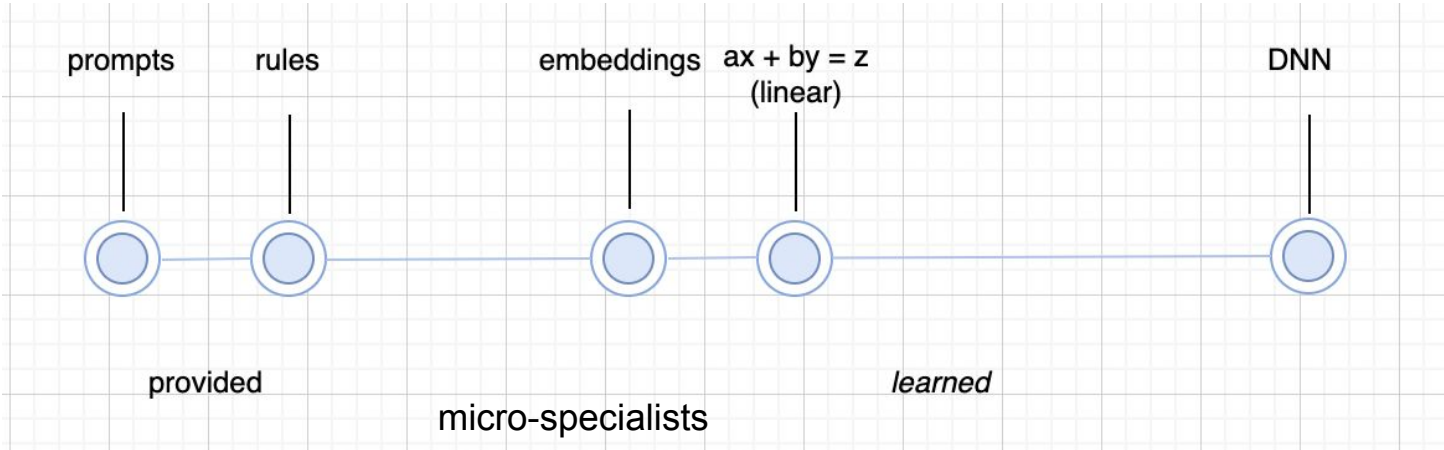
Testing and Deployment - An Interplay

- Many Deployment types - many *winning* versions (>> # models)
- a/b, multi-arm and contextual bandits, …
- Single model vs many models



Why Modular AI Matters

- In AI Modularity extends to ***Data*** + ***Objective*** + ***Quality*** *(not just code)*
- Large Models/Monoliths present difficulties *on many* ***(Sub)Objectives***
- **Modularity** Mandates ***Composition (foreign word in AI currently)***
- Modularity and Composition **scales*** (to billions of models)
- Focus shifts to managing (many) **lifecycles**
- Express (Sub)Objectives via ***Prompts***, ***Rules***, ***Representations*** *and* ***Models***

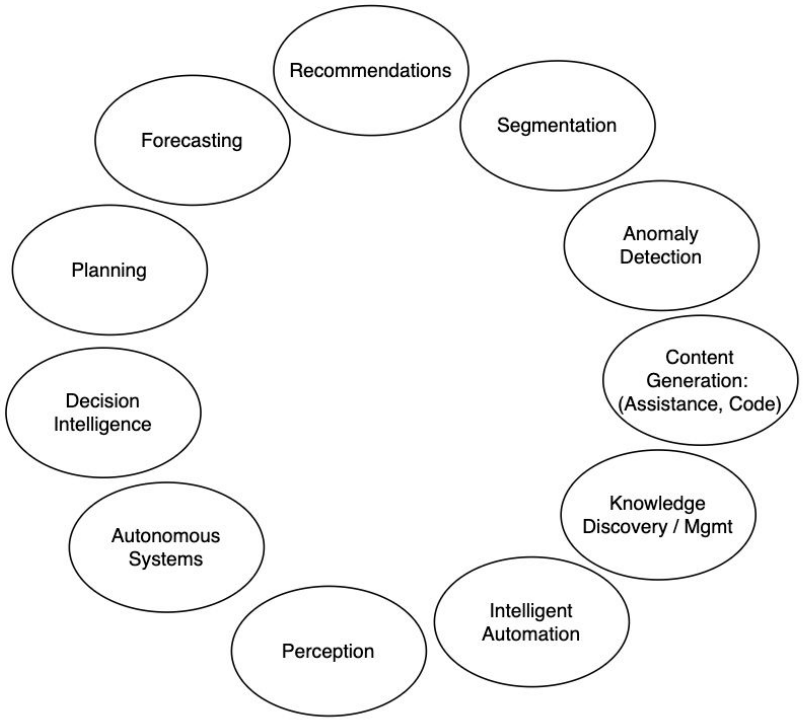


AI Journey for Organisations

- Devising an Effective AI and Data Strategy
- First Critical AI Hires
- High Level Roadmap and Commitment for Investment in AI Capabilities

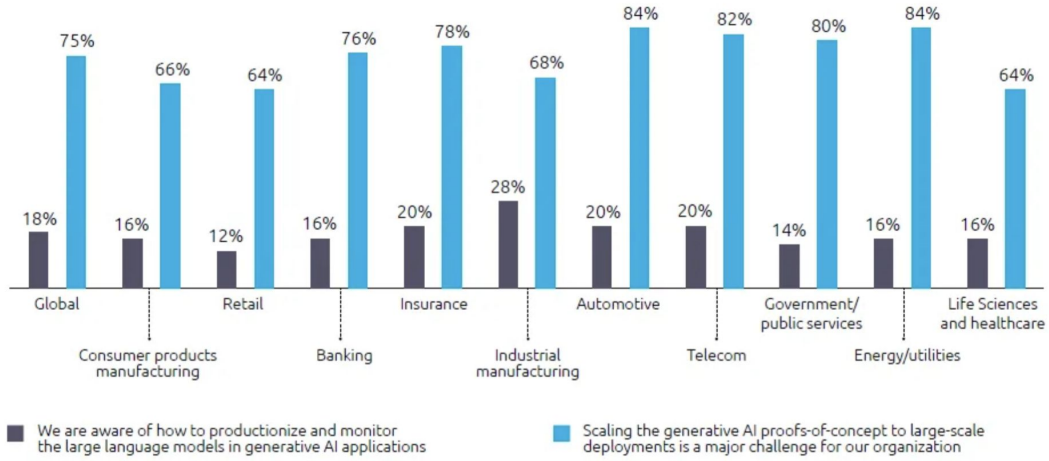
AI Journey for Organisations - Phase 1

- Identify Major AI Use Cases
  - (Workforce) Efficiency Improvements: Assistants..
  - New **Products** and new **Processes in Company**
  - Real Cases in Companies
- Validate Key Aspects of AI Value Prop
  - Implement one to two Pilot Products
  - Goal: **Validate** Go-to-Market, Gain **Competence**
- **Duration and Time**



AI Journey for Organisations - Phase 2

- Scaling Your AI Use Case(s)
  - Scaling your Team (Beyond the first few AI hires)
  - Training existing Engineers, and Product Managers
  - Data & ML Eng, Software Eng. Product
- Scaling Infrastructure
  - Beyond Pilots
  - Data Infrastructure takes shape for scale
  - Model Training / Inference Infrastructure
  - Goal is to **Validate** Go-to-Market **at Scale**
- **Duration and Time**



Scaling is Hard

Capgemini: Survey 2024 - GenAI

AI Journey for Organisations - Phase 3

- Towards Excellence at AI Use Case(s)
  - Improving your Team (Beyond the first team)
  - Training existing Engineers, and Product Managers
  - Real Cases in Companies
- Scaling Use Cases
  - Expand in Breadth
  - Excellence at Data Pipeline Infrastructure (error rates low)
  - Model Training and Inference and Test Infrastructure
  - Model **Deployment Flywheel**
  - Primary Goal is to **Excel at Scale, Efficiency and Revenue/ROI**
- **Duration and Time**

AI Journey for Organisations - Challenges for SME-s

- Hiring and Scaling AI Team
  - How to get High-Quality Hires
- Training Team
  - Improving your Team (Beyond the first team)
- Technical and Architecture Advice on Scale and Excellence
- Gain Visibility in AI Community (Local and Regional)
- **AI Leaders/Fellows on Board as Advisors along Investors**

AI Org - Inception to Excellence

Contact

AI FACTORY AUSTRIA

AI Factory Austria AI:AT
Schwarzenbergplatz 2
1010 Wien, Austria

training@ai-at.eu
info@ai-at.eu
ai-at.eu

@ai-factory-austria

THANK YOU

Funded by

EuroHPC Joint Undertaking

Funded by the European Union

Federal Ministry Innovation, Mobility and Infrastructure Republic of Austria

under discussion with

FFG Promoting Innovation.

AI Factory Austria AI:AT has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101253078. The JU receives support from the Horizon Europe Programm of the European Union and Austria (BMIMI / FFG).

AI:AT