infobip

AI @ Infobip

From Research to Production

MAY 2026

What we'll talk about today

1. Who am I?
2. What is Infobip?
3. Where does AI Fit in?
4. How we do: AI Projects at Infobip
5. What Does it Take to Build This?

Who am I?

WHO AM I?

About me

- MoEE @ ETF
- Senior AI Software Engineer @ Infobip

WHO AM I?

About me

- MoEE @ ETF
- Senior AI Software Engineer @ Infobip
- Published some research papers

WHO AM I?

About me

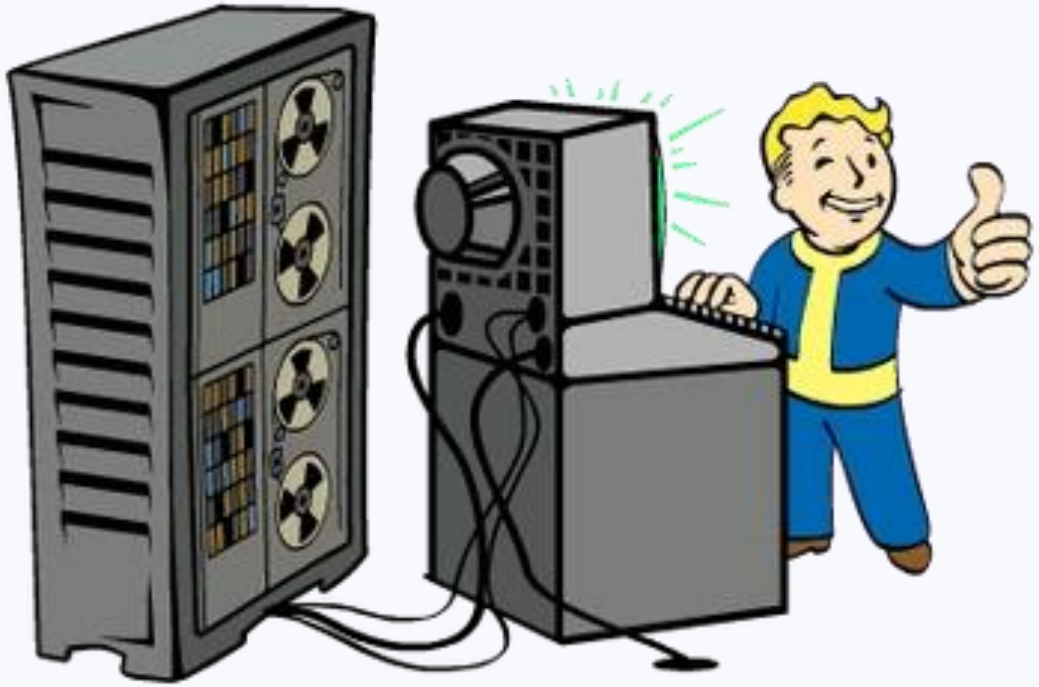
- MoE @ ETF
- Senior AI Software Engineer @ Infobip
- Published some research papers
- Some hobbies:
  - Pub quizzes
  - Sci-Fi and Fantasy
  - (Retro) Gaming

What I do at work

- AI & ML (obviously)

...but also...

- Developer Experience (DevEx)
- Systems Design and ML infra
- Tooling, observability, traceability
- API design

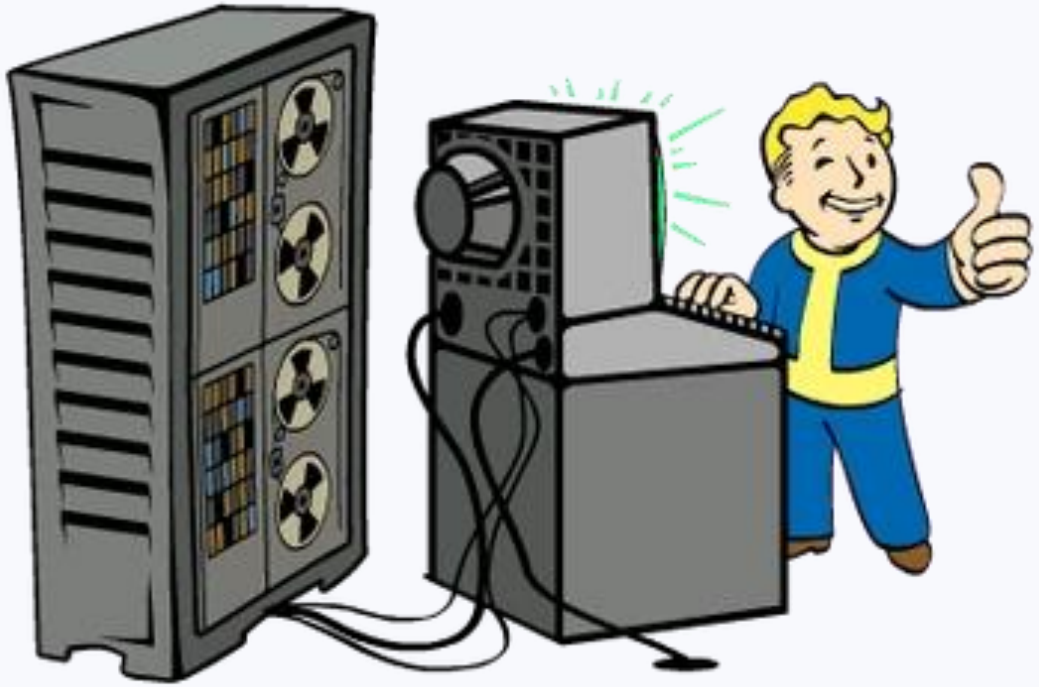


What I do at work

- **AI & ML (obviously)**

...but also...

- Developer Experience (DevEx)
- **Systems Design and ML infra**
- Tooling, observability, traceability
- API design



What is Infobip?

Infobip (TL;DR version)

- Croatian product-based IT and Telco company
- Employees: 3500+
  - Engineers: ~1000
- Offices: 75+

Infobip (TL;DR version)

- Croatian product-based IT and Telco company
- Employees: 4000+
  - Engineers: ~1000
- Offices: 75+
  - **Sarajevo office: Deals mostly with Voice (audio) processing... but also with AI!**

WHAT IS INFOBIP?

Infobip (TL;DR version)

- Croatian product-based IT and Telco company
- Employees: 4000+
  - Engineers: ~1000
- Offices: 75+
  - **Sarajevo office: Deals mostly with Voice (audio) processing… but also with AI!**
- Member of **AAIF** (Agentic AI Foundation) and Linux Foundation

WHAT IS INFOBIP?

Infobip – The Core Problem

- **Connecting Businesses to their Customers at scale**

Infobip – The Core Problem

- **Connecting Businesses to their Customers at scale**

Infobip – The Core Problem

- **Connecting Businesses to their Customers at scale**

Meta unicef

Uber Coca-Cola

Some of our clients

Infobip – The Core Problem

- Connecting Businesses to their Customers at scale

Through various channels (WhatsApp, SMS, Viber, etc.)

Up to 10 billion (10 000 000 000) messages...daily!

Meta unicef Uber Coca-Cola

Some of our clients

Infobip – Infra Powerhouse

- Global infrastructure
  - 1300+ physical servers
  - 40+ data centres
  - Microservice architecture
  - 1600 service types
  - 15000+ application instances

Where does AI fit in?

WHERE DOES AI FIT IN?

AI @ Infobip

- AI is a **force multiplier:** smarter routing, targeting, timing, etc.

WHERE DOSE AI FIT IN?

AI @ Infobip

- AI is a **force multiplier:** smarter routing, targeting, timing, etc.
- Two sides to the same coin:

Customer-facing AI

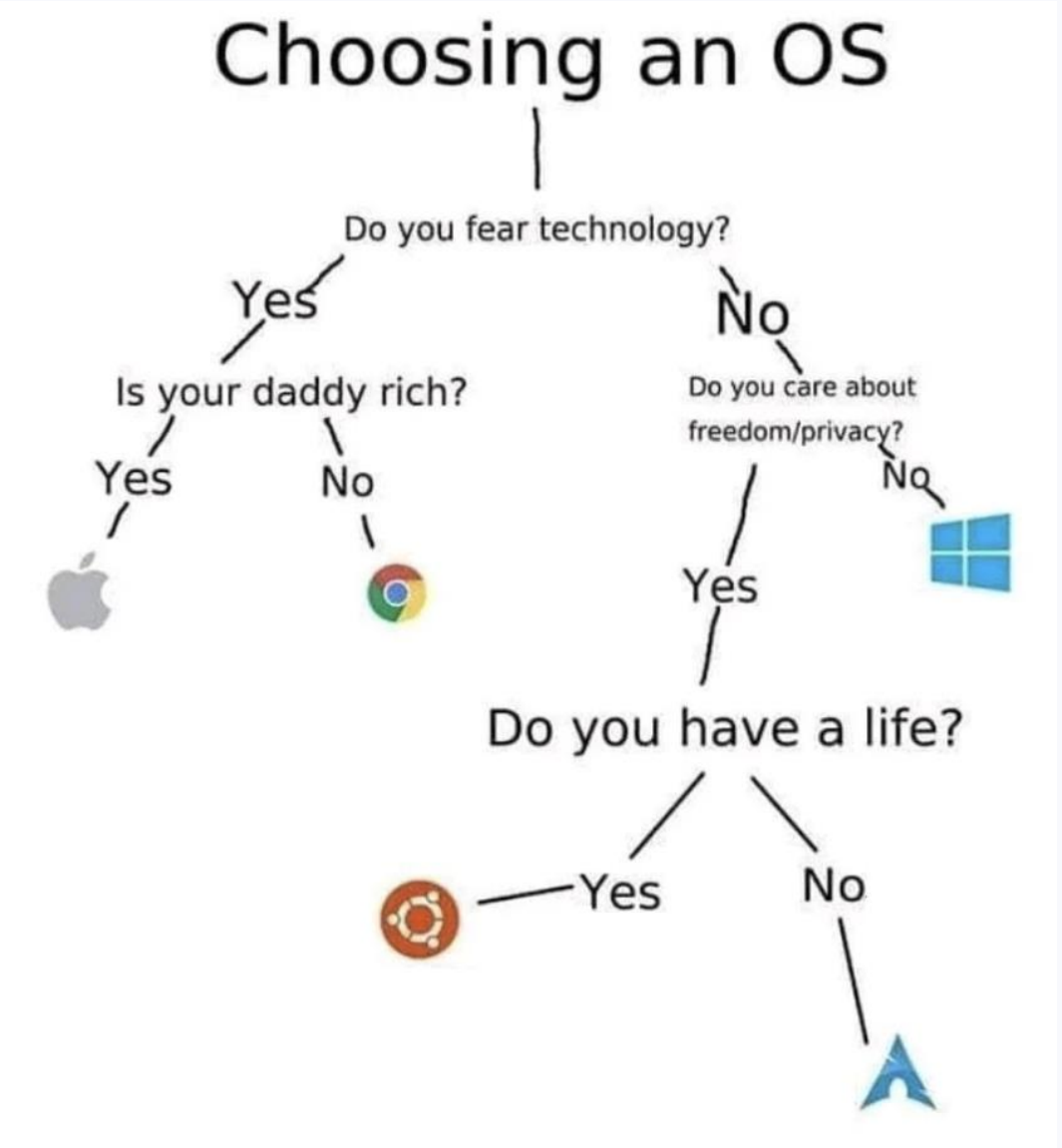
AI Infrastructure

WHERE DOES AI FIT IN?

AI @ Infobip

AI projects include, but are not limited to:

- Classic ML – Clustering, Decision Trees, etc.



WHERE DOES AI FIT IN?

AI @ Infobip

AI projects include, but are not limited to:

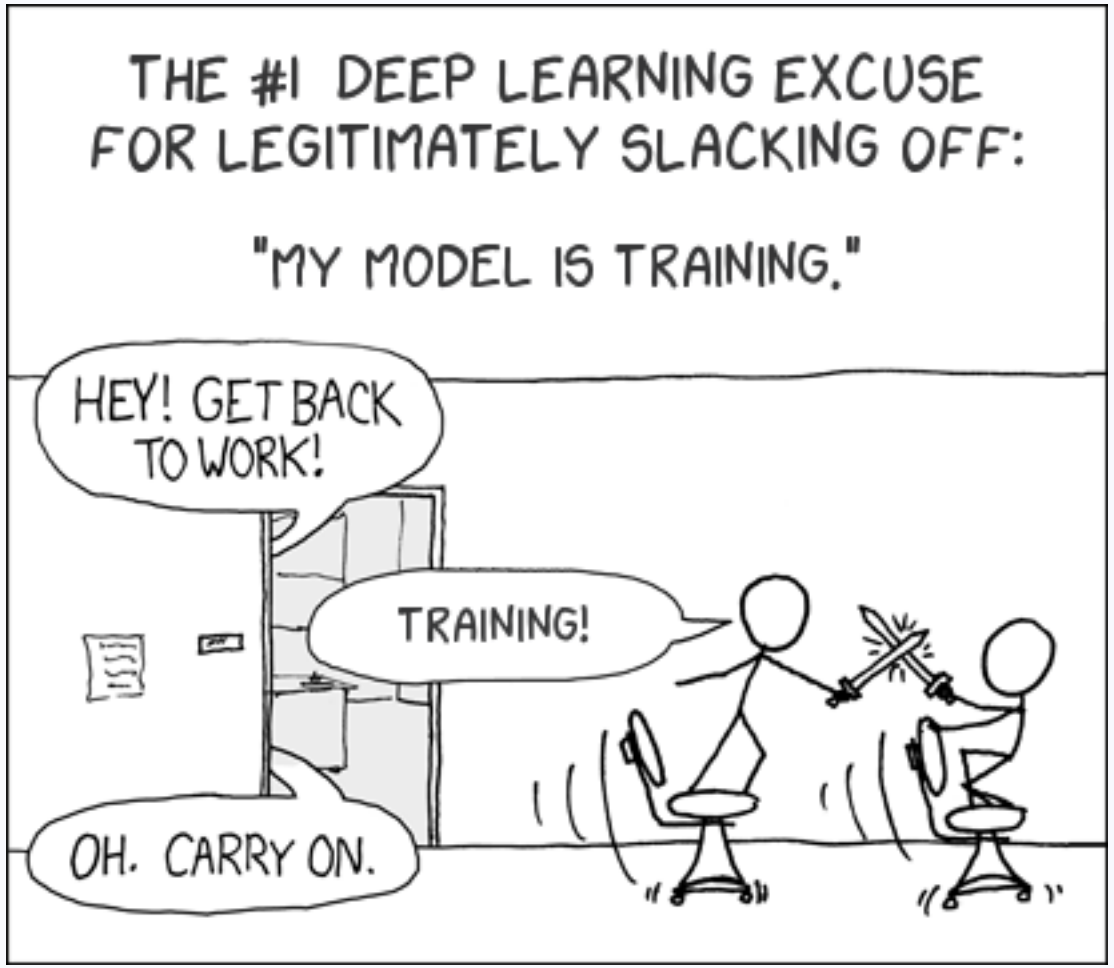
- Classic ML – Clustering, Decision Trees, etc.
- Recommender systems
  - Classical approaches
  - LLM-as-a-judge

WHERE DOES AI FIT IN?

AI @ Infobip

AI projects include, but are not limited to:

- Classic ML – Clustering, Decision Trees, etc.
- Recommender systems
  - Classical approaches
  - LLM-as-a-judge
- Deep learning (for audio, images, etc.)

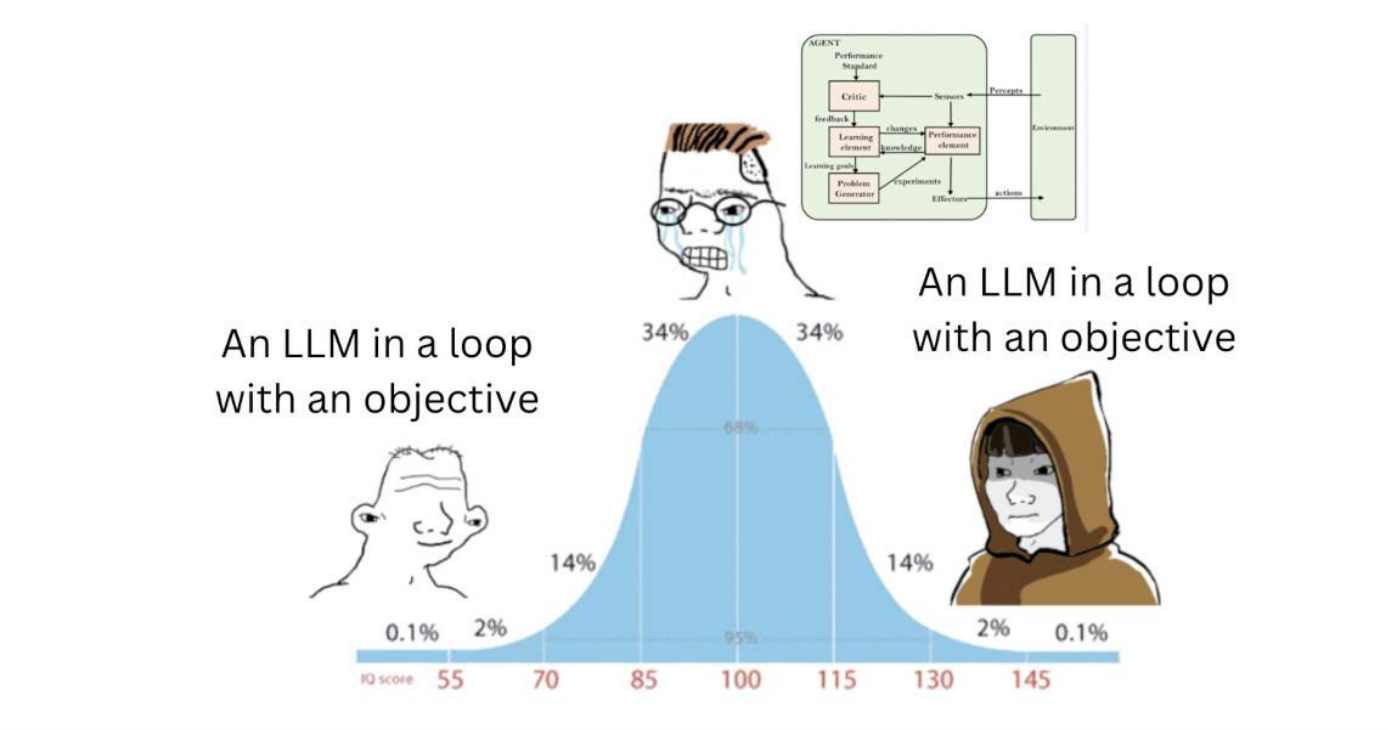


WHERE DOES AI FIT IN?

AI @ Infobip

AI projects include, but are not limited to:

- Classic ML – Clustering, Decision Trees, etc.
- Recommender systems
  - Classical approaches
  - LLM-as-a-judge
- Deep learning (for audio, images, etc.)
- GenAI
  - LLMs
  - MCPs
  - ML and Agentic system design



WHERE DOES AI FIT IN?

R&D @ Infobip

- Part of major **EU R&D initiative**
- Building next-gen **cloud + edge infrastructure**
- Focus on **secure, AI-ready communications**
- Collaborating with **industry + universities**
- **€80M+ project**, co-funded by the EU
- Engineers are encouraged to publish original research on top-tier conferences!

Important Projects of Common European Interest (IPCEI)

How we do: AI Projects at Infobip

AI PROJECTS AT INFOBIP

Destination Scoring

The Client:

- Platforms managing WhatsApp campaigns for various brands

The Problem:

- Marketing campaigns via WhatsApp get blocked by Meta

Root Cause:

- Meta scores every sender – quality score drops, sender gets blocked

The Task:

- Rank recipients by engagement likelihood, so we send to the “safest” users first, protecting sender quality score

Tech Stack:

- Python (numpy, pandas, sklearn, etc.)
- MLFlow
- Apache Kafka
- PSQL
- Kubernetes (k8s)
- FastAPI (Python Web Framework)

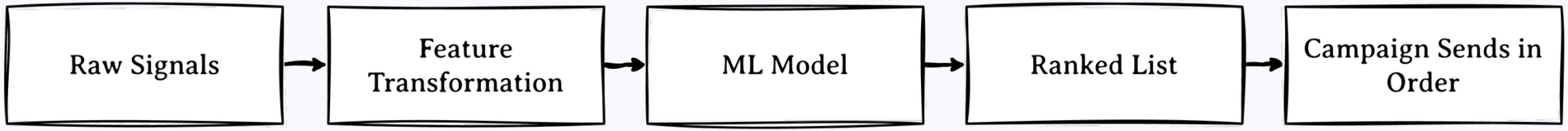
Concepts Employed:

- Large-Scale Data Analysis
- LLM-as-a-judge

Destination Scoring

- Data: Interaction history with relevant events (delivery counts, seen counts, response counts, time-to-open, etc.)
- 5 hand-crafted signals, not raw features – score gets calculated
- Scores grouped into buckets or numeric tiers
- Campaigns get sent out to users by descending score (highest to lowest)

The Result:
Campaign delivery rate increased by 13% on average



AI PROJECTS AT INFOBIP

Campaign Drifting

The Client:

- Businesses that send large-scale SMS and WhatsApp campaigns

The Problem:

- Marketing campaigns sometimes get misused – registered with one intent (e.g. sales), but promoting another (e.g. gambling)

Root Insight:

- We are interested in identifying new patterns of intent for campaigns

The Task:

- Messages need to be tracked and evaluated in real-time in order to detect drift in topic

Tech Stack:

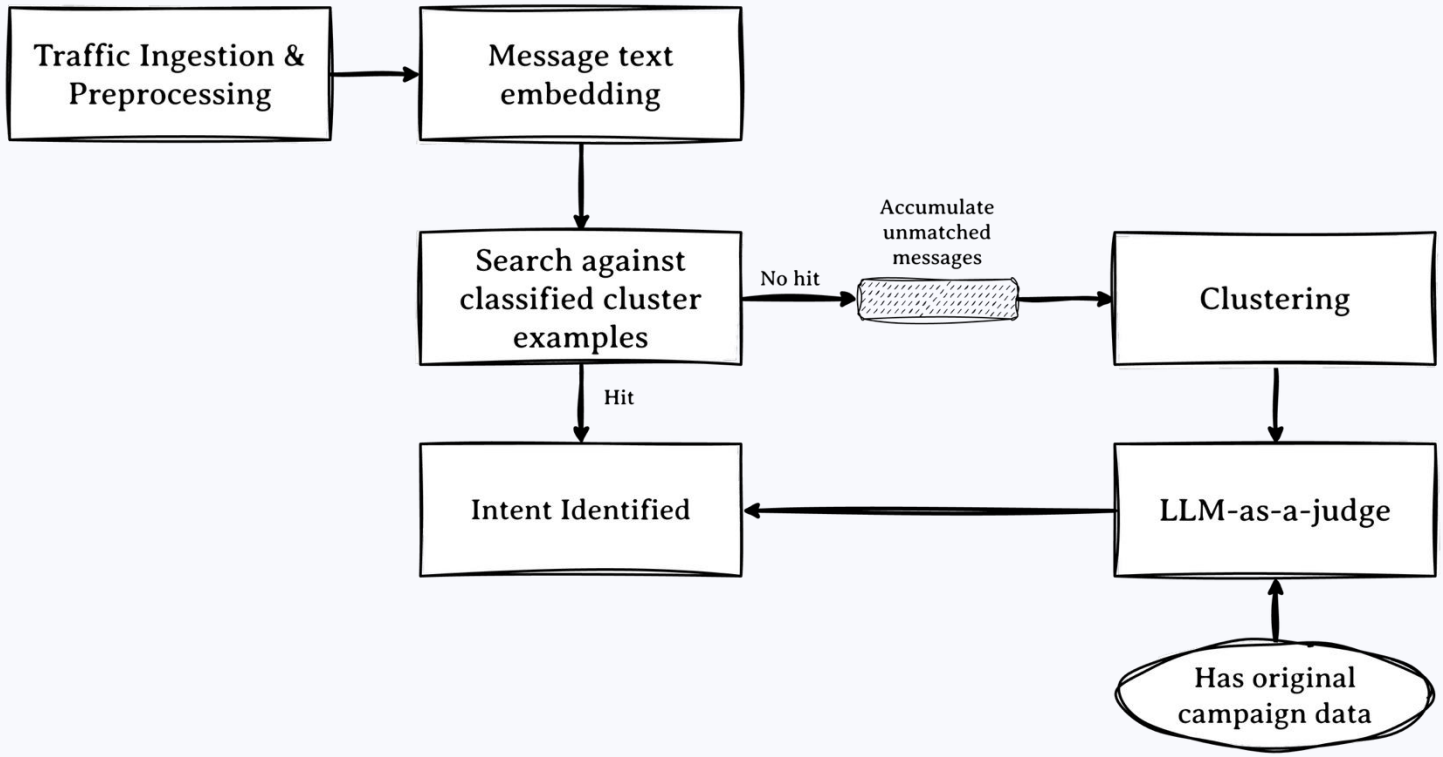
- Java (Spring Boot) + Python (FastAPI)
- Redis
- ElasticSearch

Concepts Employed:

- Data Clustering
- Text Embeddings
- Vector Search

Campaign Drifting

- Data: Combination of campaign registration metadata and message traffic tied to that campaign
- Scale: 20+ million messages… weekly!
- Unknown data gets accumulated and classified – high data variance, LLM utilized!
- Senders which abuse the system get flagged and blocked



OpenAPI MCP

MCP = Model Context Protocol: a standard for connecting LLM apps to external tools and data sources.

The Client:

- A new type of user – not a human, but an AI Agent!

The Problem:

- Making software AI-accessible requires MCP servers

Root Insight:

- Companies already have APIs – what if we can generate an MCP server from the API specifications (OpenAPI schema)?

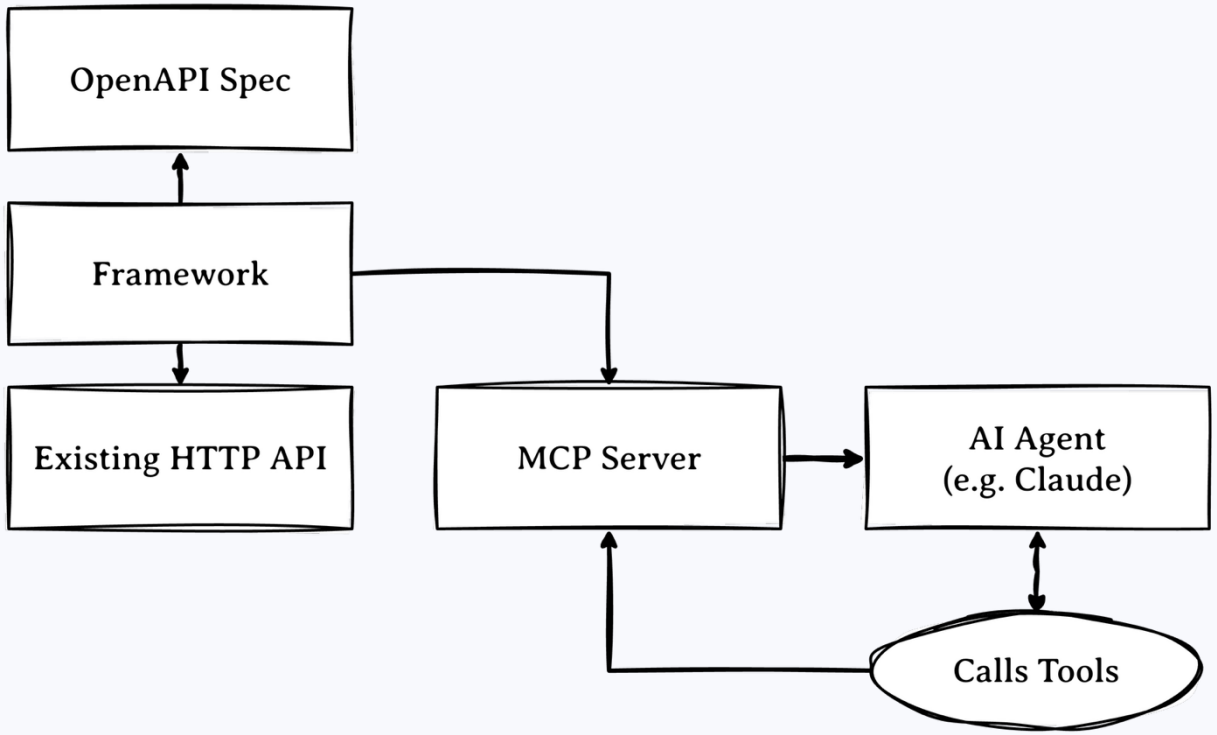
The Task:

- Create a framework for exposing any OpenAPI-documented HTTP APIs as MCP servers

OpenAPI MCP

What we've built:

- Core translation: OpenAPI operations to MCP tools (sounds simple - it's not!)
- Schema transform: OpenAPI supports some features which MCP does not, and vice versa
- Auth out of the box
- Mock/testing mode
- In production at Infobip – powers our MCP servers
- Open source, usable today: https://github.com/infobip/infobip-openapi-mcp



AI PROJECTS AT INFOBIP

Answering Machine Detection (AMD)

The Client:

- Businesses that send automated voice message campaigns

The Problem:

- Sometimes calls are answered by voicemail: it’s a waste of money to keep those calls connected

Root Insight:

- Clients should not waste time (money) by talking to answering machines

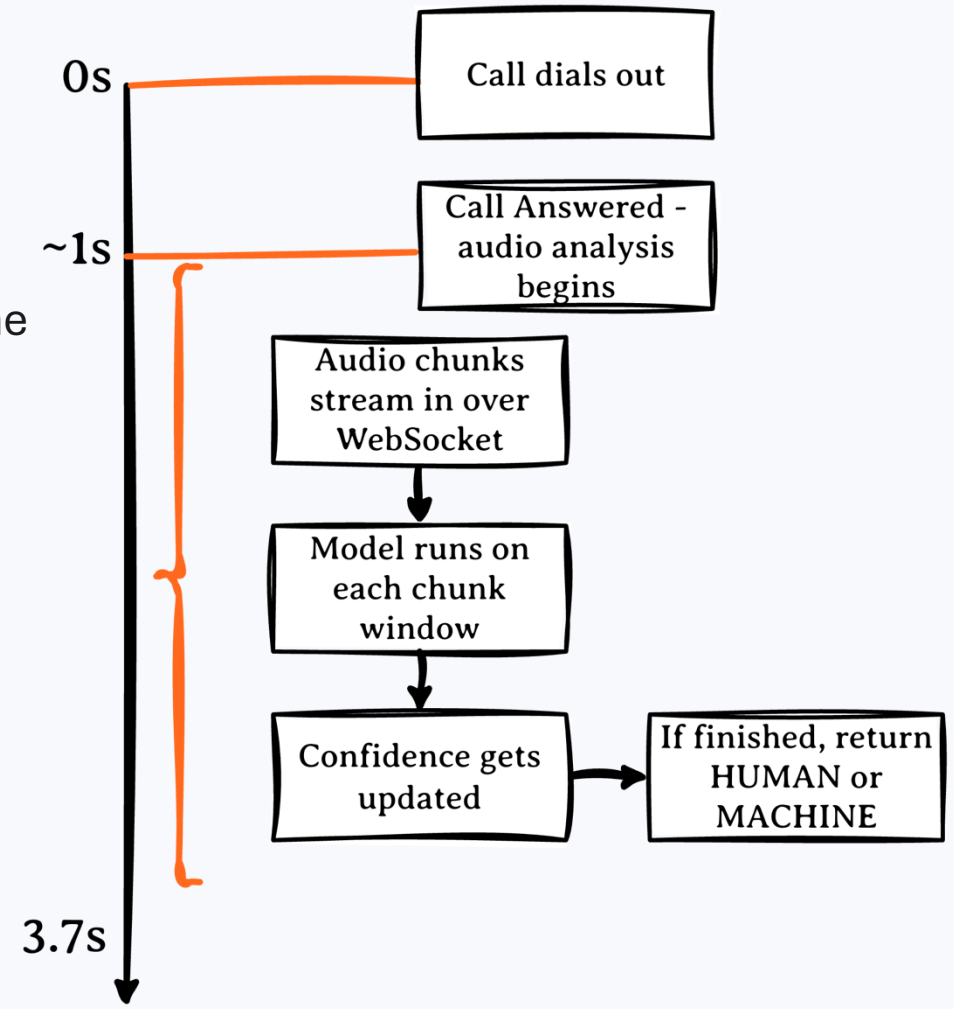
The Task:

- Within the first few seconds of a call being answered, classify it as either human or answering machine, and act accordingly

Answering Machine Detection (AMD)

What we've built:

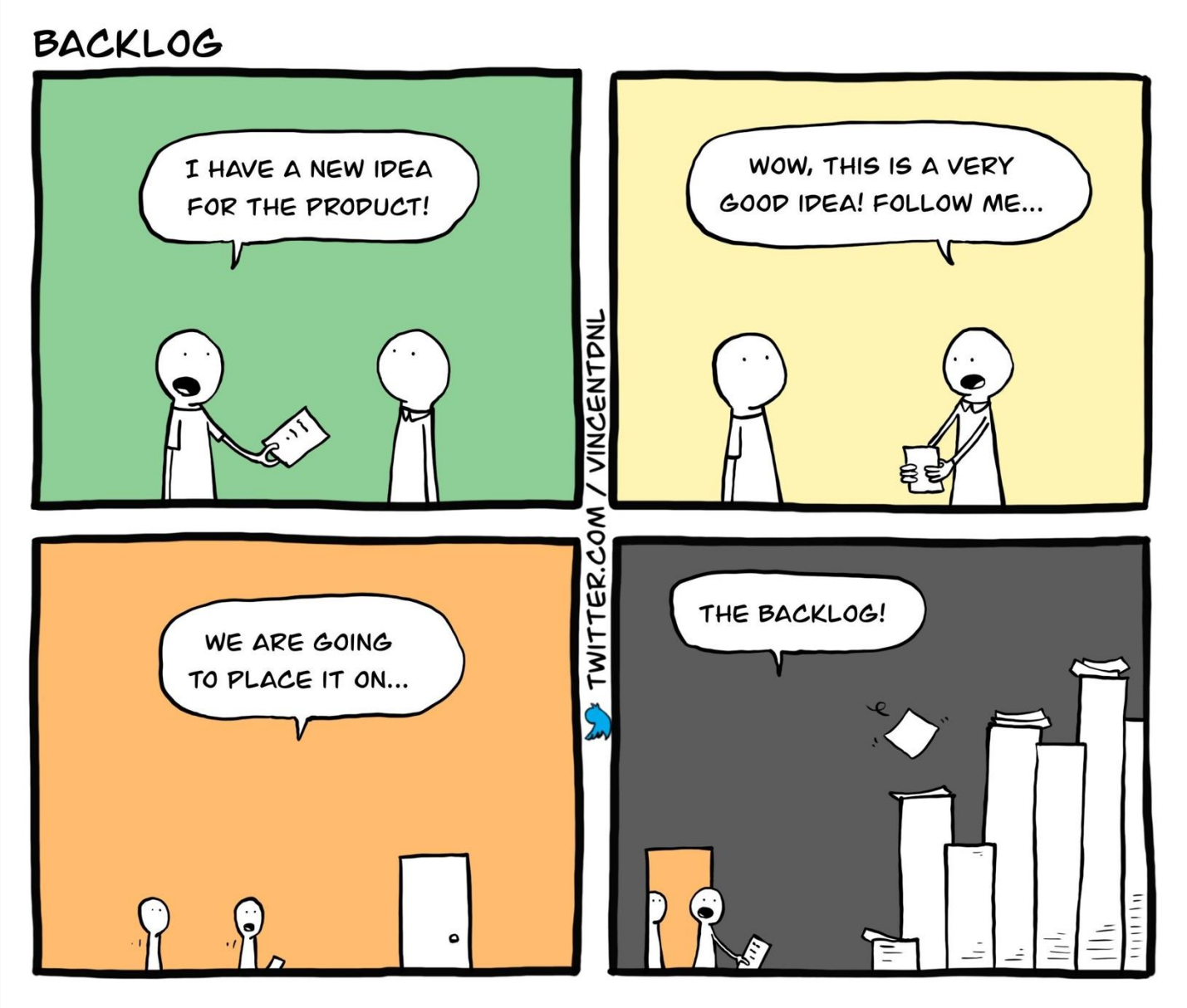
- A system that works with streaming data – audio arrives in chunks over WebSocket: The model must classify in real time
- Low-latency inference is a must-have – 3.7s for detection at best
- Output: Classification + confidence scores
- Genuinely hard cases: human-like voicemail greetings
- Full feedback loop: label data, train model, export, deploy, feed predictions back for monitoring and re-training



Many others

To name a few:

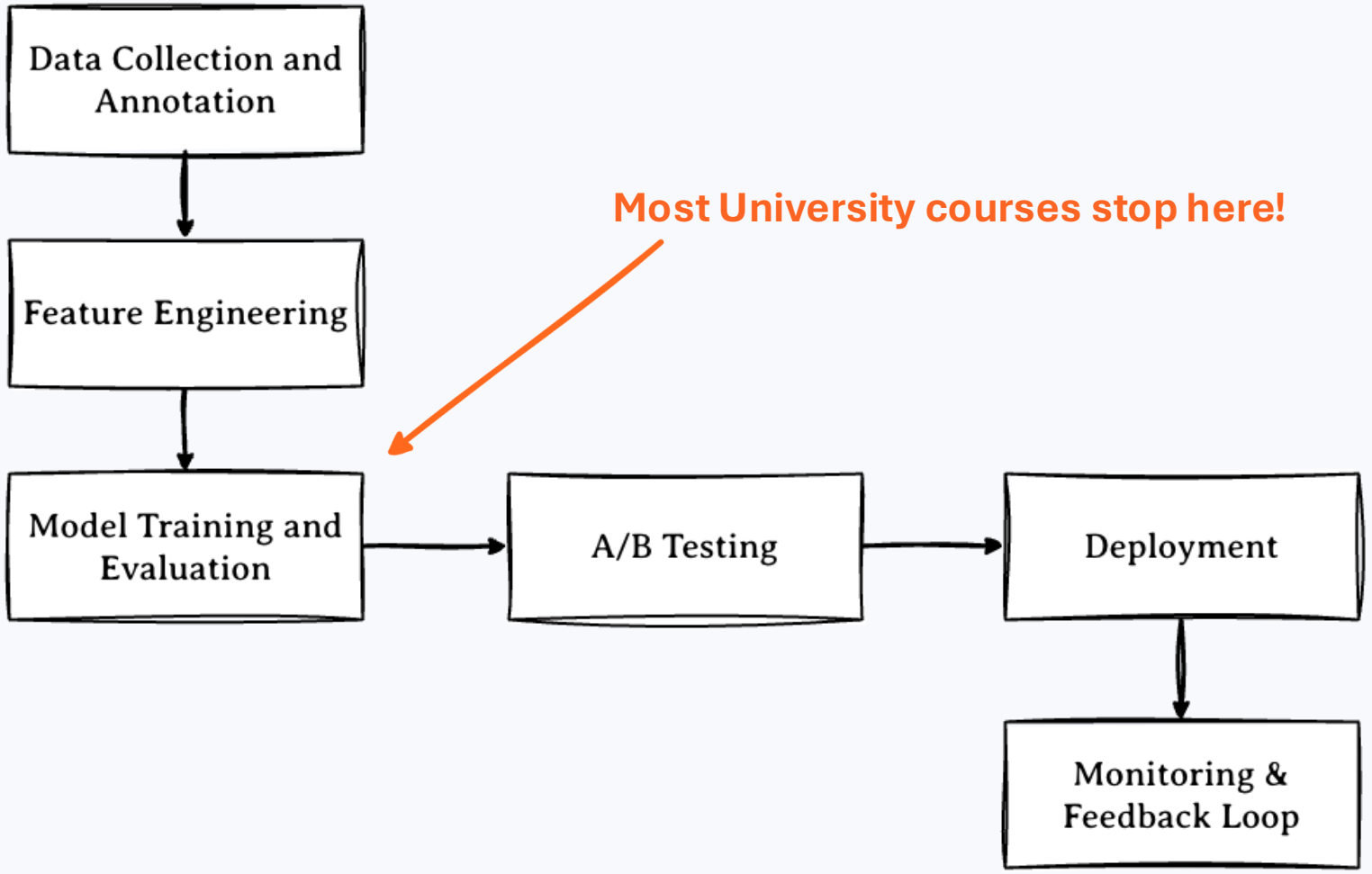
- Spam detection
- Smart routing
- AI Assistants and Agents
- Intent Detection
- Various AI Infra Projects



What does it take to build this?

BUILDING ALL OF THIS

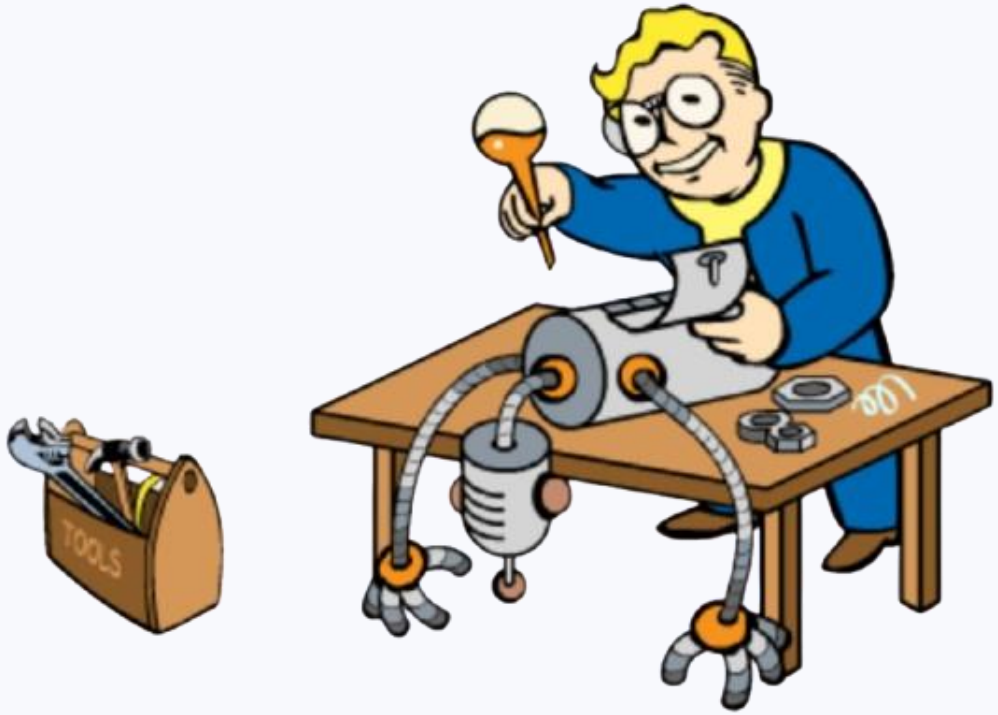
The Full Lifecycle



BUILDING ALL OF THIS

Industry Challenges

- Drift
- Latency
- Scale
- Explainability

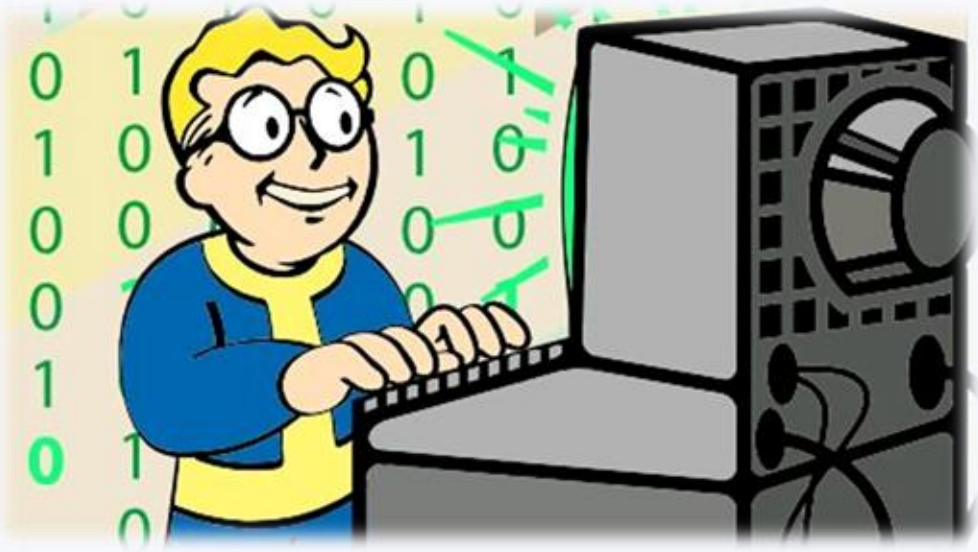


BUILDING ALL OF THIS

Skills to Build

Base Skills:

- SQL & Data Fluency
- Systems Thinking
- MLOps Basics (Docker, CI/CD, git)
- Read papers!
- Hot take: Programming Languages do not matter that much!



- After the basics: Depends where you want to specialize!

Q&A

Thank you!

Edvin Teskeredzic

Senior AI Software Engineer

Edvin.Teskeredzic@infobip.com

www.infobip.com

LET'S STAY IN TOUCH