Artificial Intelligence for Connected Industries

Big Data Technologies for Connected Industries

Daniele Miorandi & Stefano Tavonatti

Co-funded by
the European Union

Teams

1. Nika Soltani Tehrani, Leevi Haukijärvi, Kanwaljeet Kaur
2. Lluís Expósito, Victor Ceballos and Mariona Montal
3. Muhammad Haider Azam, Younes Guerfi, Iyad Chehili

Big picture

All projects are about **designing** and **implementing** big data systems. We expect you to use big data technologies you learned throughout the course (you could solve the problem without them, but this wouldn’t lead to a decent score). We welcome and reward creative, non-trivial approaches. For each project, identify potentially relevant input data and sources. Consider **real data** to understand “in the wild” challenges, but if this is not feasible, build a synthetic (yet realistic-looking) dataset and explain why. **Demos** with real data whenever possible are highly appreciated. Project details such as approaches, metrics, and functionality are **suggestions**, not mandatory.

Project delivery

- By email to Daniele & Stefano
- Deadline: ~~13.02, 11.59pm~~ **20.02, 11.59pm**
- What we expect to get:
  - Slideset (template provided, 10 slides presentation of what you did in the project)
  - Code (public git repo) → We will run it! (If it does not work ☠️)
  - (Optional) Other relevant material (screenshots, video of demos etc.)
- Use of GenAI is not allowed - it's *recommended*!
  - But mention it explicitly - what you used and how

Projects - Team 1

Web Topic Drift & Narrative Mapping

Build a big data system that analyzes large-scale web text to show how topics and narratives cluster and shift over time or across sources. Create a pipeline for ingestion, text normalization, and scalable topic modeling, and add entity extraction to link themes to people, places, or organizations. Produce topic drift indicators and a simple timeline view with representative snippets. Data source: the OpenWebText2 corpus on Hugging Face (https://huggingface.co/datasets/Geralt-Targaryen/openwebtext2). Suggested metrics: topic coherence (e.g., NPMI) and drift magnitude per time window.

Projects - Team 2

Shot Selection & Efficiency Insights (Basketball Analytics)

Build a big data system that analyzes play-by-play and shot-level events to evaluate shot quality and efficiency by player, team, or game context. Aggregate shots by zone, game phase, and score margin, and compute expected value or efficiency indicators per segment. Deliver shot charts and short comparative reports that highlight high-value patterns and potential defensive gaps. Data source: the NBA/WNBA play-by-play + shot-detail dataset on Hugging Face (https://huggingface.co/datasets/Vladislav/nba_dataset). Suggested metrics: expected points per shot zone and shooting efficiency by context (player/team/...).

Projects - Team 3

Urban Mobility Demand Hotspots (Taxi Rides)

Build a big data system that identifies spatial and temporal demand hotspots in taxi rides. Ingest trip records, aggregate by time window and zone, and detect recurring peaks as well as atypical surges. Provide a heatmap dashboard and a compact set of operational insights (e.g., high-variance zones or time windows). Data source: the NYC taxi rides dataset on Hugging Face (https://huggingface.co/datasets/TaherMAfini/taxi_dataset). Suggested metrics: hotspot persistence over time and demand-variance by zone.

Instructors

- Daniele Miorandi (theory) - daniele.miorandi@afliant.com
- Stefano Tavonatti (labs) - stefano.tavonatti@afliant.com