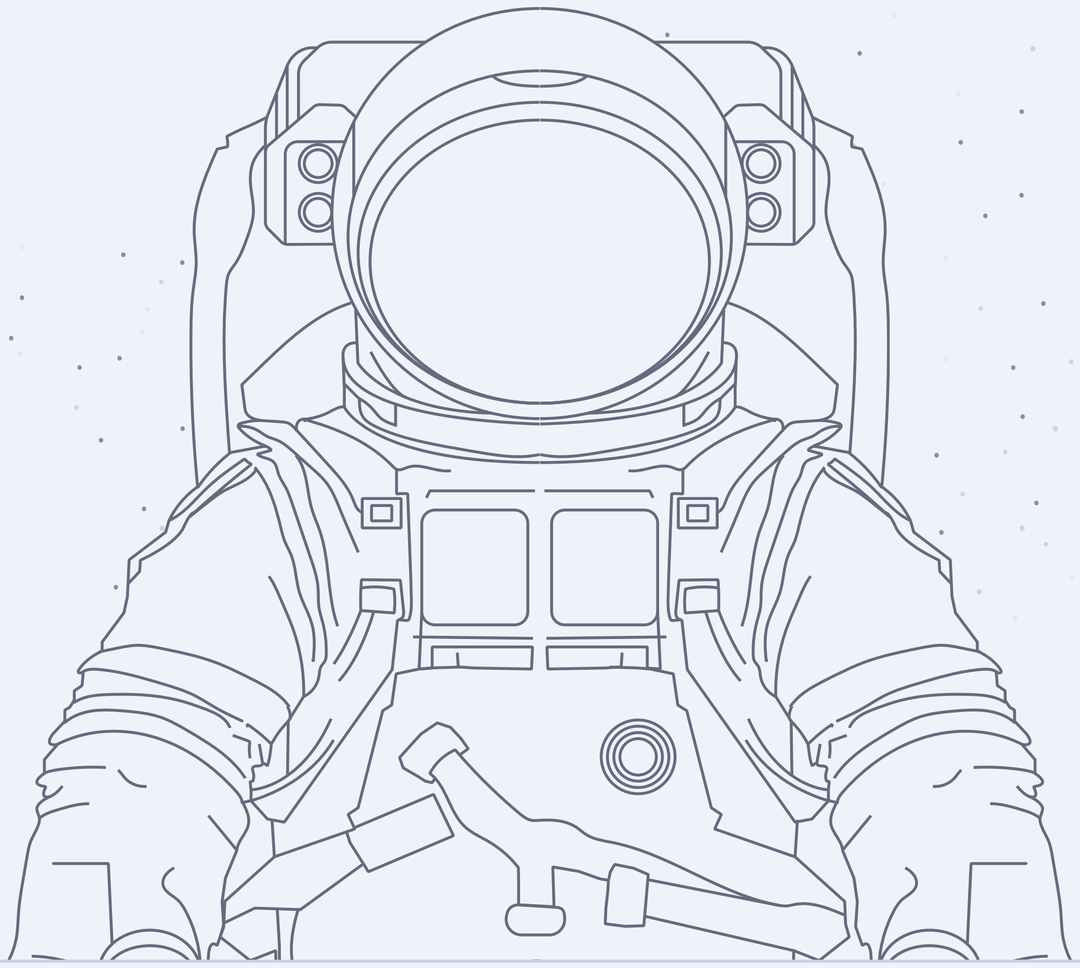
Qdrant

VECTOR SPACE DAY PROGRAMMING



9:30 - 10:55	OPENING KEYNOTES	**Qdrant** | Andre Zayarni, Andrey Vasnetsov, Neil Kanungo **Microsoft** | Robert Eichenseer **Vultr** | Kevin Cochrane **AWS** | Inaam Syed
10:55 - 11:10	BREAK	

BREAKOUT SESSIONS

TRACK A: MILKY WAY

Architectures, Infrastructure & Multimodal Retrieval

11:10 - 11:35

Building a News Sleuth for the Deep Research Paradigm

AskNews
Robert Caulk, C*O

High-stakes decisions demand news pipelines with integrity, structure, and speed. We'll present AskNews' Deep News Research approach: an expert ontology, rigorous sourcing, fair-use governance, and high-performance hybrid retrieval that together make automated investigation viable. We'll detail how raw articles become contextualized, how entities and events are linked, and how vector-lexical fusion supports exploratory and targeted queries. With this foundation, agents can triage signals, trace claims, and surface contradictions—useful for geopolitical risk monitoring, investigative journalism, prediction markets, and fact-checking. We'll discuss evaluation methods that reward precision, coverage, and timeliness, plus safeguards against misinformation. Attendees will leave with a blueprint for building AI researchers who respect journalism's standards while operating at machine speed.

TRACK B: ANDROMEDA

AI Workflows, Agents & Applications

Beyond Web Search: Real-Time Web Intelligence for AI-Native Agents

Linkup
Philippe Mizrahi, CEO

Traditional search was built for human clicks; agents need structured, real-time context. This talk introduces an AI-native search infrastructure that retrieves, normalizes, and integrates web intelligence directly into LLM applications. We'll discuss crawling and enrichment pipelines, entity and event modeling, and hybrid vector-lexical retrieval that supports both broad exploration and precise lookups. You'll see how fresh, machine-readable context enables planning, tool use, and citation—unlocking agents that can monitor markets, watch competitors, or synthesize research with traceable sources. We'll share design decisions, latency budgets, and safeguards that keep results reliable. Attendees will leave with a clear view of what becomes possible when the web is rewired for agents, not ad clicks.

	TRACK A: MILKY WAY	TRACK B: ANDROMEDA
11:35 - 12:00	**How to Cheat at Benchmarking Search Engines** Delivery Hero Roman Grebennikov, Principal Engineer Every week, a new search engine claims to be blazingly fast, backed by benchmarks, of course. But these benchmarks are often run on different hardware, with different configurations, and on different datasets, making fair comparisons between engines impossible. Even when you fix all the variables across competing engines (like ANN-benchmark does), you're still not comparing apples to apples. Each engine has its own set of optimal parameters, which skews the results. In this talk, we'll share the story of building a reproducible benchmarking harness and a public leaderboard designed to fairly compare the performance of modern search engines.	**Building Scalable AI Memory for Agents Across Graphs and Vectors** Cognee Vasilije Markovic, Founder Connecting company data to LLMs looks simple—until scale, speed, and reliability collide. In practice, teams face data consistency issues, brittle contracts, microservice sprawl, and developer experience gaps. This talk presents cognee: a scalable AI Memory Python SDK that abstracts storage and retrieval over multiple graph and vector databases while integrating cleanly with various LLMs. We'll show how memory is modeled, segmented, and synchronized across structured and unstructured sources, using a finance example that combines documents, records, and relationships. You'll learn design choices behind multi-backend support, how embeddings and graph context complement each other, and patterns for low-latency retrieval with robust versioning. We'll share approaches to modularity, testing, and DX that reduce fragility and accelerate iteration. You'll leave understanding what "AI memory" truly means, how to give agents durable, relevant context, and how to evolve your architecture without lock-in.
12:00 - 12:25	**Hands-On GraphRAG: Using Knowledge Graphs to Improve Retrieval Grounding** Neo4j Martin O'Hanlon, Technical Curriculum Developer Generative AI needs reliable grounding to be truly useful. This practical session introduces GraphRAG—combining knowledge graphs with RAG to add structure, relationships, and provenance to retrieved context. You will discover how to build a knowledge graph from unstructured text and see how graph context improves retrieval precision and explainability. You'll see where vector search is effective, where it struggles, and how graph traversal can improve context. You will see how GraphRAG tools can be integrated into a simple LangChain agent. You will learn how graphs can help your GenAI project by providing richer and transparent context.	**Evaluate Your Qdrant-RAG Agents with No-Code n8n Evaluations** n8n Marcel Claus-Ahrens, Automation Expert, n8n Ambassador Great RAG agents need continuous evaluation—not just intuition. In this live, no-code session, we'll build an agent with n8n and Qdrant, index a small knowledge base, and validate it using n8n's native evaluation methods. We'll demonstrate LLM-as-a-Judge with an example set of ideal answers, show how to track regressions over time, and wire alerts for quality drift. You'll learn to design test sets, measure citation faithfulness, and interpret failure cases like missing context or brittle prompts. By session end, you'll have a repeatable workflow to ship agents with confidence, even without writing custom evaluation infrastructure.
12:25 - 12:45	**Beyond Text-Only: Llamaindex Retriever with Superlinked's Mixture of Encoders** Superlinked Filip Makrauli, ML Engineer, DevRel This session surveys Google DeepMind's latest embedding solutions for retrieval: the state-of-the-art Gemini Embedding and the open EmbeddingGemma models. We'll explain task-type controls, dimensionality options, and how to integrate these models into vector search stacks from basic prototypes to tuned production systems. You'll learn patterns for prompt-task alignment, indexing choices that balance accuracy and latency, and knobs for optimizing domain performance. By the end, you'll know where each model shines and how to deploy them confidently for advanced retrieval tasks.	**Self-Improving Evaluations for Agentic RAG: Tracing and Feedback Loops** Arize AI Dat Ngo, AI Architect & Director Solutions, EMEA GoodData's engineering team will open up their production-grade Python microservices and RAG architecture that power natural-language analytics. We stream semantic objects—datasets, visualizations, dashboards—into Qdrant in near-real-time, then apply high-throughput similarity search and re-ranking to assemble precise prompts that ground LLM responses. The session covers practical ingestion patterns, schema design for analytics artifacts, and retrieval optimization techniques that balance latency, recall, and cost. We'll share how Langfuse-based reliability testing exposes failure modes early, plus the metrics and tuning levers that moved the needle in production. Expect hard-won lessons on scaling, observability, and guardrails for consistent insights. Whether you're building an analyst copilot or embedding AI into BI workflows, you'll learn reusable patterns for context construction, prompt packaging, and response validation. Join us to see how real-time data, vector search, and disciplined evaluation combine to deliver robust, context-aware AI assistants for modern analytics.

12:45 - 1:45	LUNCH DOWNSTAIRS

	TRACK A: MILKY WAY	TRACK B: ANDROMEDA
1:45 - 2:10	**Vision-Language Models: A New Architecture for Embedding Models** Jina AI Michael Günther, Senior Research Scientist Vision-Language Models (VLMs) use tranformer architectures to learn from mixed text-image inputs and increasingly serve as strong backbones for embedding models such as jina-embeddings-v4. We'll share training insights for VLM-based embedding models that support both dense (single-vector) and late-interaction (multi-vector) retrieval across domains, tasks, and languages. Particular attention will go to images containing both text and diagrams, UI screenshots, and illustrations, for which VLMs excel. We'll unpack factors like image resolution, retrieval objectives, the influence of the so-called "modality gap" on retrieval performance, and how we evaluated the model. Operational efficiency also matters, so we'll compare post-training quantization and quantization-aware training, outlining trade-offs between footprint, throughput, and accuracy.	**Vector Databases Power Workflow Engineering for Context-Rich AI Systems** LlamaIndex Clelia Astra Bertelli, Open Source Engineer AI has shifted from prompt tweaks to orchestrating full, context-aware workflows. As applications grow more complex, it's no longer enough to craft better prompts—you must structure, persist, and retrieve the right context at the right time. This talk frames "workflow engineering" as the discipline that operationalizes AI systems with guardrails, observability, and reliable state across runs. We'll focus on two capabilities vector databases unlock. First, state management and persistence: storing and fetching workflow context to enable resilient, repeatable behavior and recovery. Second, long-term memory for LLMs and agents: capturing and retrieving durable memories so systems improve over time and adapt to evolving tasks. Along the way, we'll connect high-level architecture to production-minded patterns, including indexing strategies, memory chunking, and retrieval design with concise code examples. You'll leave with a practical model for using vector databases as the backbone of workflow engineering—building AI that is better orchestrated, stable in production, and measurably more helpful.
2:10 - 2:35	**Practical Multimodal Embeddings: Video Recommendations and Cross-Modal Search** TwelveLabs Hrishikesh Yadav, Developer Advocate This session walks through real developer workflows that embed text, audio, images, and video to power recommendations and semantic search. Using the TwelveLabs Embed API (Marengo-retrieval-2.7), we'll generate multimodal embeddings and store them in Qdrant for fast similarity queries with metadata filtering. We'll build a video recommendation engine that replaces shallow metadata matching with semantic relevance, then explore use-cases for the cross-modal retrieval such as image-to-video and audio-to-video. Along the way, we'll cover chunking strategies for long video, labeling tips, and evaluation methods that reflect user intent. You'll leave with actionable patterns for indexing pipelines, schema design, and retrieval queries that deliver meaningful results across modalities. Prerequisites are basic embedding concepts and Python familiarity; the content targets intermediate practitioners.	**Agent-Powered Retrieval with Haystack and Qdrant: Promise and Pitfalls** deepset Bilge Yücel, Developer Relations Engineer Are agents the future of retrieval or an overcomplication? This session compares traditional retrieval pipelines with agent-powered approaches built using Haystack and Qdrant. We'll construct a traceable, debuggable agent that plans tool usage, selects retrieval strategies, and explains its steps. Then we'll evaluate latency, accuracy, and robustness against a strong non-agent baseline on a movie dataset stored in Qdrant. You'll learn to read agent traces and decide when agents add value—complex multi-step tasks, ambiguous queries, or dynamic toolsets—and when a simpler pipeline is preferable.
2:35 - 3:00	**High-Throughput, Low-Latency Embedding Pipelines for Real-World Applications** Baseten Rachel Rapp, Head of EMEA Embeddings power RAG, search, agents, and recommendations—but production reality is a different story. This talk distills patterns from companies running embedding inference at scale. We'll map where latency and throughput degrade and discuss architectural fixes, as well as model selection trade-offs, dimensionality, and quantization considerations. Finally, we'll share open-source tools that can boost any embedding API, along with deployment tips for compound AI systems where multiple models and tools coordinate. You'll leave able to diagnose bottlenecks, design resilient pipelines, and ship faster systems without overspending.	**Scaling Real-Time RAG for Analytics** GoodData Jan Soubusta, Field CTO GoodData's engineering team will open up their production-grade Python microservices and RAG architecture that power natural-language analytics. We stream semantic objects—datasets, visualizations, dashboards—into Qdrant in near-real-time, then apply high-throughput similarity search and re-ranking to assemble precise prompts that ground LLM responses. The session covers practical ingestion patterns, schema design for analytics artifacts, and retrieval optimization techniques that balance latency, recall, and cost. We'll share how Langfuse-based reliability testing exposes failure modes early, plus the metrics and tuning levers that moved the needle in production. Expect hard-won lessons on scaling, observability, and guardrails for consistent insights. Whether you're building an analyst copilot or embedding AI into BI workflows, you'll learn reusable patterns for context construction, prompt packaging, and response validation. Join us to see how real-time data, vector search, and disciplined evaluation combine to deliver robust, context-aware AI assistants for modern analytics.

	TRACK A: MILKY WAY	TRACK B: ANDROMEDA
3:00 - 3:20	**Vector Search with Gemini Embedding and Open EmbeddingGemma Models** Google DeepMind Patrick Löber, Senior Developer Advocate This session surveys Google DeepMind's latest embedding solutions for retrieval: the state-of-the-art Gemini Embedding and the open EmbeddingGemma models. We'll explain task-type controls, dimensionality options, and how to integrate these models into vector search stacks from basic prototypes to tuned production systems. You'll learn patterns for prompt-task alignment, indexing choices that balance accuracy and latency, and knobs for optimizing domain performance. By the end, you'll know where each model shines and how to deploy them confidently for advanced retrieval tasks.	**Redefining Long-Term Memory Ingestion for Streaming, Autonomous Agents** Equal M K Pavan Kumar, Distinguished AI Architect Autonomous and multi-agent systems need durable memory without blocking real-time work. We present a fully asynchronous, streaming-driven memory platform powered by Qdrant. Conversations and events are ingested in real time, then structured into full histories, semantic summaries, and rich metadata. Retrieval becomes layered and fast: filter by metadata first, surface summaries for context, and deep-link into detailed history only when needed. The architecture plugs into any streaming engine–Kafka, RabbitMQ, ActiveMQ–and emphasizes low latency, backpressure resilience, and scalable retention policies. You'll learn how this design improves decision quality, reduces prompt bloat, and delivers enterprise-grade responsiveness for agents operating continuously.

3:20 - 3:40	BREAK

LIGHTNING TALKS

3:40 - 4:40	**From Massive Text Streams to Searchable Knowledge with Apache Kafka and Qdrant** **bakdata** Jakob Edding & Raphael Lachtner, Software Engineers **Optimizing GDPR Compliance Retrieval with Hybrid Graph-Augmented RAG Systems** **KI Reply** Ramakant Agrawal, Senior Manager **Multimodal Vectors for Instant, Personalized Product Discovery** **iCompetence** Marc-André Lampe & Lionel Schulz, Head of Experience Orch. & Prod. Innovation **Voice-First Multimodal Search: Raiva and Qdrant as the Query Interface** **Raiva Technologies** Hermann del Campo, CTO **Stories from the AI Search Frontier** **Superlinked** Daniel Svonava, Founder

CLOSEOUT

4:40 - 5:00	Think Outside the Bot Hackathon Awards Closing Remarks
5:00 - 10:00	After Party 🎉