

Title: AI-Powered Data
Validation: Defending Open Data
with Intelligence

Speakers: Davide Montesin ·
Michele Magri · Hamza Jamil

AI-POWERED DATA VALIDATION

Defending Open Data with Intelligence





Title: A Lifelong Passion for Computer Science and Hacking

- I've gone from counting the bits and bytes that fit into CPU register in the '90 to working with AI.
- Always curious about how things work under the hood.
- Founder of Catch Solve, a tech company focused on software and data quality through automated testing made simple



Hamza Jamil

I'm a software developer, passionate about free and open-source software.

- Loves experimenting with new frameworks and modern technologies.
- Enjoys turning complex ideas into simple, usable projects.
- Builds seamless API integrations.

GitHub: [Hamza5955](#)



Michele Magri

Passionate about anything that has tech in it, my software engineering experience started here at the NOI Hackathon.

- Experience with CI/CD, Cloud Computing and Self-Hosting.
- Passionate about hardware engineering and PCB design.
- Advocate for automation

Contact: commercial@syswhite.dev

GitHub: [@SysWhiteDev](https://github.com/SysWhiteDev)

Title: The operational imperative for data integrity

- On Open Data Hub, quality underpins analytics, data exchange, and decisions.
- Poor quality increases friction, causes costly errors, and weakens governance alignment.
- High integrity data is essential for reliable performance and system stability.
- Data quality isn't optional, it's foundational to trust and adoption.

Title: Pillars to evaluate data quality before solutions

- **Completeness:** No essential data missing (nulls, mandatory fields).
- **Accuracy:** Alignment with trusted reference sources.
- **Consistency:** Uniform representation across systems and datasets.
- **Validity:** Types, ranges, patterns, and domain rule compliance.
- **Uniqueness:** No duplicate entities or overlapping keys.
- **Integrity:** Correct relationships, referential links, and structure.

Metadimensions for usability: **Freshness** (uptodate) and **Usefulness** (interpretable, accessible, aligned to needs).

Title: Opensource, AI-powered validation for data and images

- Data is the foundation of innovation; quality is key.
- Catch Solve, with AtomHR, presents Data Defender.
- Automated checks for datasets and images with explainable results.
- Open-source to adapt, extend, and trust.



Title: Proven on Open Data Hub, portable everywhere

- Demonstrated on Open Data Hub (ODH) datasets and APIs.
- Plugandplay adapters for REST, files, and streams.
- Customize rules and models to your domain.
- Integrate with CI/CD and monitoring platforms.

Title: Benefits of largescale anomaly scanning

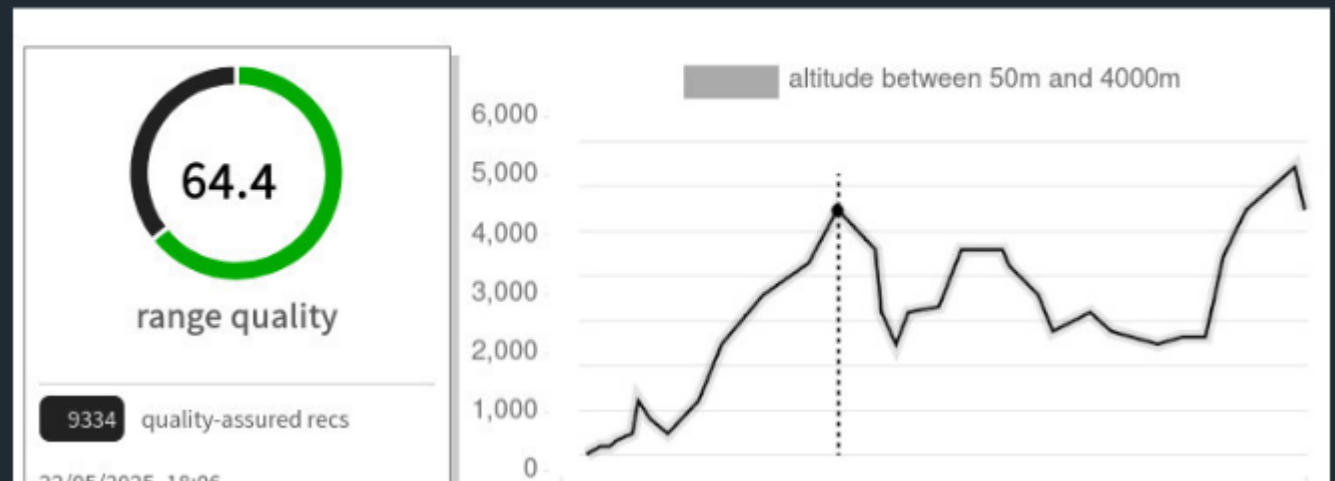
- Analyzes massive volumes to surface rare patterns and outliers.
- Finds anomalies that manual or sampled checks would miss.
- Cuts investigation time by prioritizing what matters.
- Early alerts and trends: quality and reliability improve over time.



Title: AI + automation + datadriven philosophy

- Continuous monitoring of freshness, schema, and content quality.
- Early detection of anomalies, drifts, and image defects.
- Guided, efficient corrective actions and notifications.
- Faster cycles and higher trust in data assets.

Tracking data evolution



Title: Open data, enterprise APIs, internal systems

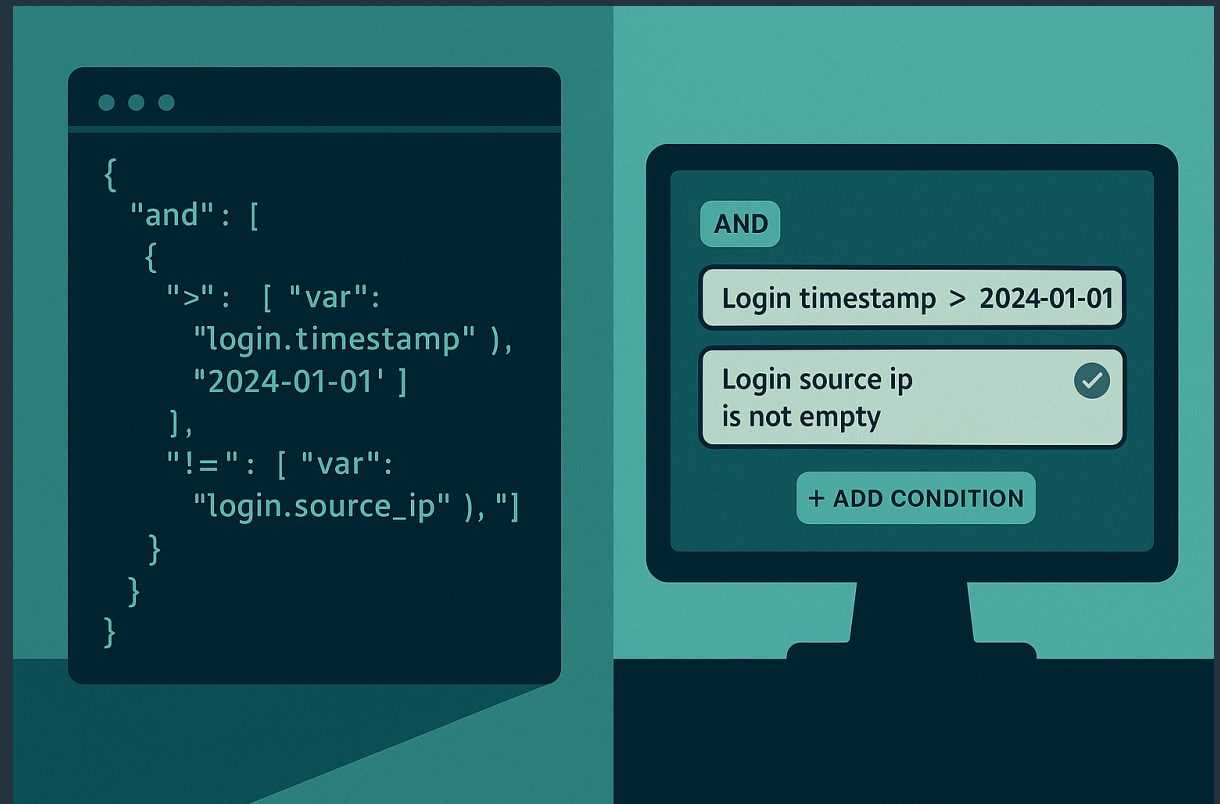
- Public open datasets and portals.
- Mission-critical enterprise and partner APIs.
- Internal data lakes, warehouses, and ETL pipelines.
- Image catalogs and computer-vision pipelines.

Title: Simpler validation, better quality over time

- Opensource + AI working together to validate.
- From checks to insights and actionable guidance.
- Less firefighting; more real problemsolving.
- Smart, sustainable validation accessible to all.

Title: Nocode rules with a visual interface

- Design validations using JsonLogic expressions in a UI.
- Enable nondeveloper users to author and review rules.
- Preview results on sample data; explain failures inline.
- Store rules as JSON for portability, audit, and governance.



Title: Flexible rules with code

- Write validations as simple JS functions/assertions.
- Reuse helpers for dates, geo, schemas, and images.
- Version control rules with tests and CI feedback.
- Compose rule sets per dataset, API, or pipeline step.

```
<check>
  <name>GpsInfo must be inside
  <category>GpsInfo must be
  <description>This rule checks
    This rule checks that
    Scope filters are available
  </description>
  <data_scope_filters>
    <data_scope_filter>
      data_quality
      <rule_expression>
        !Array.isArray(
          $.Latitude
          $.Longitude
          $.Longitude
        )
      </rule_expression>
    </data_scope_filter>
  </data_scope_filters>
</check>
```

Validation logic

FILTER

Gps Info



is in a list

34 <= Latitude <= 72

**-25 <= Longitude
<= 45**

Title: From structural correctness to semantic understanding

- LLMs assess whether data makes contextual sense, not just format compliance.
- Example: plausibility of EV charger positions and status across sources.
- Expands checks from syntax to meaningimproving relevance and trust.

Title: Visualcontent consistency checks

- Verify captions and descriptions match what's visible in the image.
- Detect missing or mismatched objects, scenes, or attributes.
- Provide confidence scores and short explanations.
- Improve accessibility: assess alt-text quality and color contrast.

Ensuring text and images tell the same story



Mix-up:
description
says "In the
city centre"

Title: Multilingual translation quality

- Validate semantic equivalence across languages for titles and descriptions.
- Flag mistranslations, omissions, or tone inconsistencies.
- Return scores and concise rationales for issues.

Rating 0.0: {'de': '-', 'en': '-', 'it': '-'}

Rating 5.0: {'de': 'Zirbenzimmer', 'en': '..', 'it': '..'}

Rating 5.0: {'de': 'Frühstück', 'en': 'Frühstück', 'it': 'Frühstück'}

Rating 5.0: {'de': 'Marende', 'en': 'foot', 'it': 'merenda'}

Rating 10.0: {'de': 'Lounge', 'en': 'Lounge', 'it': 'Lounge'}

Title: Learn normal patterns; surface outliers

- Models detect unusual values, shifts, and drifts that rules may miss.
- Automated retraining workflows keep models aligned with changing realities.
- Monitoring dashboards track performance and alerts.

Complements deterministic rulesbroad coverage with adaptive intelligence.

Title: Synthetic data and stress testing

- Generate realistic test data and hard negatives to probe rule robustness.
- Automated quality scoring with concise explanations of issues.
- Guided remediation suggestions to close the loop.

Title: Hybrid approach with managed LLMs

- **Hybrid design:** Deterministic rules as enforcement backbone; LLMs add context.
- **Manage probabilistic behavior:** Prompt versioning and response traceability.