

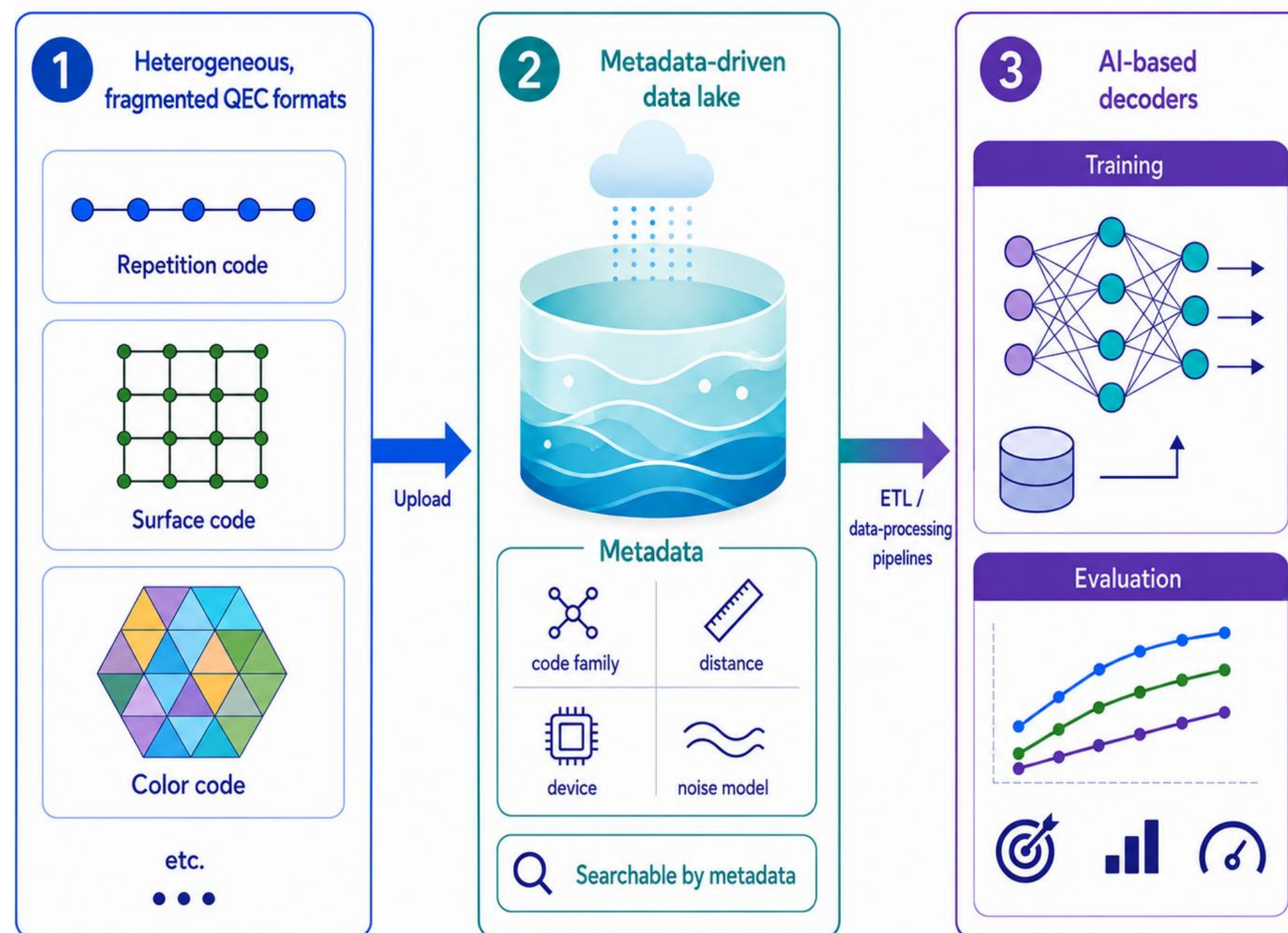
1. Why a domain-specific data lake for QEC?

- QEC is required for fault tolerant quantum computing.
- There is a shift from classical decoders to AI-based decoders.
- This necessitates infrastructure to support QEC data.

CHALLENGES

QEC datasets have

- heterogeneous schemas,
- are **fragmented** (different file formats), and
- real datasets are **scarce**.



What does QEC data look like?

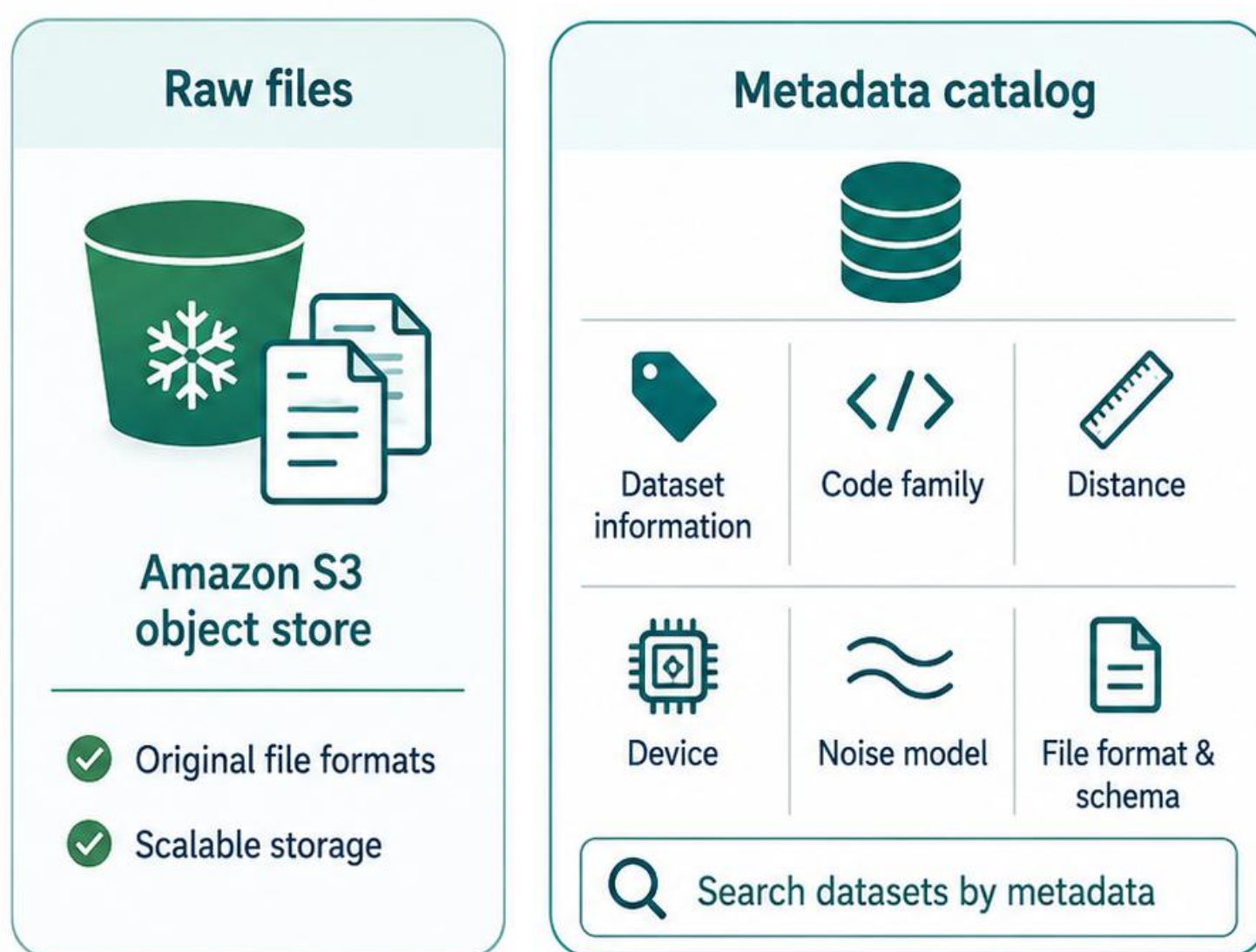
$d^2 - 1$ stabilizers soft information

shot_id	round	s_1	s_2	...	s_{d^2-1}	LLR	p_read
001	0	0	1	...	1	1.85	0.94
001	1	1	0	...	0	2.31	0.91
002	0	1	1	...	0	—	—
002	1	0	1	...	1	—	—

Our contributions

- An infrastructure for QEC dataset storage and management.
- Facilitate QEC data processing: ready for decoding, calculating statistics, etc.
- One place for all sources of QEC data (real, simulated, synthetic).

2. Storing and Managing QEC data



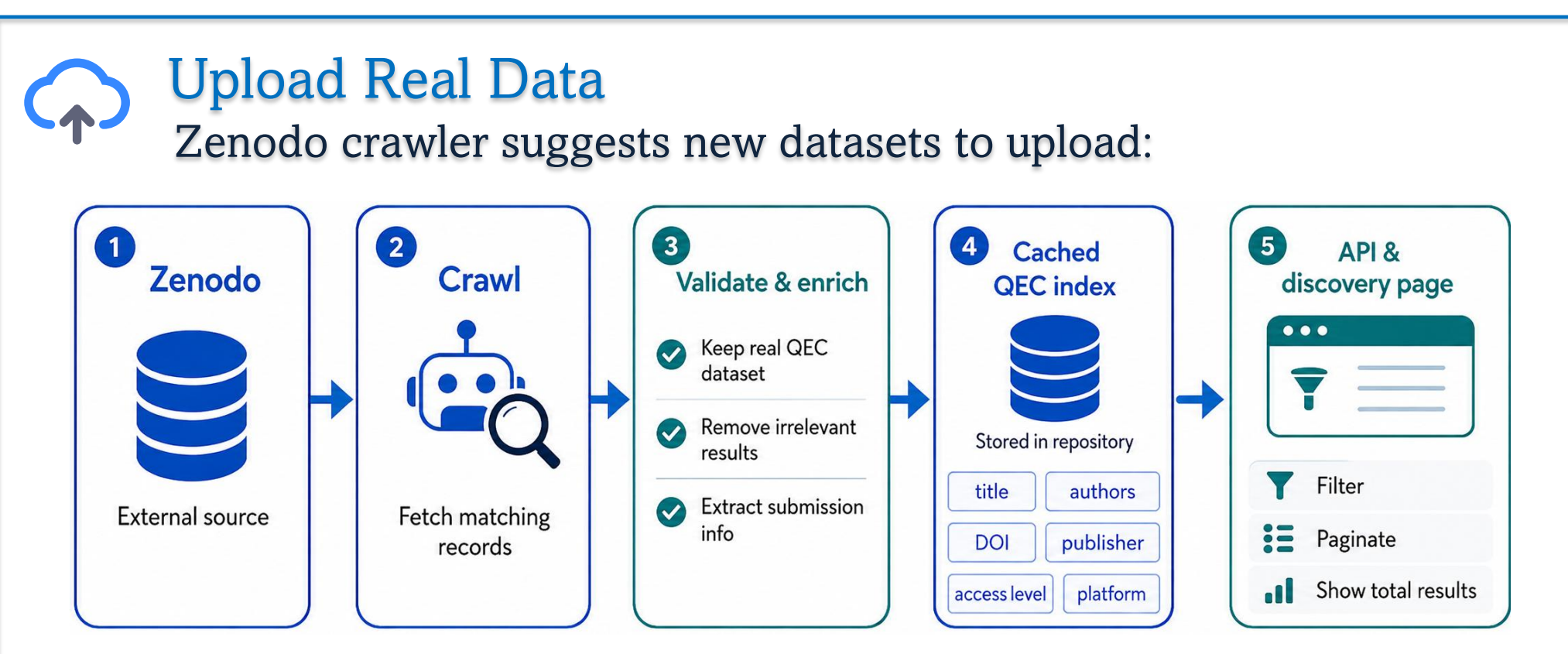
- Raw datasets are stored in their original file formats.
- A separate metadata catalog stores searchable dataset attributes, allowing datasets to be found without scanning the raw files.

3. Data Processing

Pre-defined ETL pipelines save time on manual data processing

- Select** an existing or upload a new dataset to process.
- Normalize**: sort and ensure identity columns are present.
- Derive** syndrome or detector data if missing.
- Output** format selection (e.g. flat tabular).
- Register** the newly processed data in the catalog for reuse.

4. Filling the Lake



Generate Data

- Simulate datasets with STIM and custom noise models.
- Expand real datasets with synthetic data using a diffusion model: WaveStitch [3].

5. When do I need the QEC Lake?

- I need a simulated color code dataset**
Search the catalog by keywords (color code), filter by metadata (simulated).
- What does this dataset actually contain?**
Ask the QEC Assistant to explain its code, noise model, syndrome representation, and other metadata in natural language.
- I want to know how different AI-based decoders perform on my dataset**
Use our mini model zoo to quickly get AI-based decoder results and compare with classical baselines.
- QEC data is needed for my own processing pipeline**
Export your (processed) data to the desired format (flat tabular, sequence, ...)

Takeaway: A well-engineered data lake enables efficient data discovery, management, and domain-specific computation.

- [1] Aadi Patwardhan, Max de Groot, Andrei Ilinescu, Aditya Shankar, Yuandou Wang, Wenbo Sun, Floris Geerts, and Rihan Hai. QEC lake: Data Lake for AI-enhanced Quantum Error Correction. PVLDB, 19(12), 2026.
- [2] Rihan Hai, Christos Koutras, Christoph Quix, and Matthias Jarke. Data lakes: A survey of functions and systems. TKDE, 35(12):12571–12590, 2023.
- [3] Aditya Shankar, Lydia Chen, Arie van Deursen, and Rihan Hai. WaveStitch: Flexible and fast conditional time series generation with diffusion models. SIGMOD, 3(6):1–25, December 2025.

