

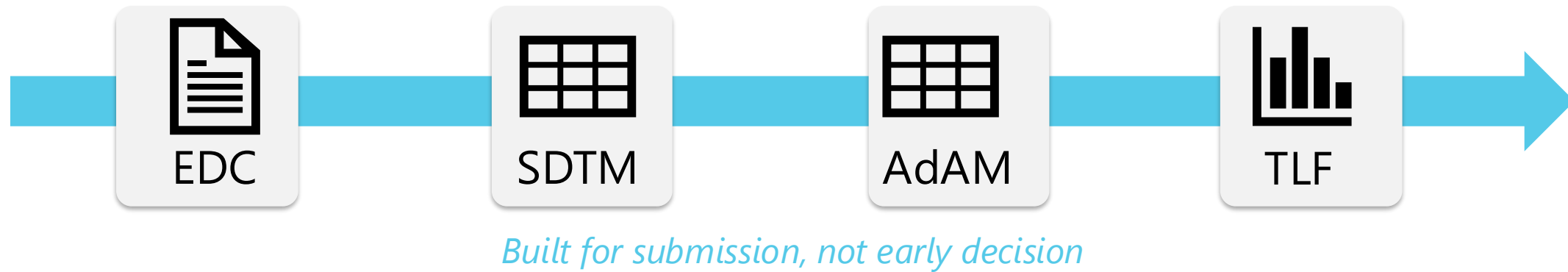


What Pharma Can Learn from Data Science Outside the Industry

Improving decision speed without compromising compliance

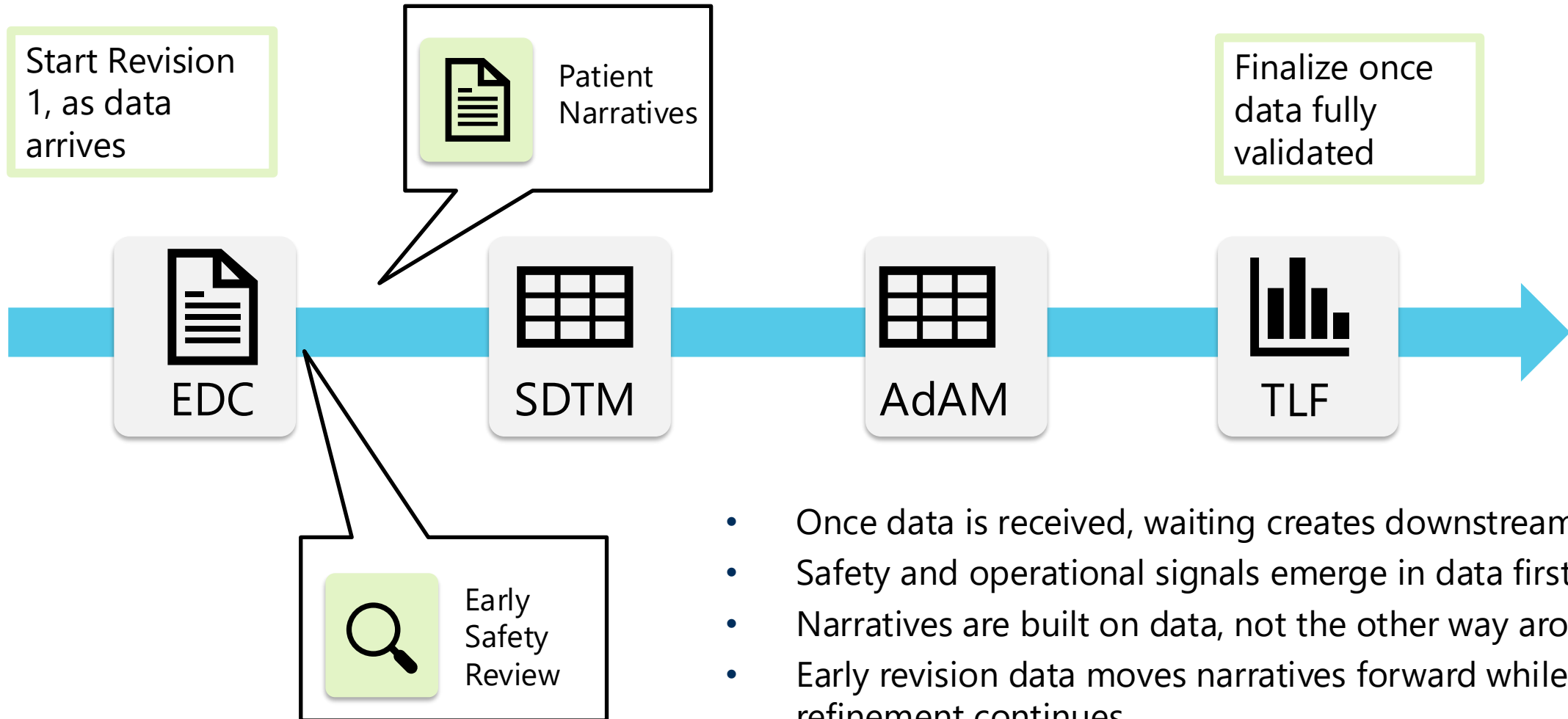


The Core Problem



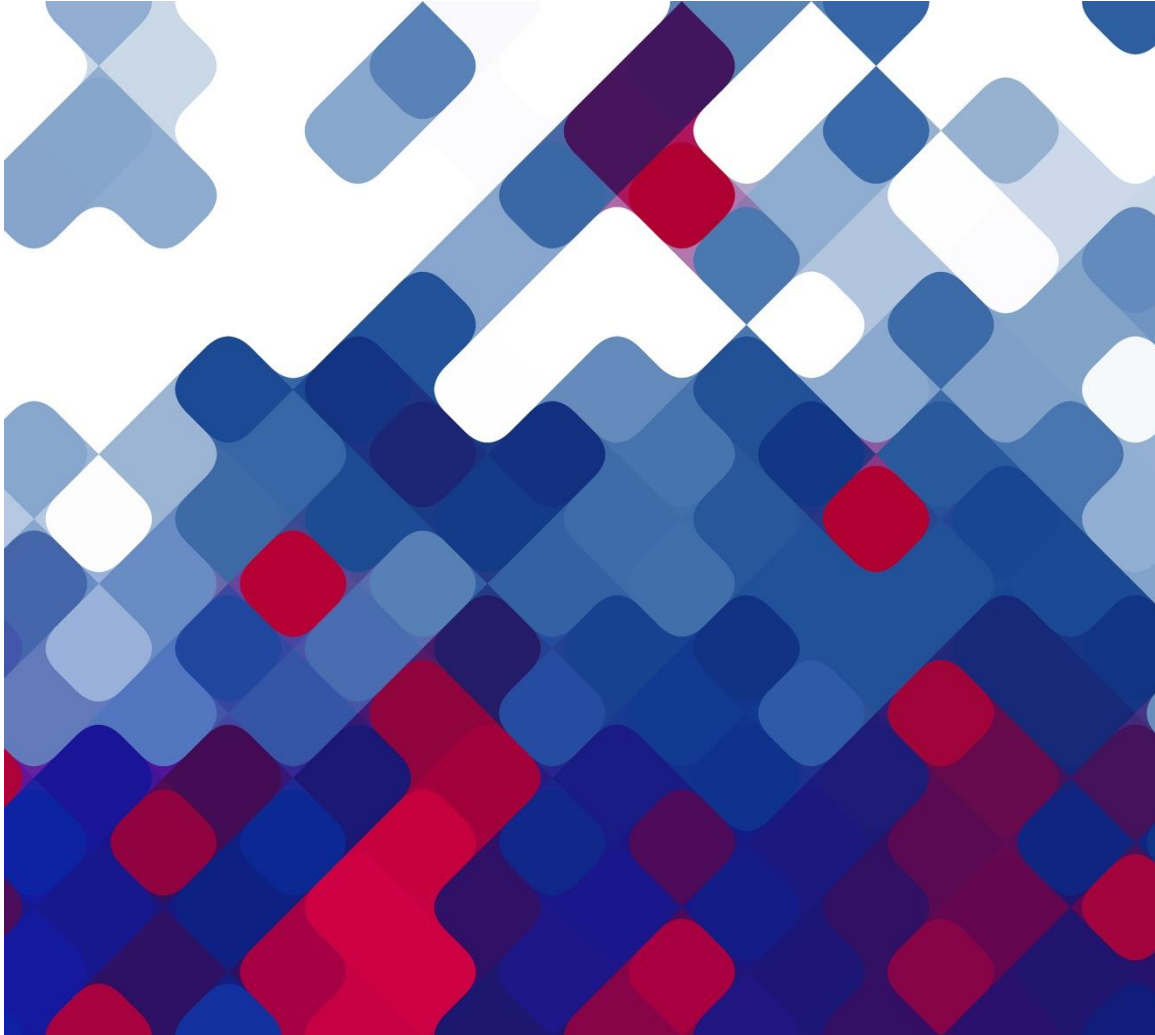
- Pharma data systems prioritize integrity and traceability
- Speed of insight is often compromised
- Waiting for SDTM and ADaM slows decision speed
- Growing data volume makes this increasingly risky

Why Decision Speed Matters



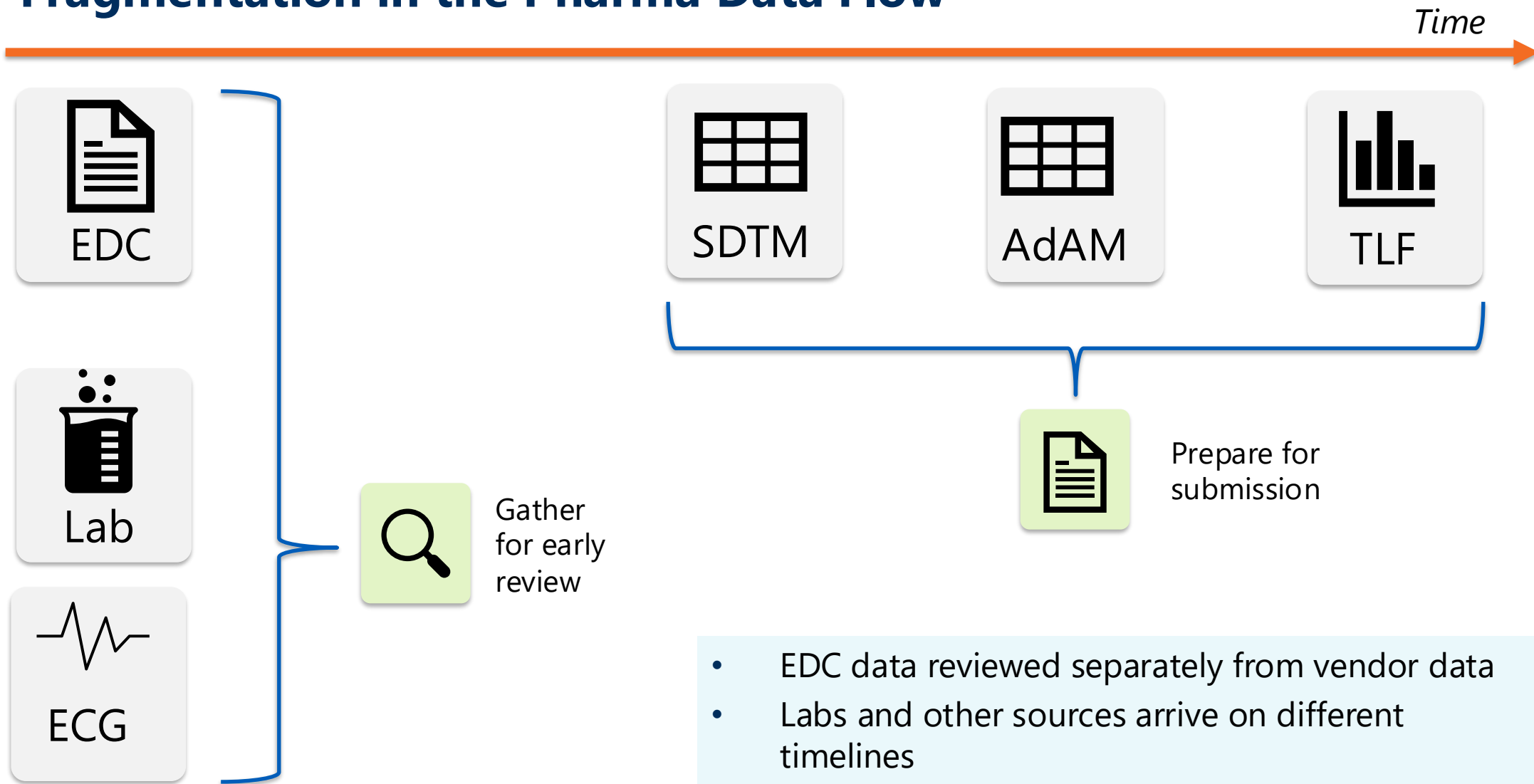
- Once data is received, waiting creates downstream drag
- Safety and operational signals emerge in data first
- Narratives are built on data, not the other way around
- Early revision data moves narratives forward while refinement continues
- Speed and quality coexist with proper engineering and guardrails

Decision Speeds depends on When Data is Ready



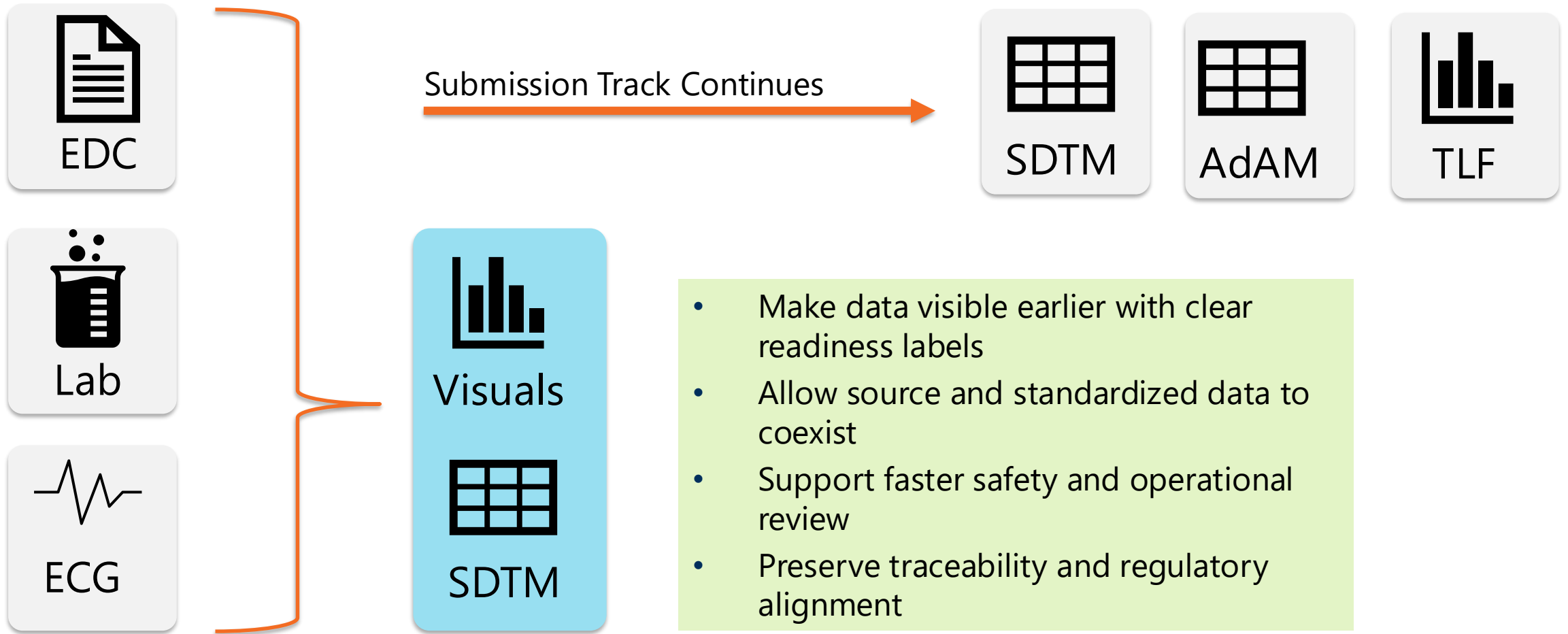
- Decisions are easier when data is clear and consistent
- Early reviews use data that is still changing
- It can be hard to see the full picture at once
- As studies move faster, data arrives more often
- Delays in review increase safety risk and extend study timelines

Fragmentation in the Pharma Data Flow

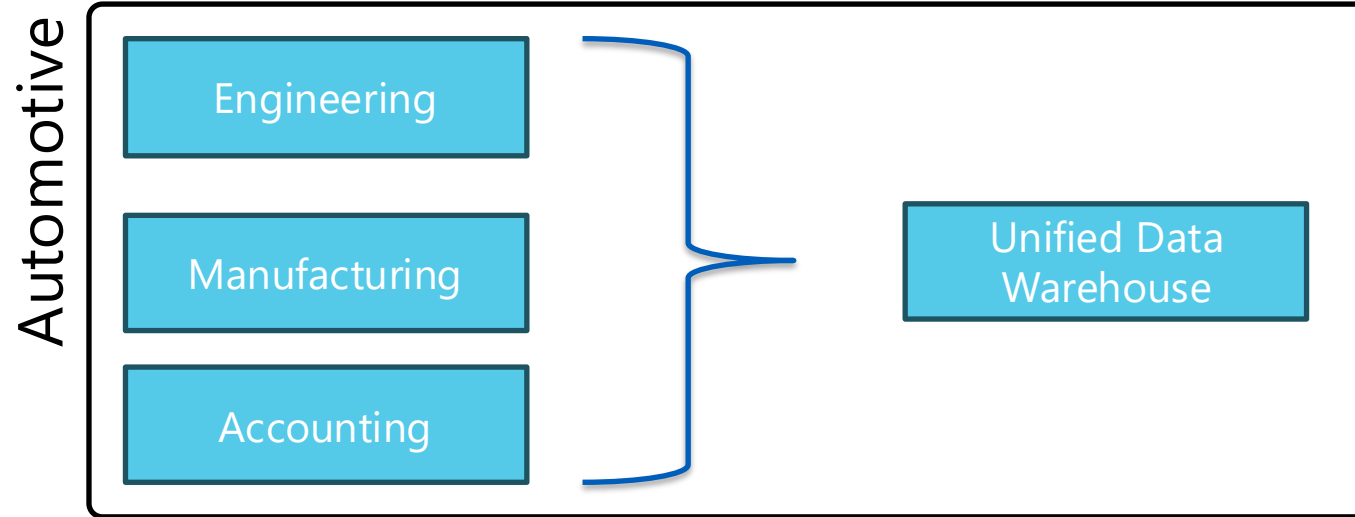


- EDC data reviewed separately from vendor data
- Labs and other sources arrive on different timelines
- SDTM and ADaM are viewed later and elsewhere
- No single, continuous view of study data

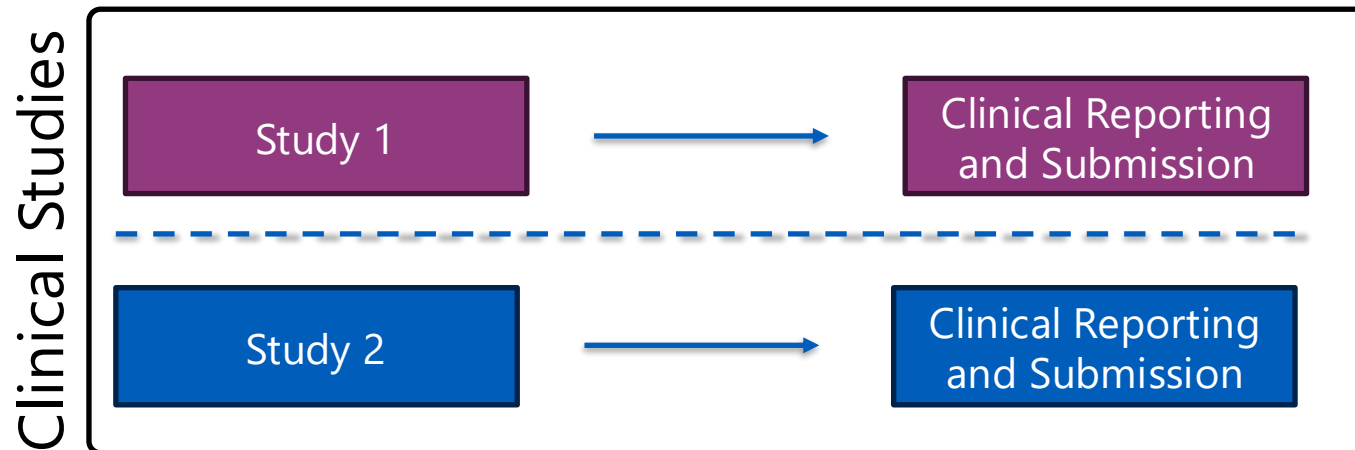
The Opportunity



Differences in Data Culture



- Automotive and manufacturing integrate data continuously
- Multiple systems, inconsistent identifiers, shared reporting
- Pharma prioritizes study isolation and scientific rigor
- Strong controls are essential but can delay insight



Early Data Does Not Mean Low Quality

- Early data can be validated for correctness and repeatability
- Not intended for regulatory submission
- Supports timely safety monitoring and oversight
- Comparable to process validated operational data in manufacturing



Data Engineering's Role in Data Science

- Data science depends on engineered, trusted data
- Ensures repeatability, quality, and operational guardrails
- Owns end to end data movement, from intake to analysis
- Goes beyond file conversion or tooling choices



Why Integration Is Hard (Everywhere)

System #1 – Record Changes

Period	Inventory	Change
Week Begin	100	
Day 1		+10
Day 2		-5
Day 3		+20

System #2 – Records Amount

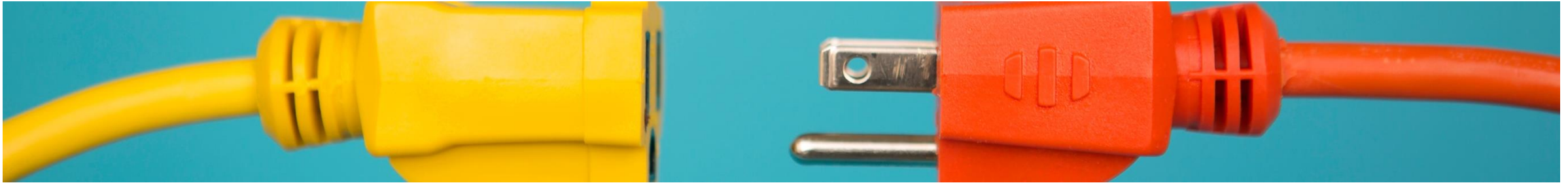
Period	Inventory
Week Begin	100
Day 1	110
Day 2	105
Day 3	125

- Source systems use different formats and meanings
- Data must be transformed, not just combined
- Alignment requires intentional processing

Same Part, different Numbers

System	Part Number
Customer	XF2T-14401-AB
Engineering	PR-1002890
Manufacturing	M-0091002

Why Integration Is Still Hard in Pharma



- Studies are intentionally isolated
- Cross study consistency adds complexity
- Different vendors, formats, and coding rules
- Source systems cannot be fully controlled

Focus on Process, Not Tools

- Do not force a single vendor or platform
- Standardize ingestion and transformation pipelines
- Embrace variation in source data
- Control meaning, outcomes, and provenance



The Medallion Architecture



- Framework for progressive data refinement
- Raw, bronze, silver, and gold layers
- Each layer has a clear purpose
- Enables repeatability and recoverability

Medallion Architecture - Raw



Raw:

- Data is organized by Sponsor/Project/Study
- Receive Data as-is from data source
- If data is received as zip, we store as zip.
- Data is stored by dates, so we retain all copies of data received.
- No alterations or modifications are made to data or data files

Goal: Preserve original data received.

Medallion Architecture - Bronze



Bronze:

- Data is organized by Project/Study within Sponsor Bronze data lake
- Convert received data into “tabular” format (rows and columns)
- Data is converted unpacked from zip and separate converted and stored as separate files.
- Files are converted from CSV, TXT, SASBDAT7, SAS XPT, Excel XLXS formats.
- Various encoding schemes are handled.
- Add tracking fields to each row (load date, provenance id, etc).

Goal: Store uniformly formatted single file data organized by date received.

Medallion Architecture - Silver



Silver:

- Data within the Silver data lake is organized by project under each sponsor, with studies identified by study id column.
- Align data sources to CDISC SDTM naming standards: Tables and Variables.
- **The majority of data mapping occurs at this stage.**
- Lab data may originate from both EDC systems and regional laboratories; these are harmonized and merged within the Silver layer.
- Additional operations include aligning data types, preserving original values, and stacking all study datasets within each project into unified Silver tables.

Goal: Align data into CDISC familiar tables and variable names, stack study data.

Medallion Architecture - Gold



Gold:

- Prepare data for final fit and finish for use in dashboards.
- Create subject dimension for use across all domains. Based on DM, ADSL, and others.
- Create combined disposition tables based on various silver sources. This may get pushed to silver someday.
- Derive values such as AESTDY, which is based on the days between the start of and study start.
- Derive study or project specific variables.

Goal: Finalize data for use within dashboards. This is the “public” view of the data.

What Pharma Can Do Now vs Later?

Do Now

- Adopt clear data layering
- Improve data engineering practices
- Deliver usable data earlier

Plan for Later

- Shared decision ownership
- Semantic governance and cross team alignment

Key Takeaway

- Faster decision speed does not require lower standards
- Build a phased, realistic roadmap

Contact Information

Kris A. Wenzel

MMS Holdings, Inc.

880 Commerce Blvd, Canton Township, MI 48187

kwenzel@mmsholdings.com

www.mmsholding.com

linked in @ <https://www.linkedin.com/in/kriswenzel/>