Quilt

Tech Talks

Building a Successful Data Lake For Biotech

Eric Goldbrener
Principal Consultant
@Goldbrener Software & Media

Aneesh Karve
Chief Technology Officer
@Quilt Data

Talk Overview

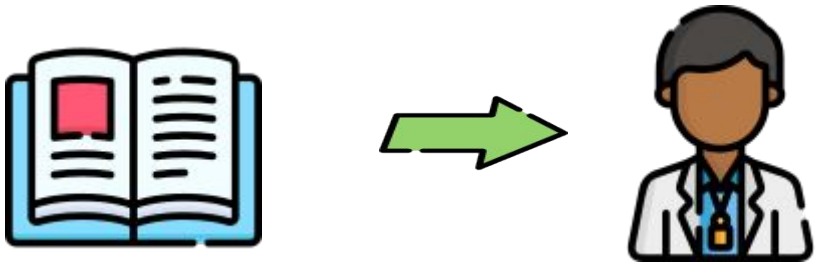
- Data-centric challenges in biotech
- The role data lakes play in the drug discovery process
- Tips, tricks, and ROI for creating a self-organizing data lake

Data Lakes & Drug Discovery

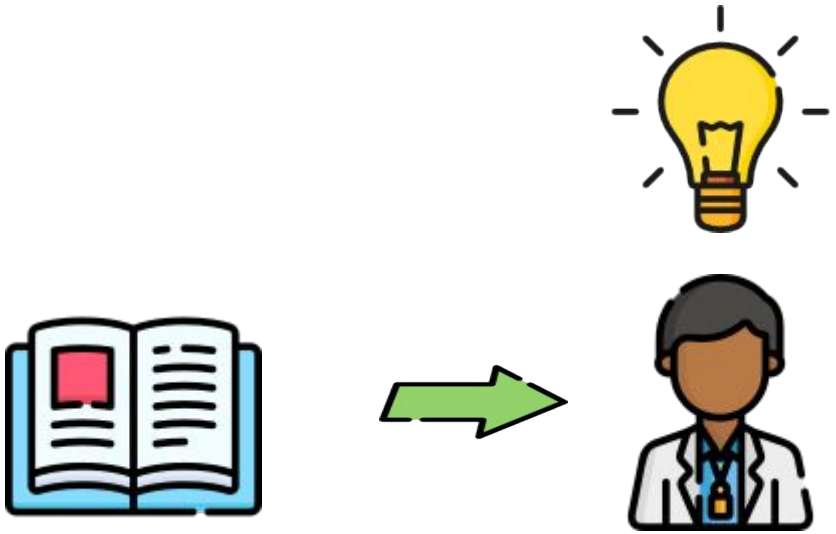
Eric Goldbrener

TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

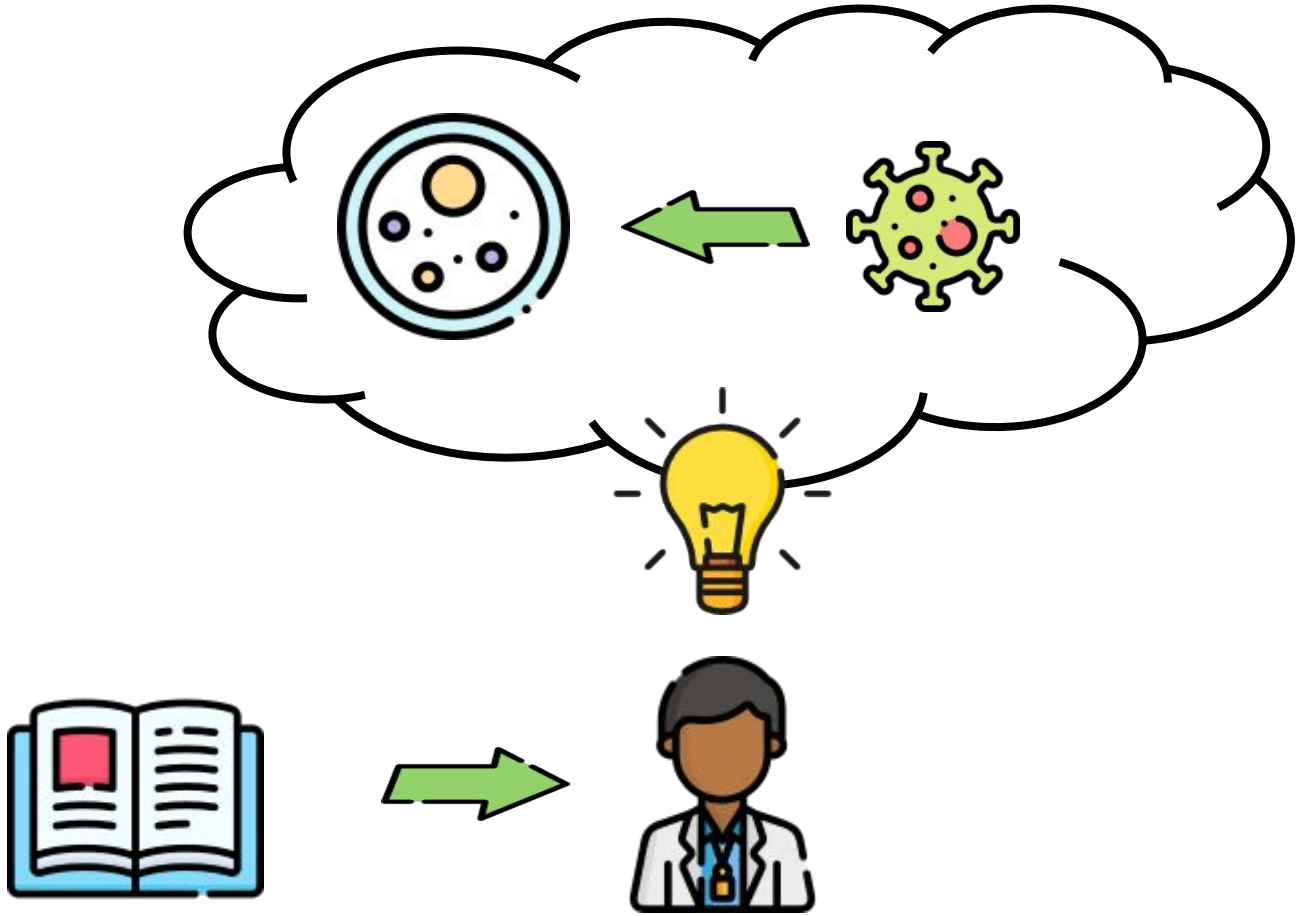
TARGET	*HIT 2 LEAD*	*EARLY DEV*	*CLINICAL*	*MARKET*



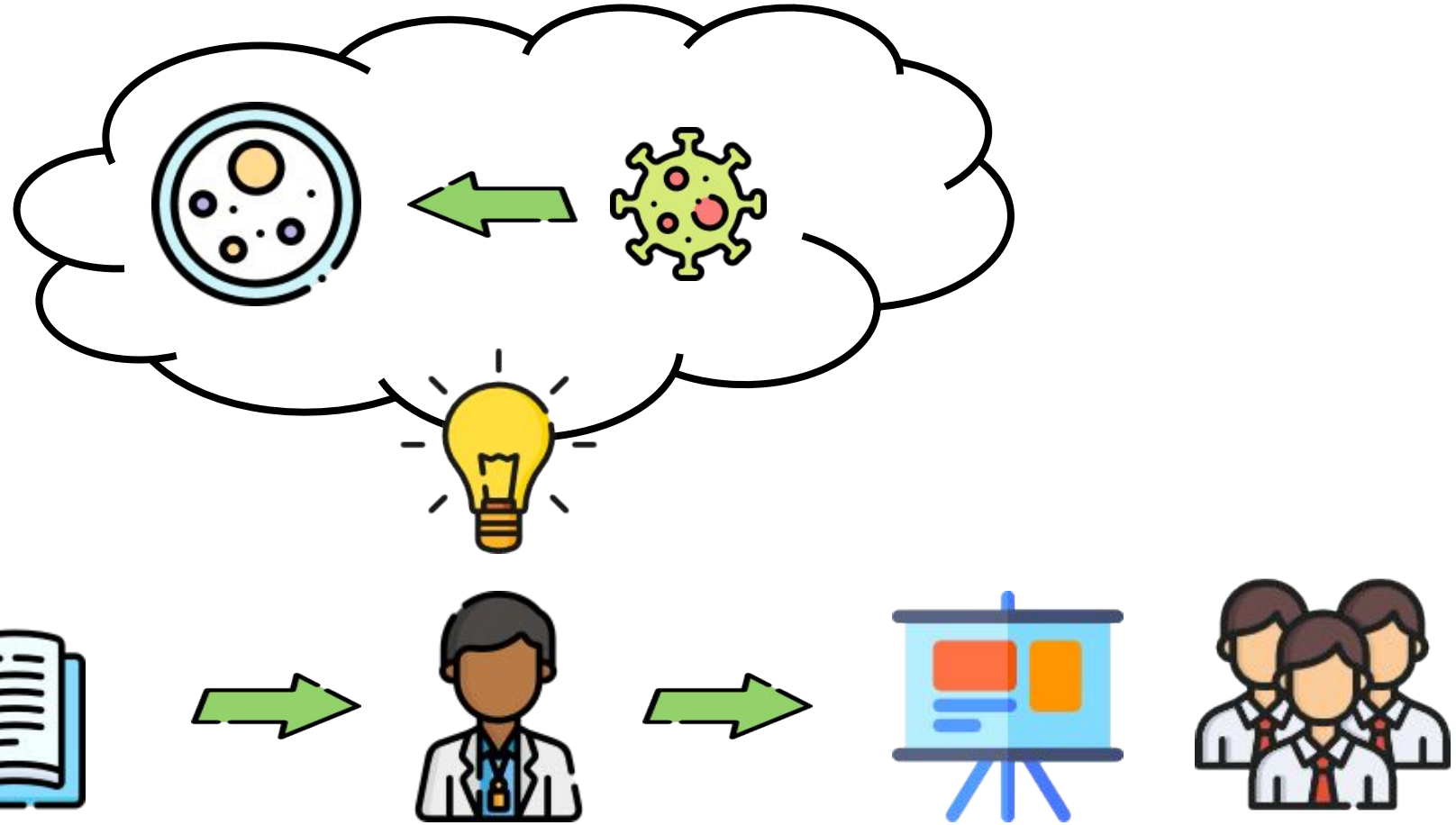
TARGET	*HIT 2 LEAD*	*EARLY DEV*	*CLINICAL*	*MARKET*



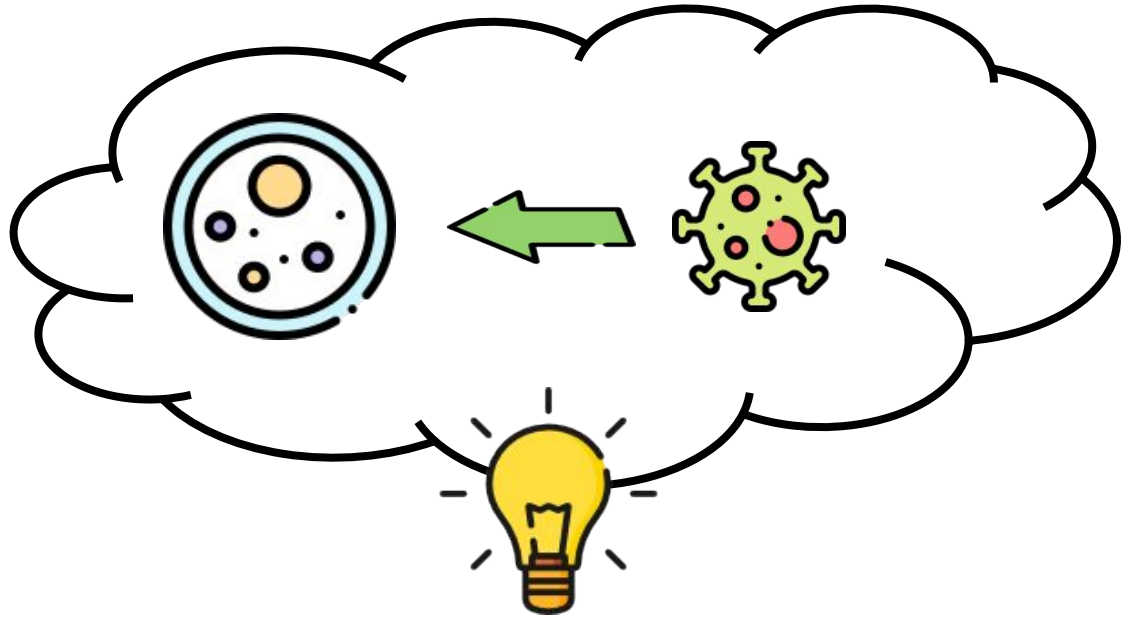
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

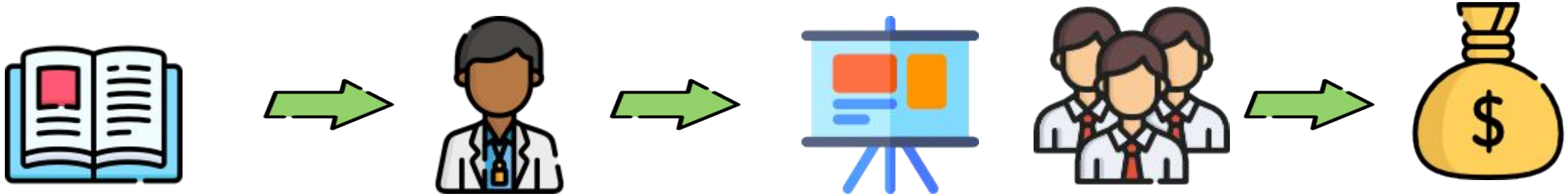


TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

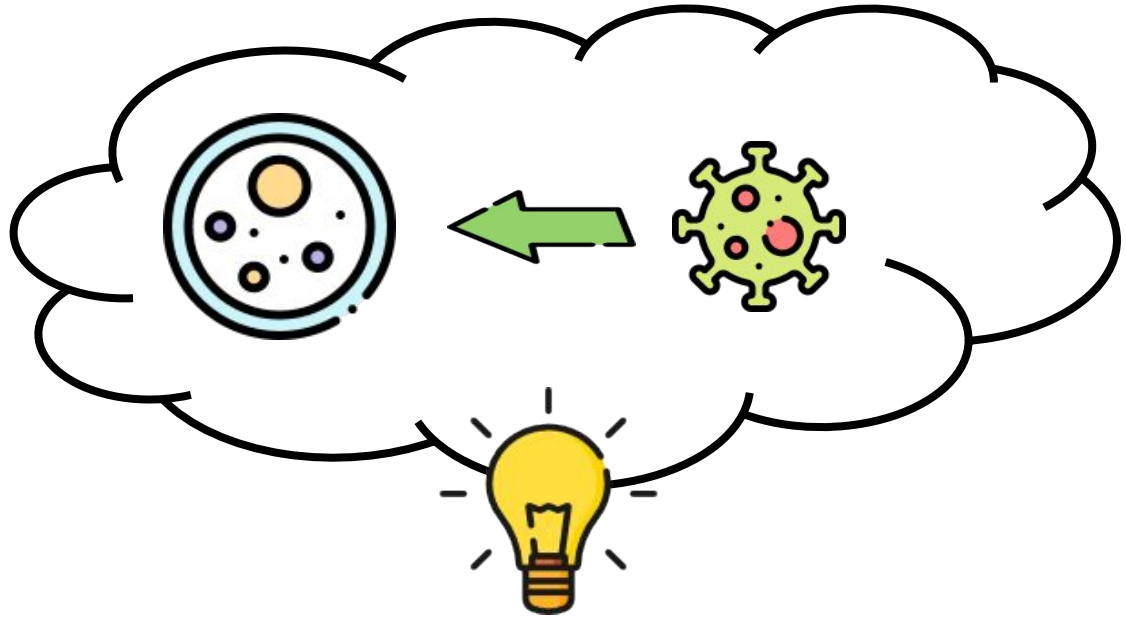


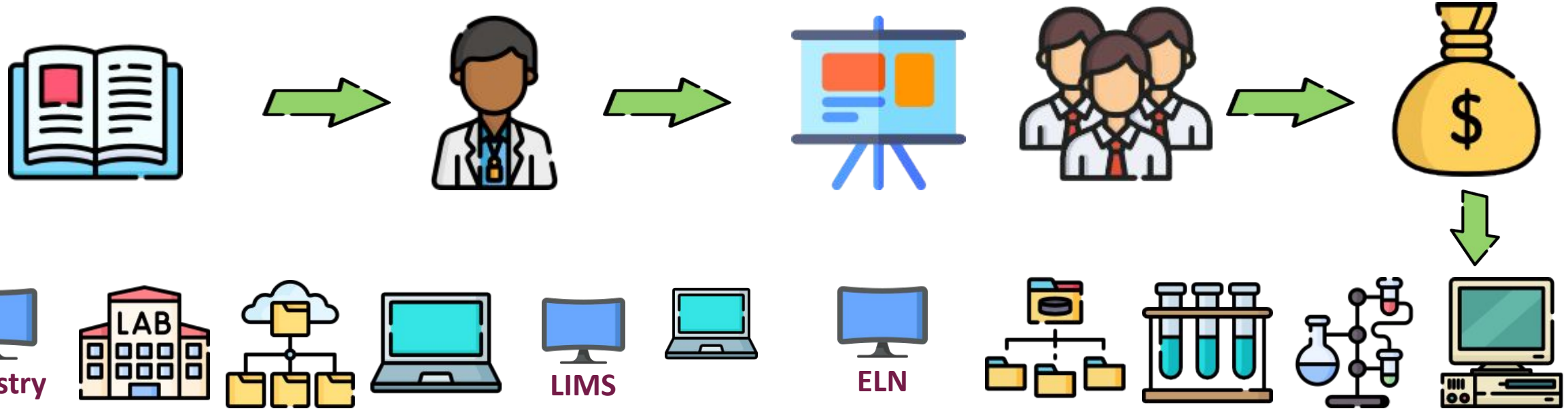
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

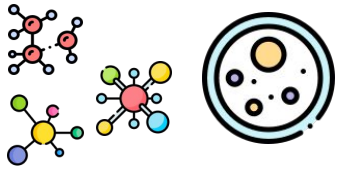


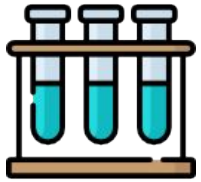


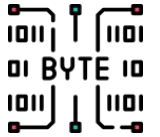
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET





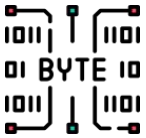




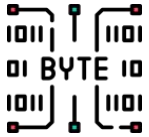


RESEARCH

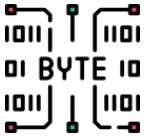
Registry

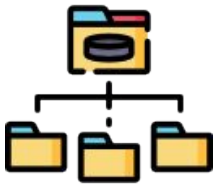


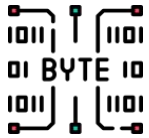
LIMS

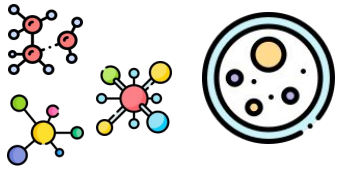


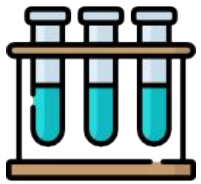
ELN

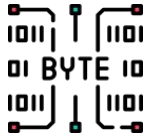






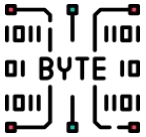




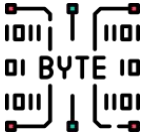


RESEARCH

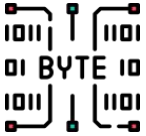
Registry

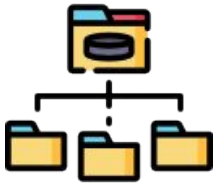


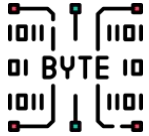
LIMS

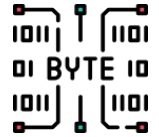


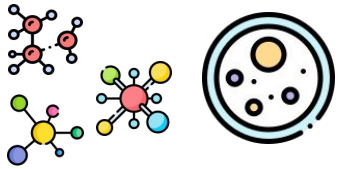
ELN

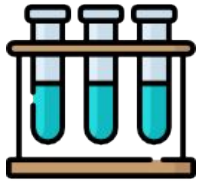






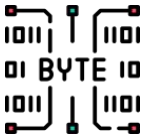




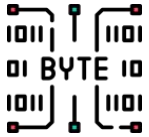


RESEARCH

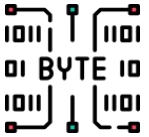
Registry

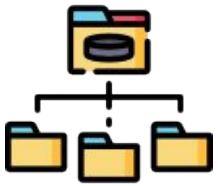


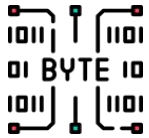
LIMS

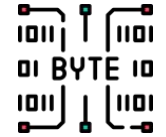


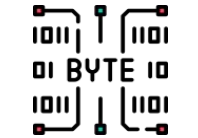
ELN

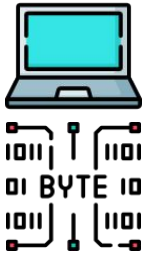






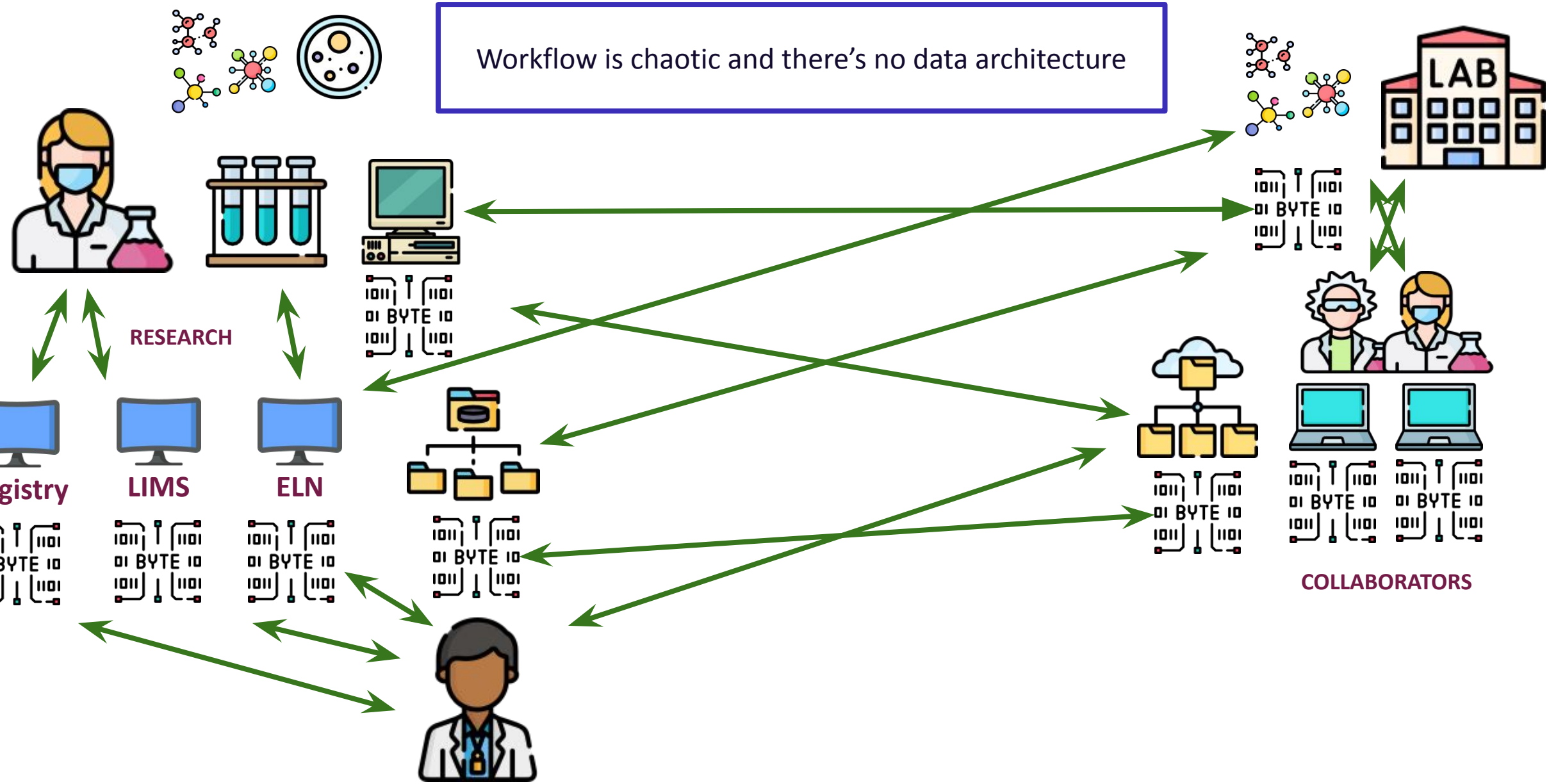






COLLABORATORS

Workflow is chaotic and there's no data architecture



PROBLEMS

No SOP for collaboration

Workflow too complicated

Apps and storage are a major expense but not used consistently throughout the enterprise

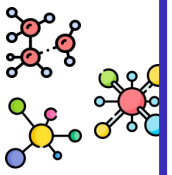
Data is saved in multiple locations with no authoritative reference copy

Scientists email or copy files back & forth

No genealogy or version controls

Access controls hinder data usage while failing to provide robust security

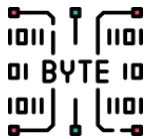
Researchers can’t find important data

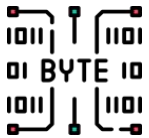


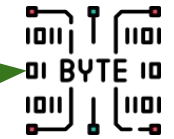
RESEARCH

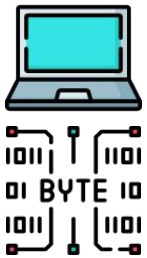
Registry

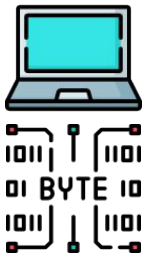
LIMS



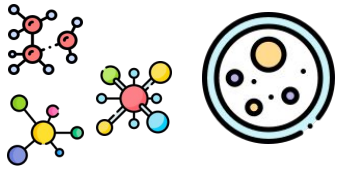


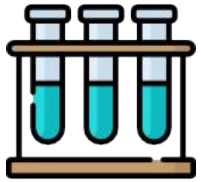






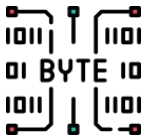
COLLABORATORS

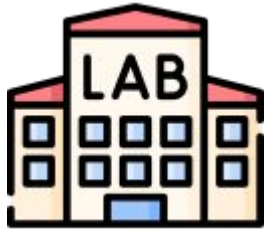


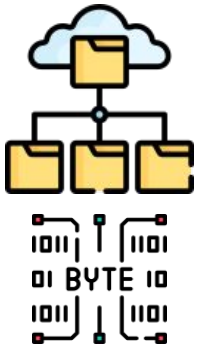


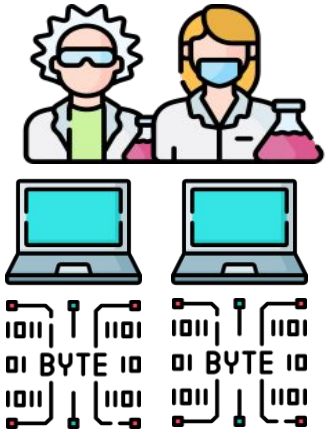
RESEARCH

Here's what you ought to do!



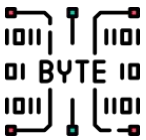




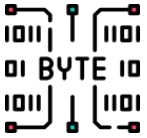


COLLABORATORS

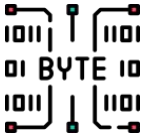
Registry

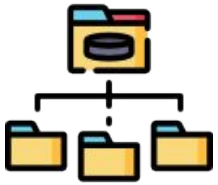


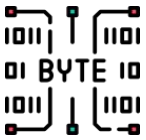
LIMS

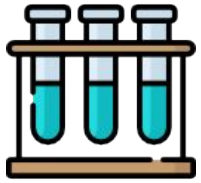


ELN



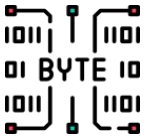




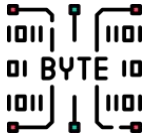


RESEARCH

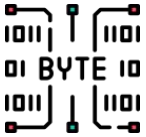
Registry



LIMS



ELN



DATA LAKE in the CLOUD

Throw ALL raw data and files into a bucket

Assign each document a URL

Tag each document with metadata (*source, format, timestamps, versions, content, ownership*)

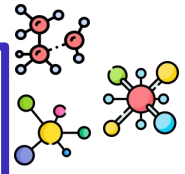
Access Control Lists (*ACL*)

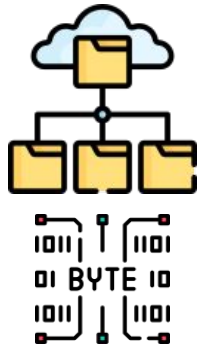
Content catalog

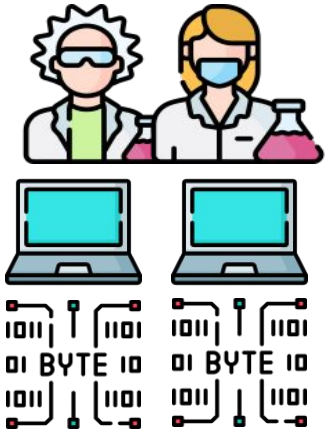
Search by metadata keyword

Derive metadata automatically from content

Work with 3rd party apps (*API*)





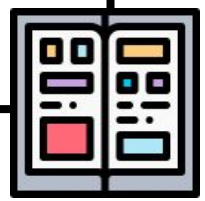


COLLABORATORS


Registry
LIMS
ELN
BYTE

DATA CATALOG

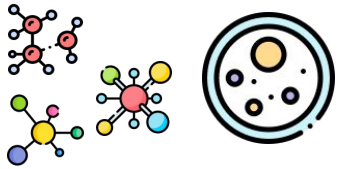
Assay Type
Cell Line
Chemical Compound
File Type
Gene
Lab Instrument
Organism
Protein
Structure
Tissue

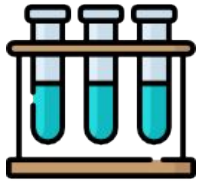


Search by synonym

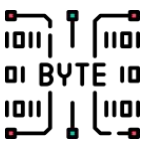
Browse Data Collection

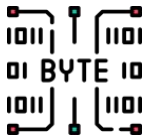
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

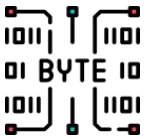


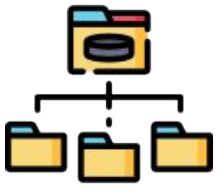


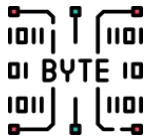
RESEARCH

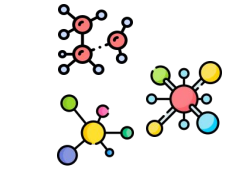


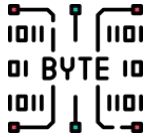


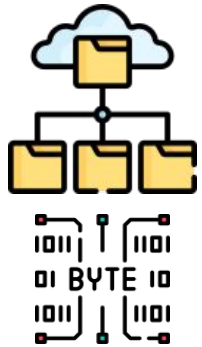


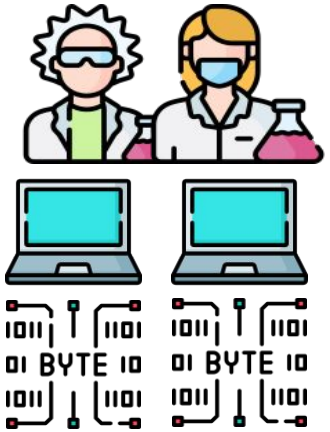






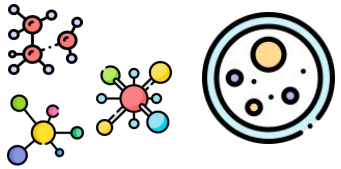


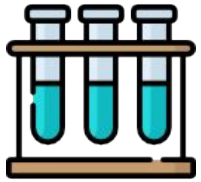




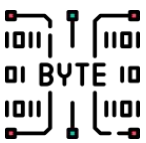
COLLABORATORS

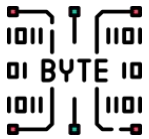
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

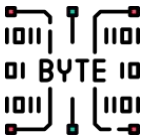


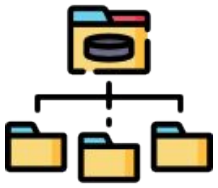


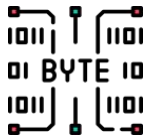
RESEARCH

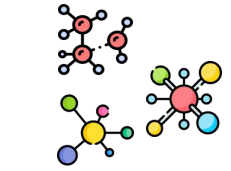


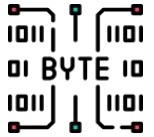


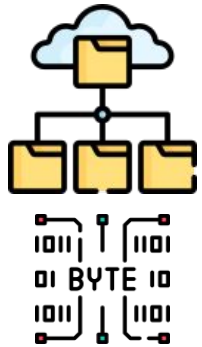


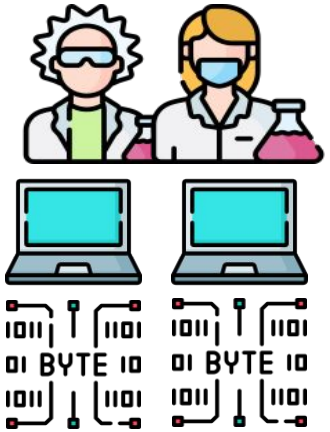




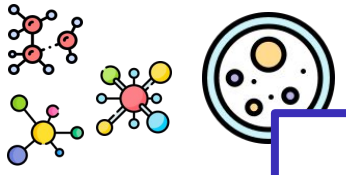


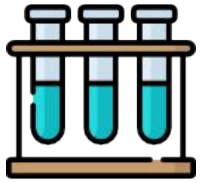






COLLABORATORS



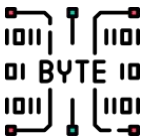
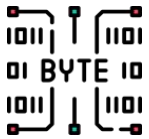
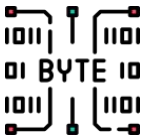


RESEARCH


Registry


LIMS


ELN

TRUTH

(1)

The data process of drug discovery is necessarily complex and expensive

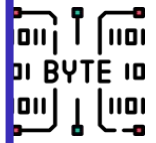
(2)

The system exists primarily to produce documents for regulatory approval

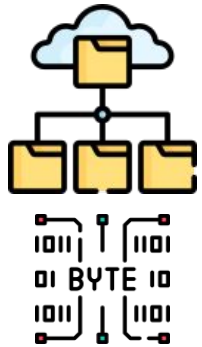
(3)

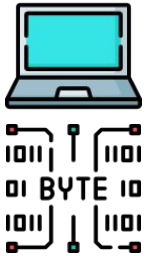
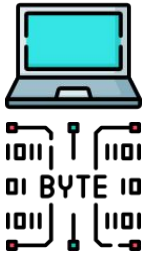
The system is not designed to enable reuse of data



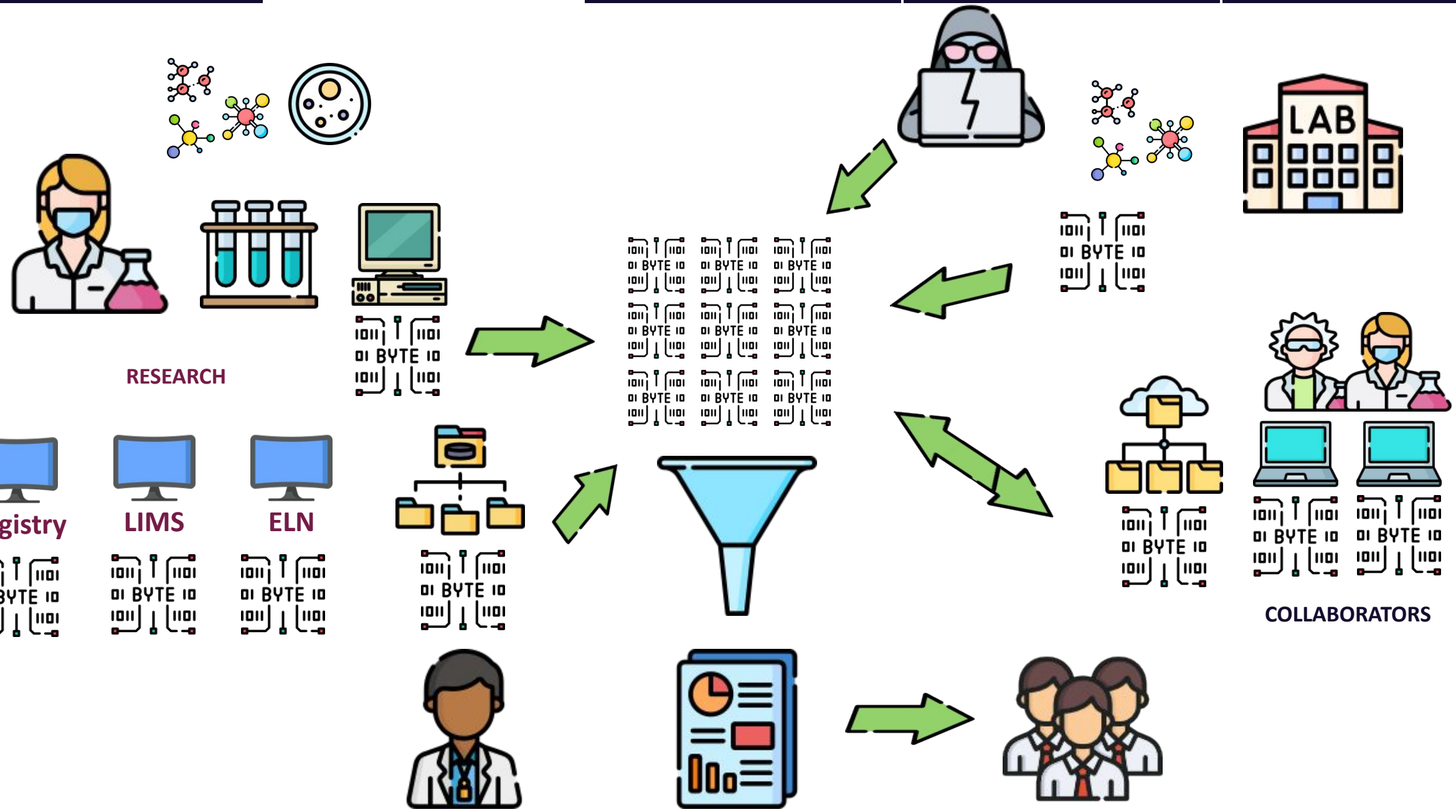
LAB



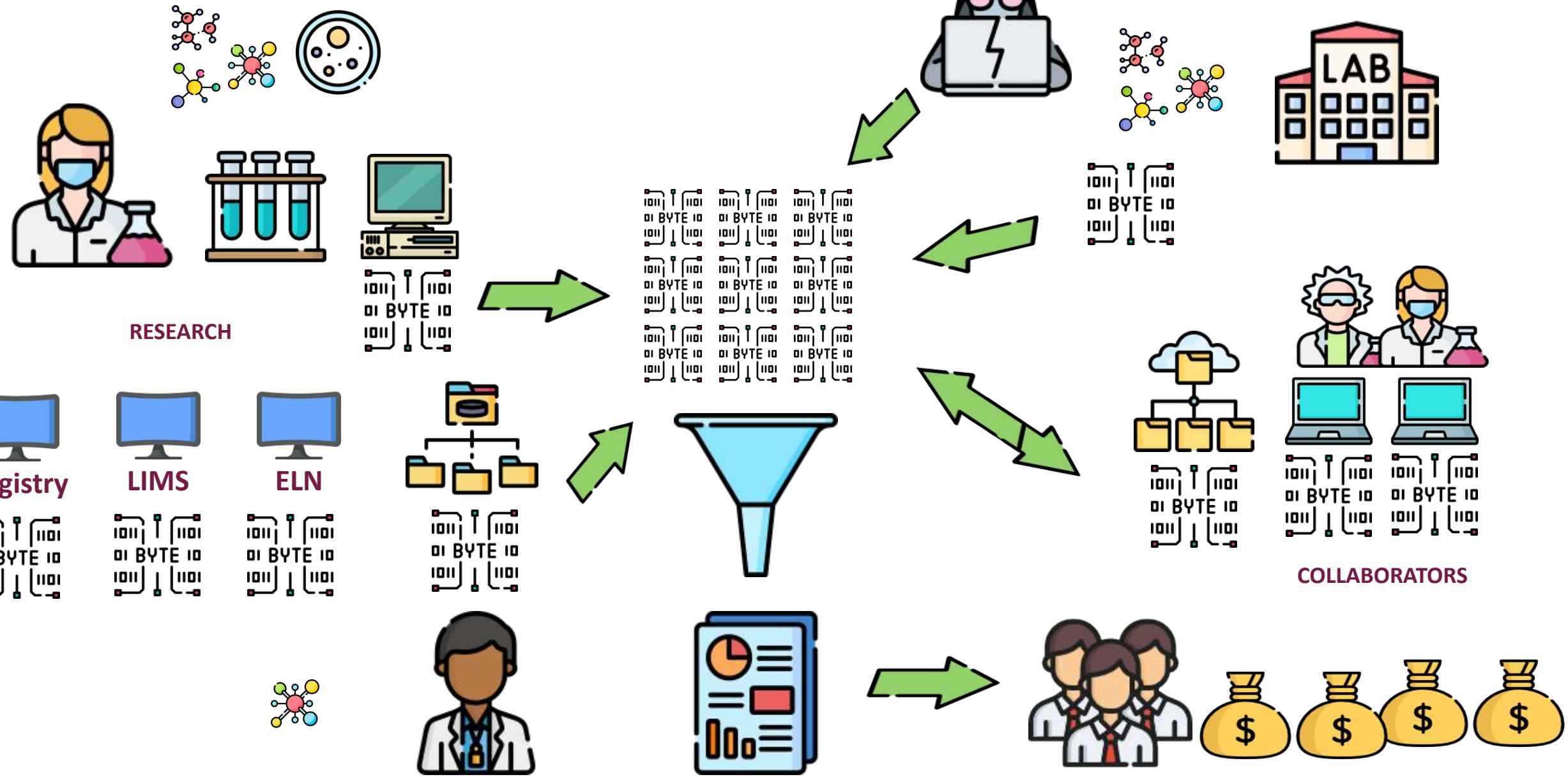




COLLABORATORS

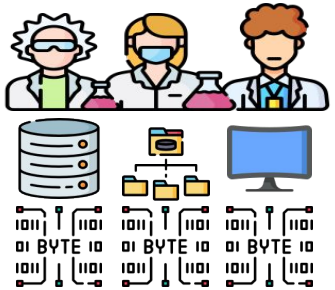
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

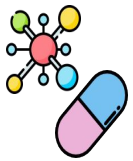


TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET



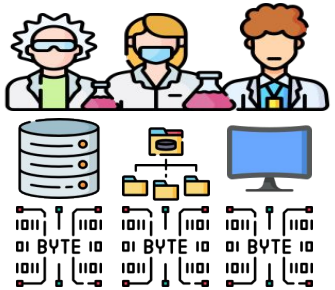
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

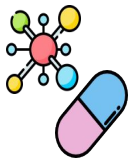




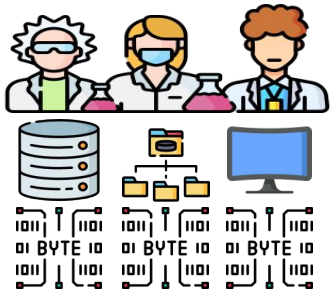
PHARM

TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET



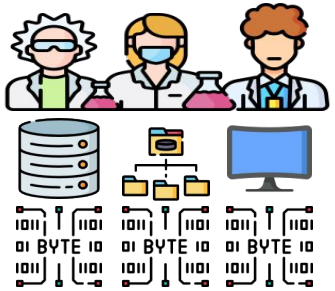


PHARM

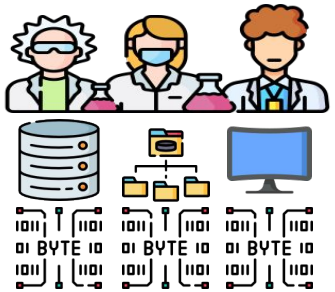


CMC

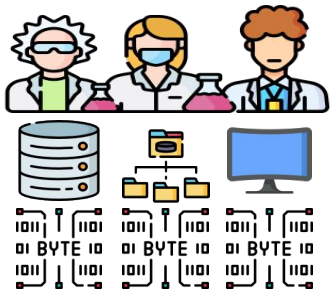
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET



PHARM

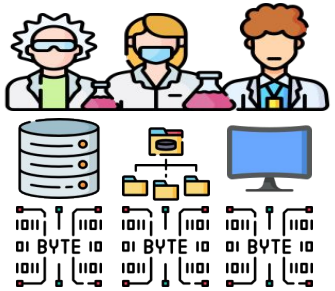


CMC

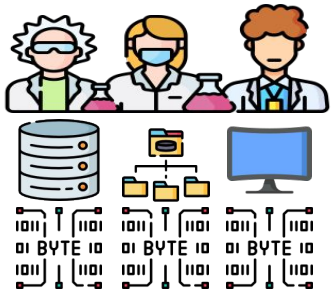


VIVO

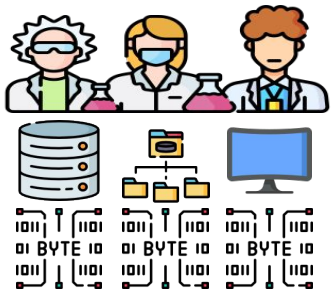
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET



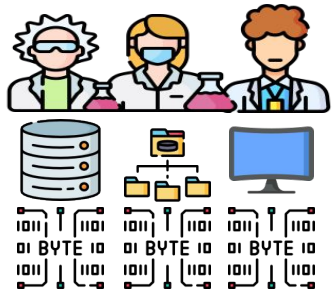
PHARM



CMC

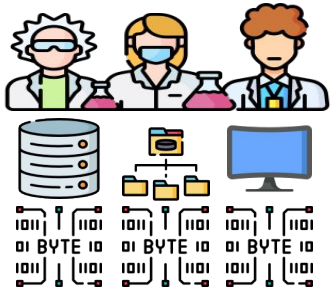


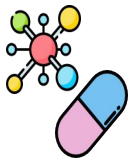
VIVO



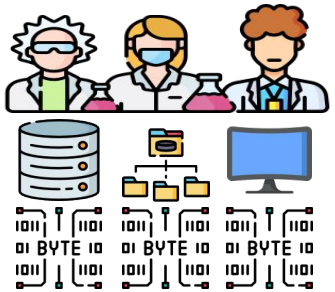
PKDM

TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

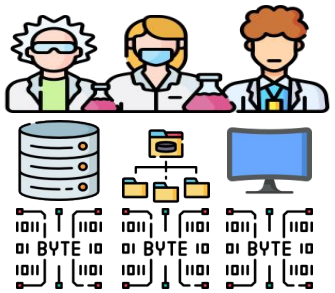




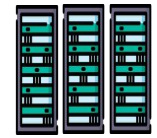
PHARM



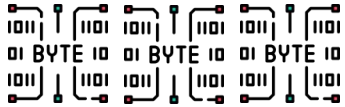
CMC

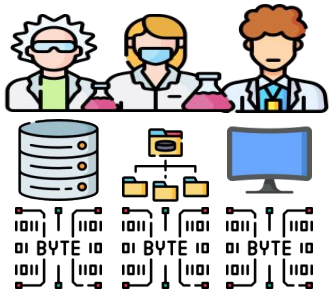


VIVO



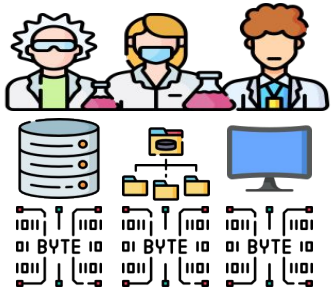
NONMEM

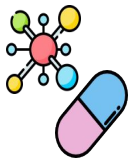




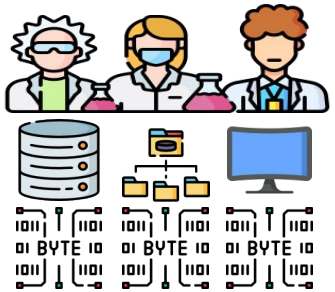
PKDM

TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

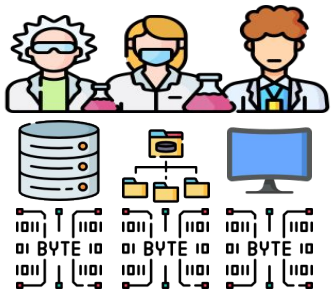




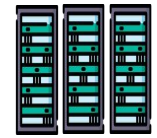
PHARM



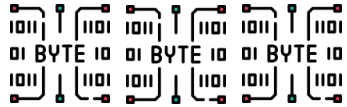
CMC

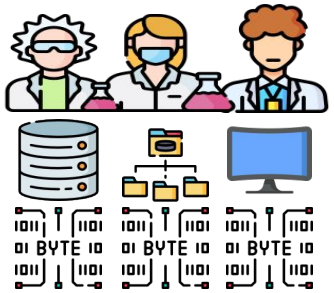


VIVO

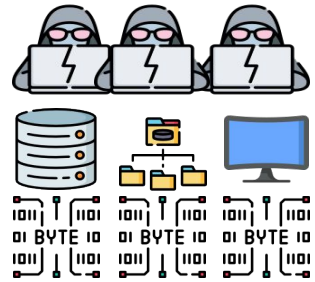


NONMEM



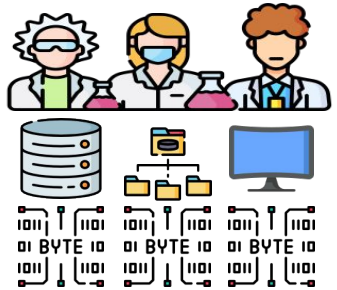


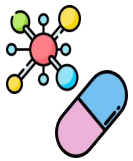
PKDM



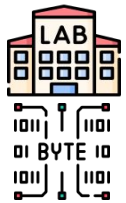
PROGRAMMING

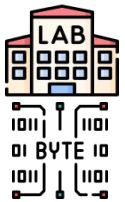
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

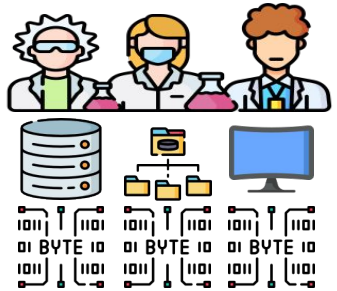




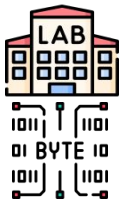
PHARM

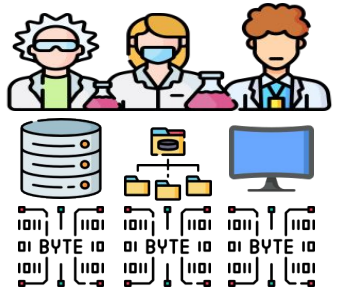




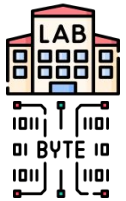


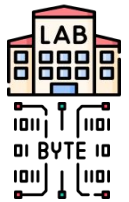
CMC

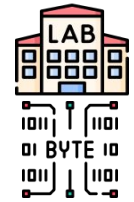


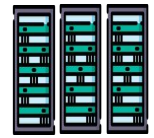


VIVO

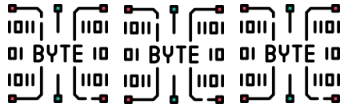


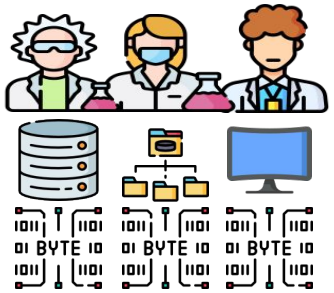




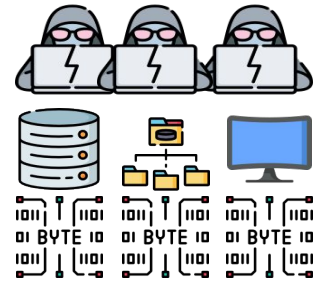


NONMEM

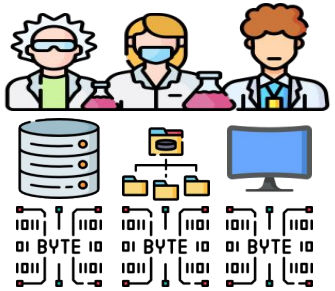


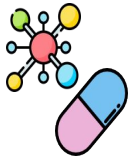


PKDM

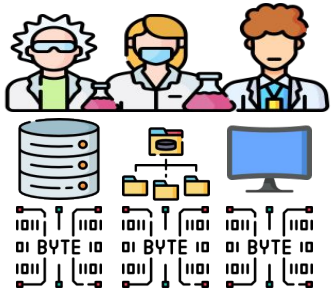


DATA SCIENCE

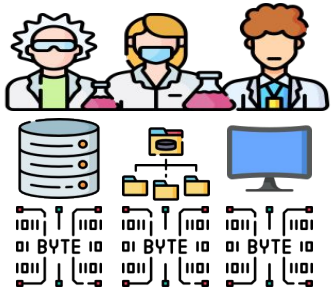




PHARM



CMC



VIVO

PROBLEMS MAGNIFIED

Same issues as before only larger scale

Data siloed by function

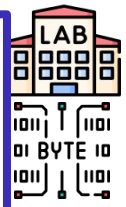
Byzantine governance processes

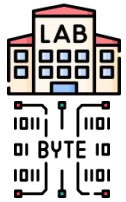
No consistent data standards

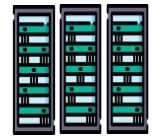
Growing demand for access to working data by functional areas. Response times are too long

The system has grown so big and unwieldly that it's impossible to fix

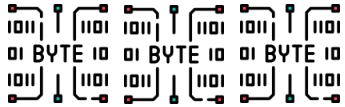
The company spends lots of money to work around deficiencies in the data process

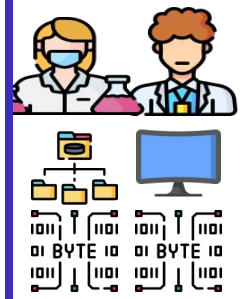




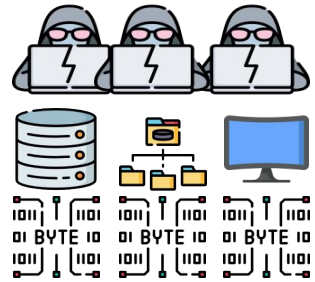


NONMEM

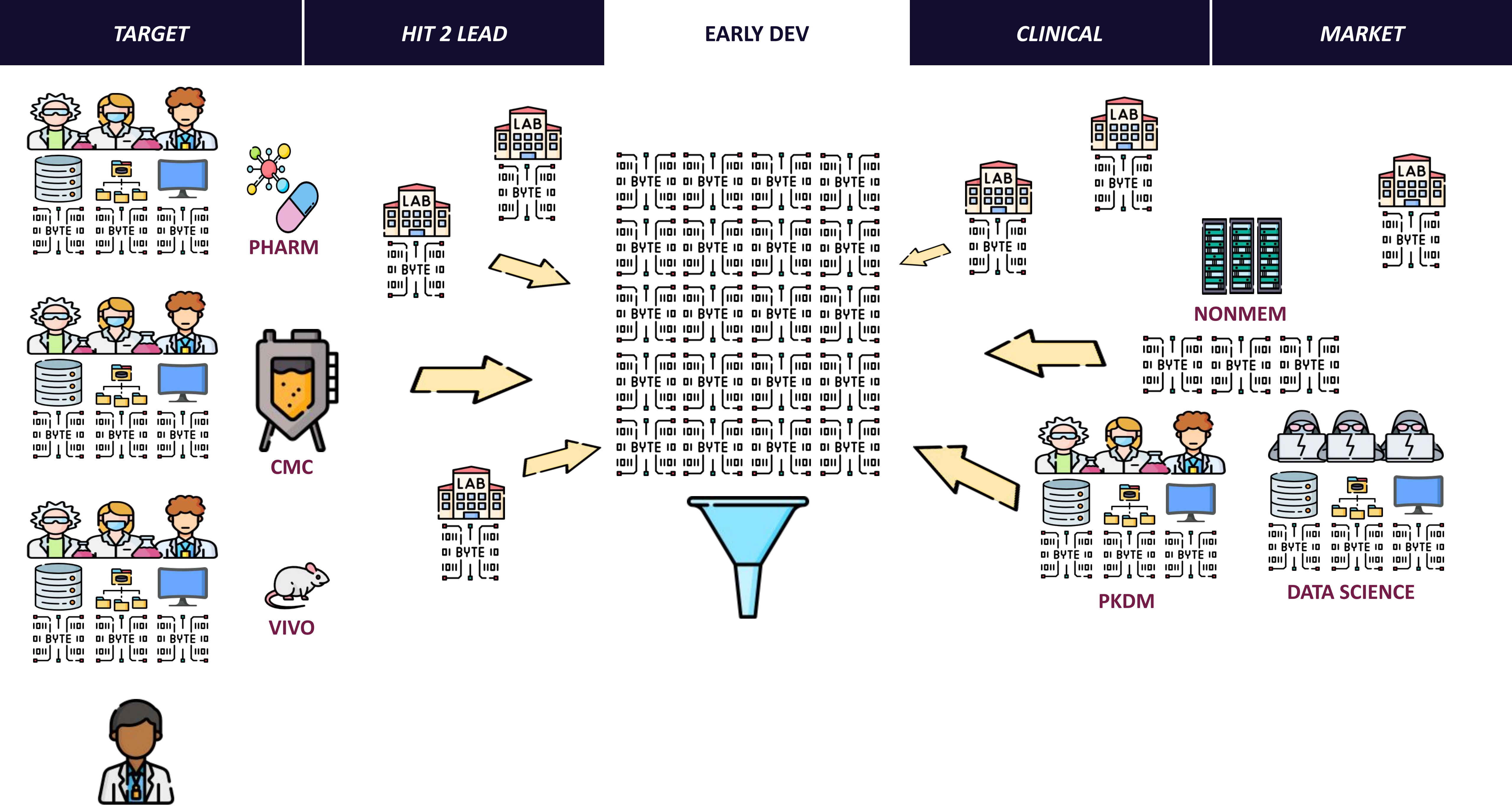




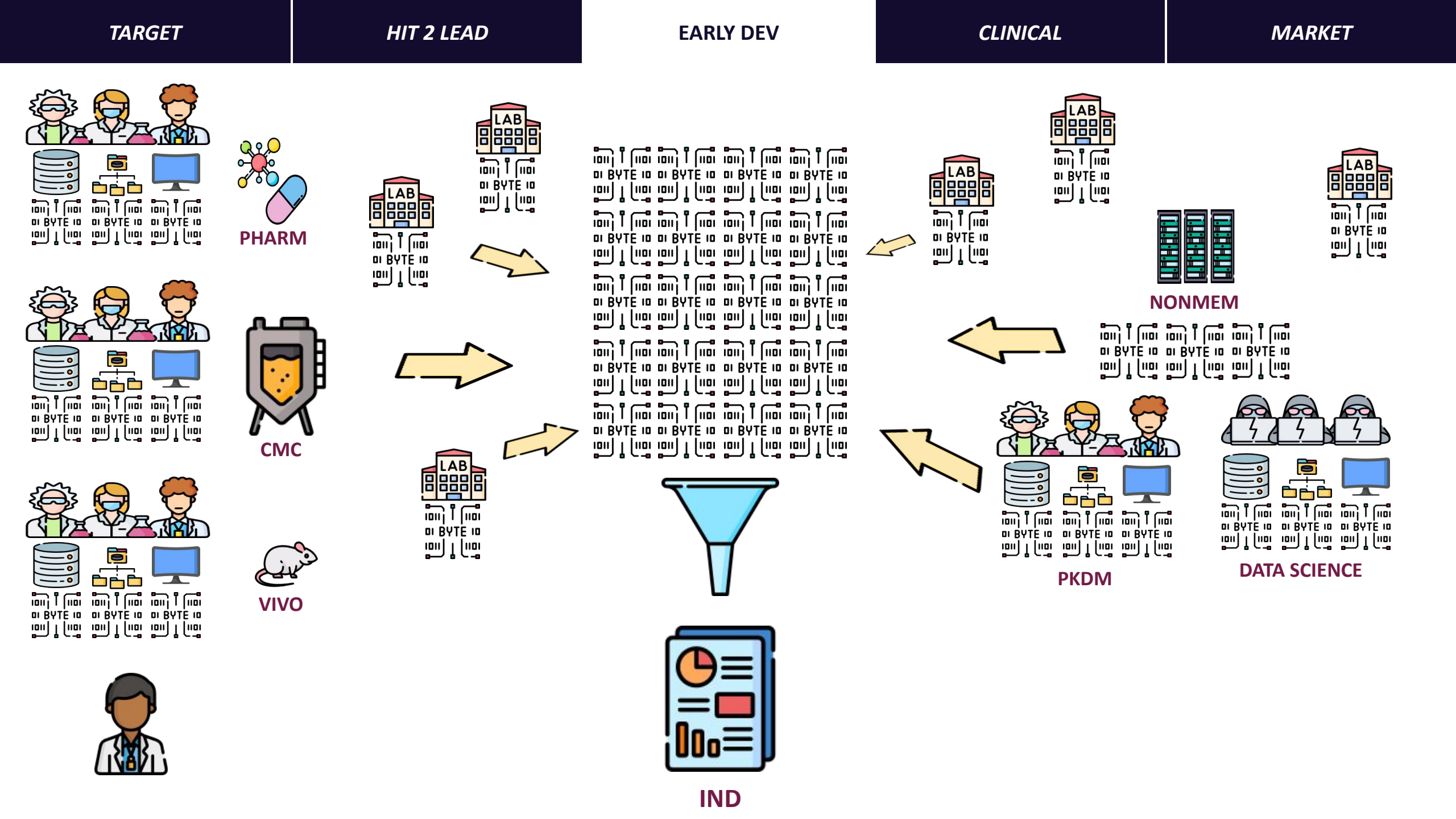
PKDM



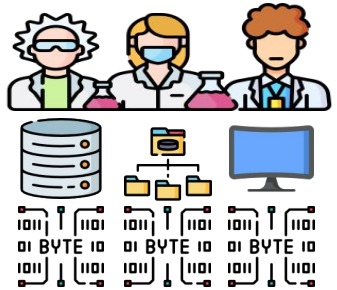
DATA SCIENCE

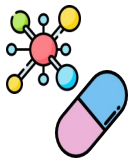


TARGET
HIT 2 LEAD
EARLY DEV
CLINICAL
MARKET
PHARM
CMC
VIVO
NONMEM
PKDM
DATA SCIENCE

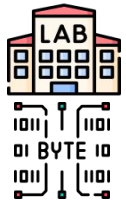

TARGET
HIT 2 LEAD
EARLY DEV
CLINICAL
MARKET
PHARM
CMC
VIVO
LAB
NONMEM
PKDM
DATA SCIENCE
IND

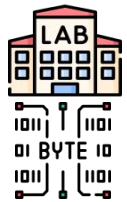
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

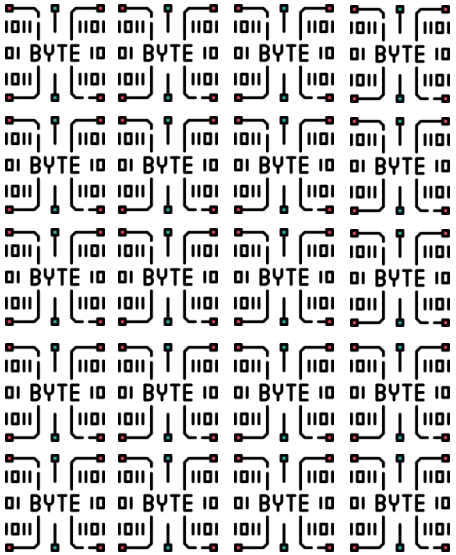


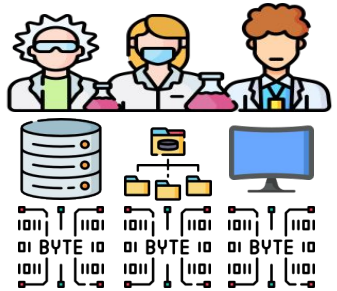


PHARM

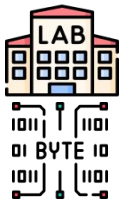


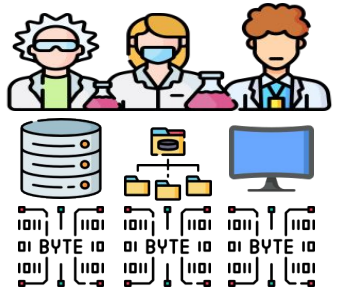






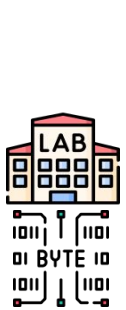
CMC

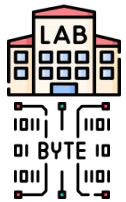


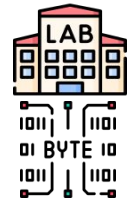


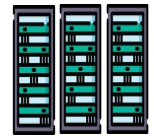
VIVO

IND

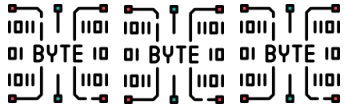


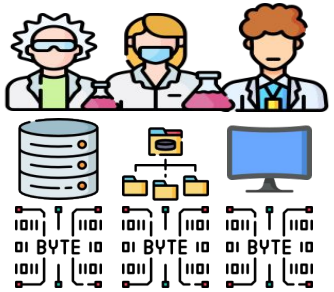


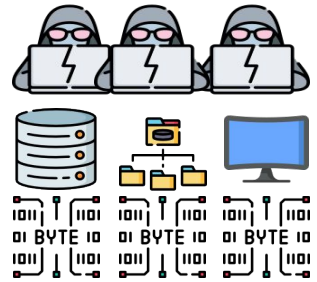




NONMEM



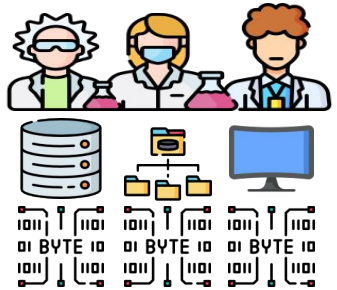


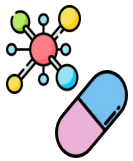


PKDM

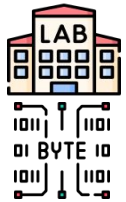
DATA SCIENCE

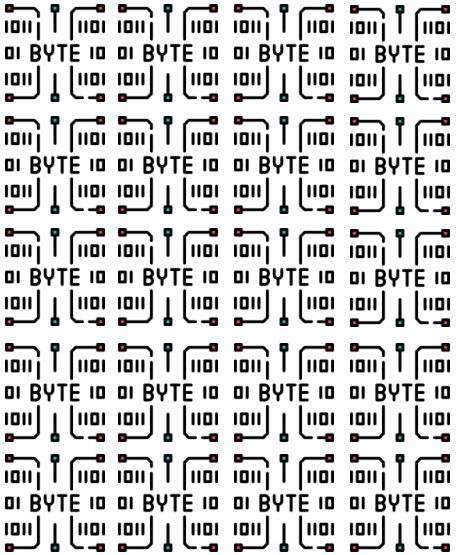
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET

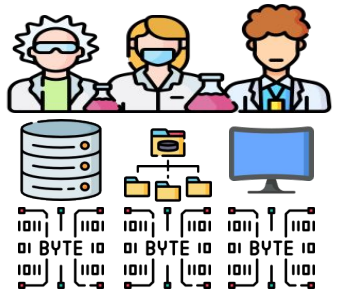




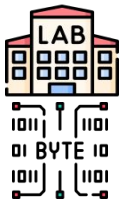
PHARM

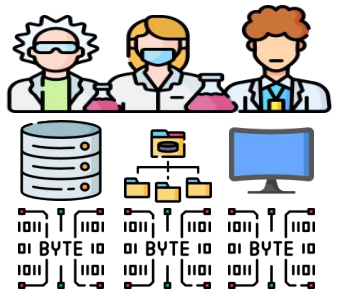




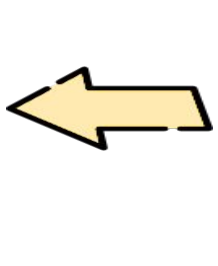


CMC

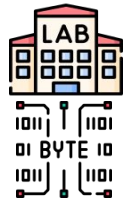


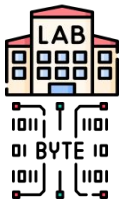


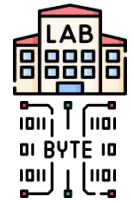
VIVO

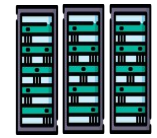


IND

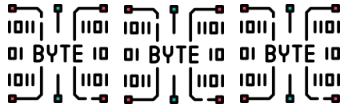


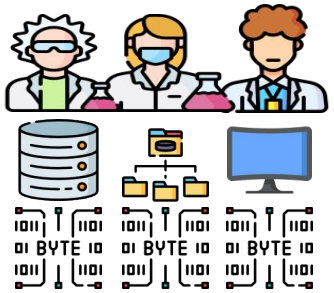




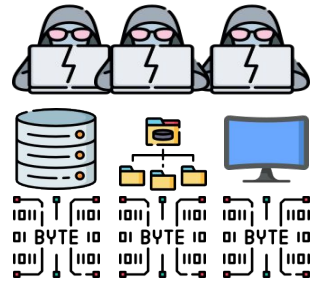


NONMEM



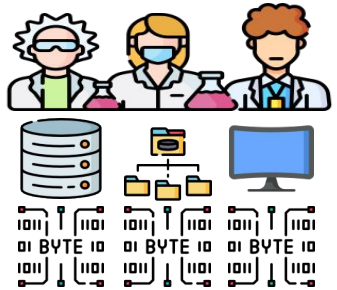


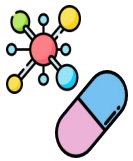
PKDM



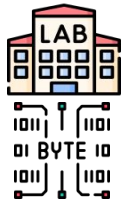
DATA SCIENCE

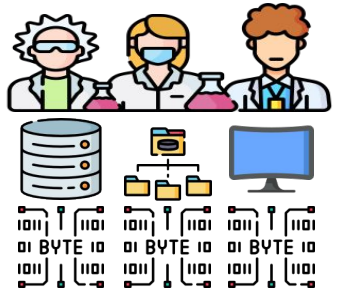
TARGET	HIT 2 LEAD	EARLY DEV	CLINICAL	MARKET



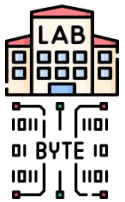


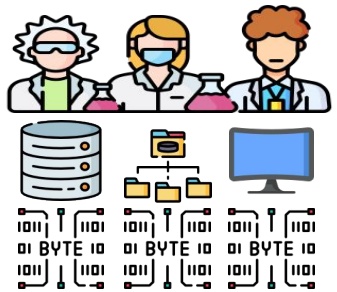
PHARM



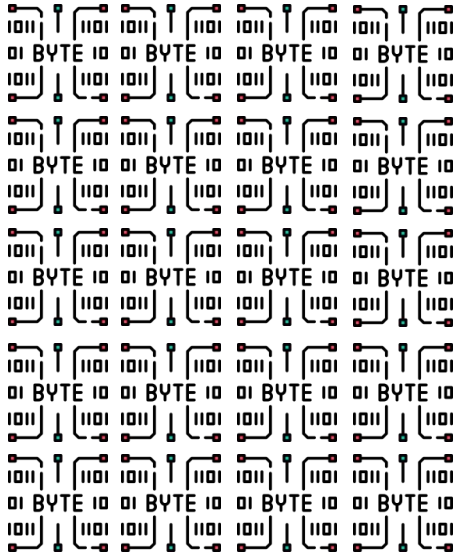


CMC

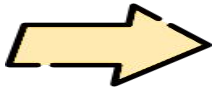


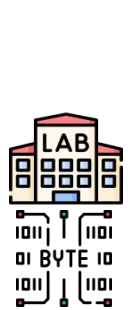


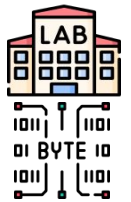
VIVO

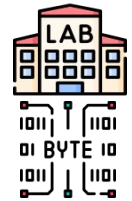


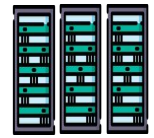
IND



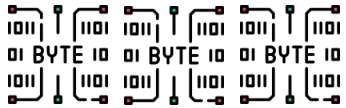


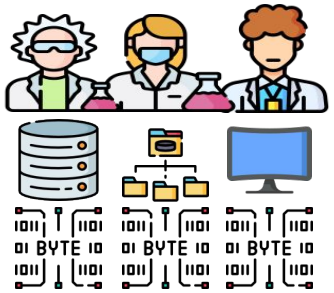




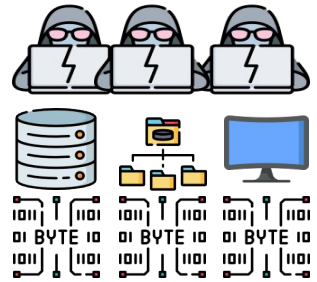


NONMEM





PKDM



DATA SCIENCE

What Happened

- The enterprise invested millions of dollars in applications, storage and data processing but has not appreciated the value of their data

What’s Going to Happen

- Functional areas want to use existing data infrastructure (ELN, LIMS, MES) but cannot obtain licenses, permissions or connectivity
- Scientists want to analyze animal data from all drug programs where a certain buffer is used in the formulation
- Management acquires a related drug program from another company and plans to merge development systems and operations
- An investor wants to finance clinical trials in exchange for a stake in the drug so the company must hand over all related data
- Scientific team needs to defend their research after an academic fails to reproduce work that was published in a journal
- Regulators request detailed information

Recommendations

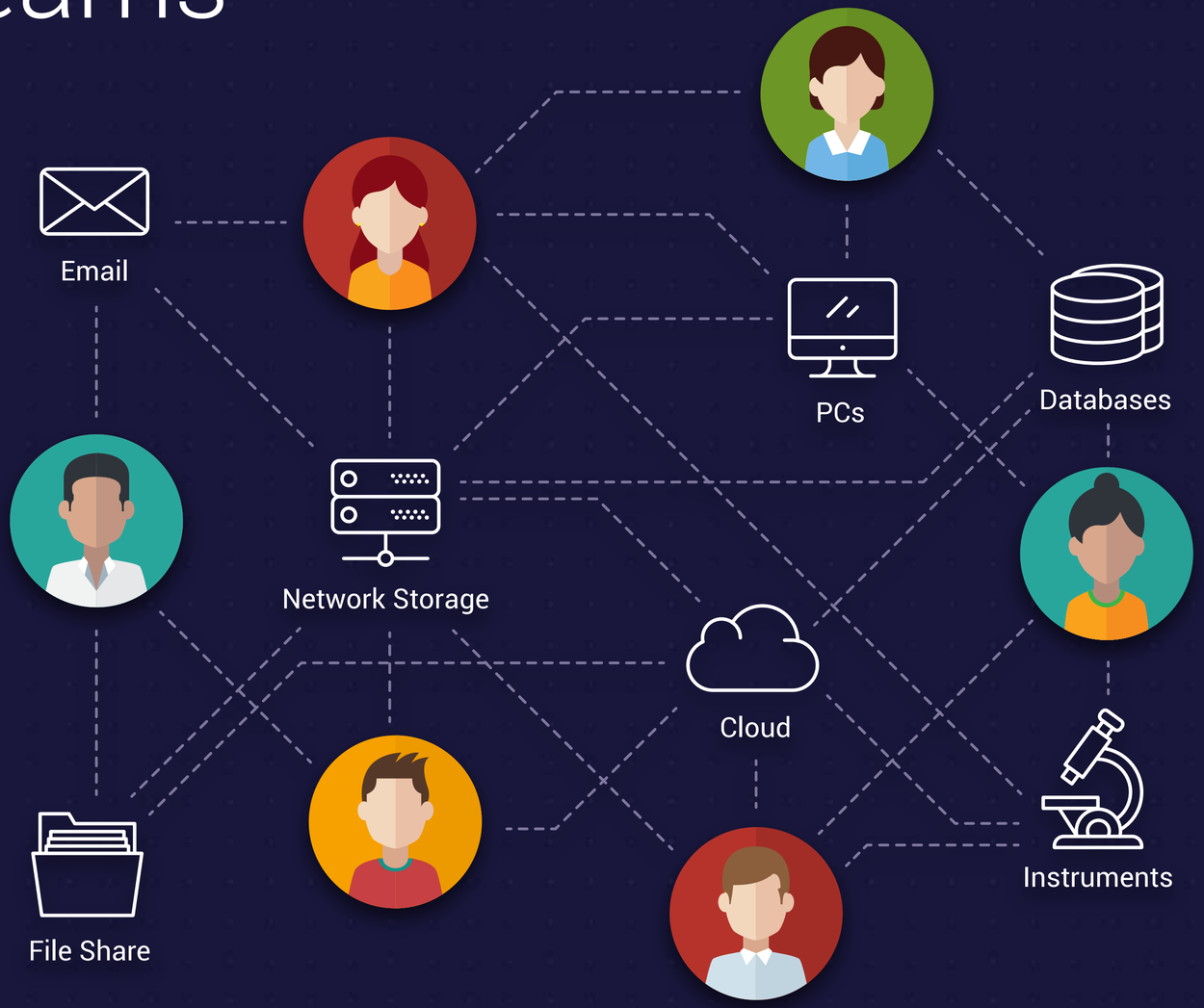
1. Create a scientific data lake in the cloud at the beginning
2. Make the data lake the central repository of reference data for all applications
3. Define a workflow process in which users read source data from the data lake and publish reports and analytics back to the data lake
4. The data lake must generate its own human-readable content catalog with browse and search capabilities
5. Build controls and security for regulatory compliant applications

Creating a Scientific Data Lake

Aneesh Karve

Fragmented data teams

- Different systems, silos, and skill sets
- No data quality lifecycle
- Box is a dead-end for modeling
- Data science can’t reach or model all of the data



Structural barriers to a single source of truth

- Schemas unknown at write time
- Schemas change and evolve
- Query needs discovered long after collection
- Folder structures and file names just don’t scale
- Write-only databases (ELNs, Box)
- Little if any data documentation
- No one system everyone can use

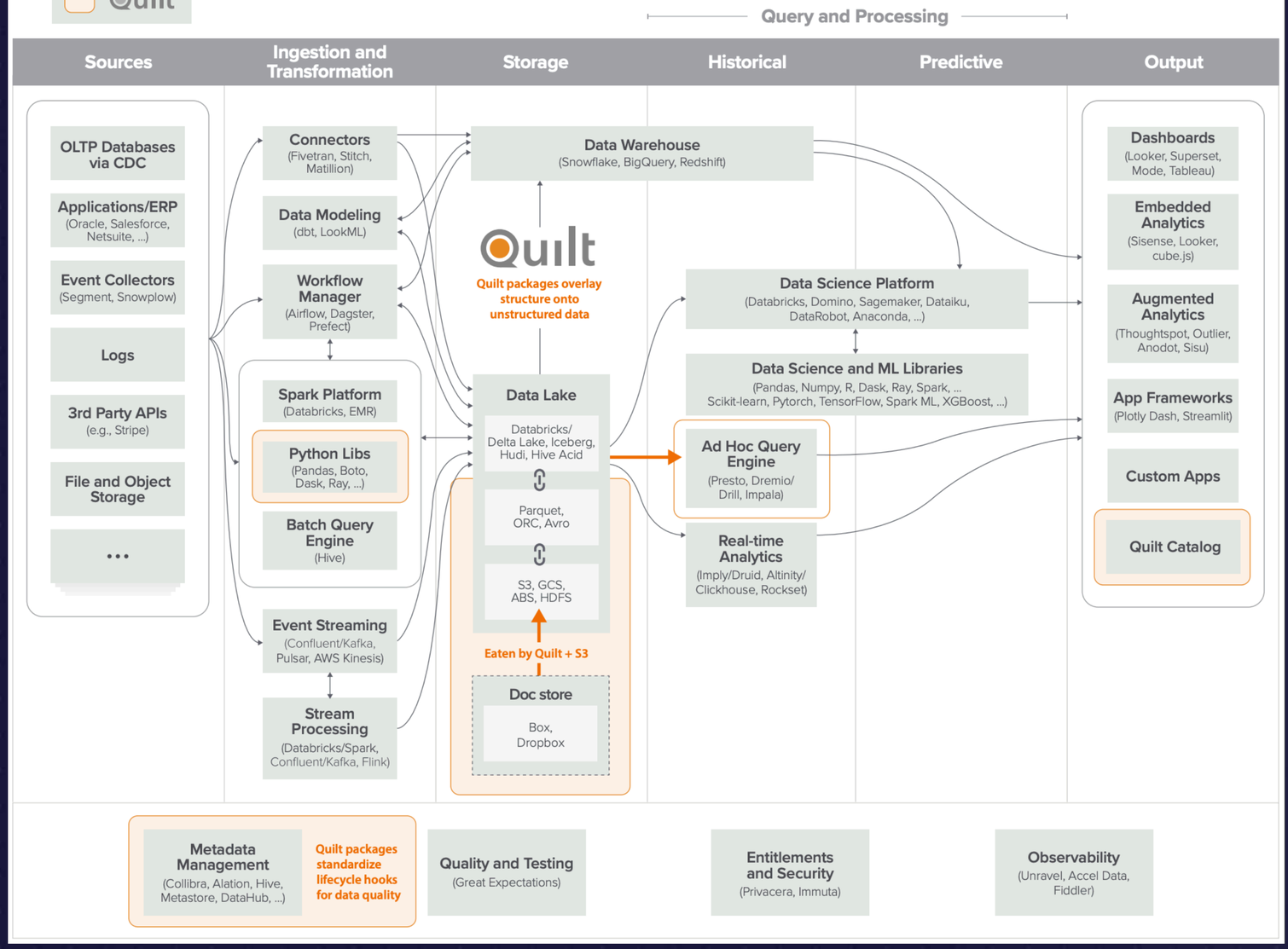
In a perfect-world, a self-organizing data lake

- Schemas discovered, not defined
- Quality gates let you go from swampy to curated and trusted
- Documentation embedded in the data
- Every revision immutable

A Unified Data Infrastructure Architecture

Quilt

Adapted from "Emerging Architectures for Modern Data Infrastructure."
Matt Bornstein, Martin Casado, and Jennifer Li



Principles for self-organizing data lakes

1. There is no schema
2. Always have a schema
3. Use blob storage as the core
4. Define a data lifecycle
5. Gate your lifecycle with quality checks or “workflows”
6. Prefer schema on read databases
7. Embed documentation with data
8. Index metadata for discovery
9. Immutability means reproducibility, means speed of iteration
10. Role-based access over buckets for permissions and compliance

There is no schema

- You won’t know the optimal schema at the outset
- Data sets are multi-modal
- Flexibility makes for speed in the early phases

Always have a schema

- A *manifest* enforces a schema on arbitrary data
- Define and enforce schemas for experiments and assets

“Data sets” transcend schemas and physical layout

- Separate location from intent
- Infinite logical views
- Structures unstructured data
- Immutable

Logical keys	Physical keys	Metadata
`data.parquet`	`s3://bucket/folder/101.parquet`	`(hash, size, intent)`
`parse.json`	`s3://another/folder/paper-81acded.json`	`(hash, size, intent)`
`plates.tiff`	`s3://images/foo/plate_1238.tiff`	`(hash, size, intent)`
…	…	…
ANY LOGICAL VIEW	ANY LOCATION	ANY INTENT

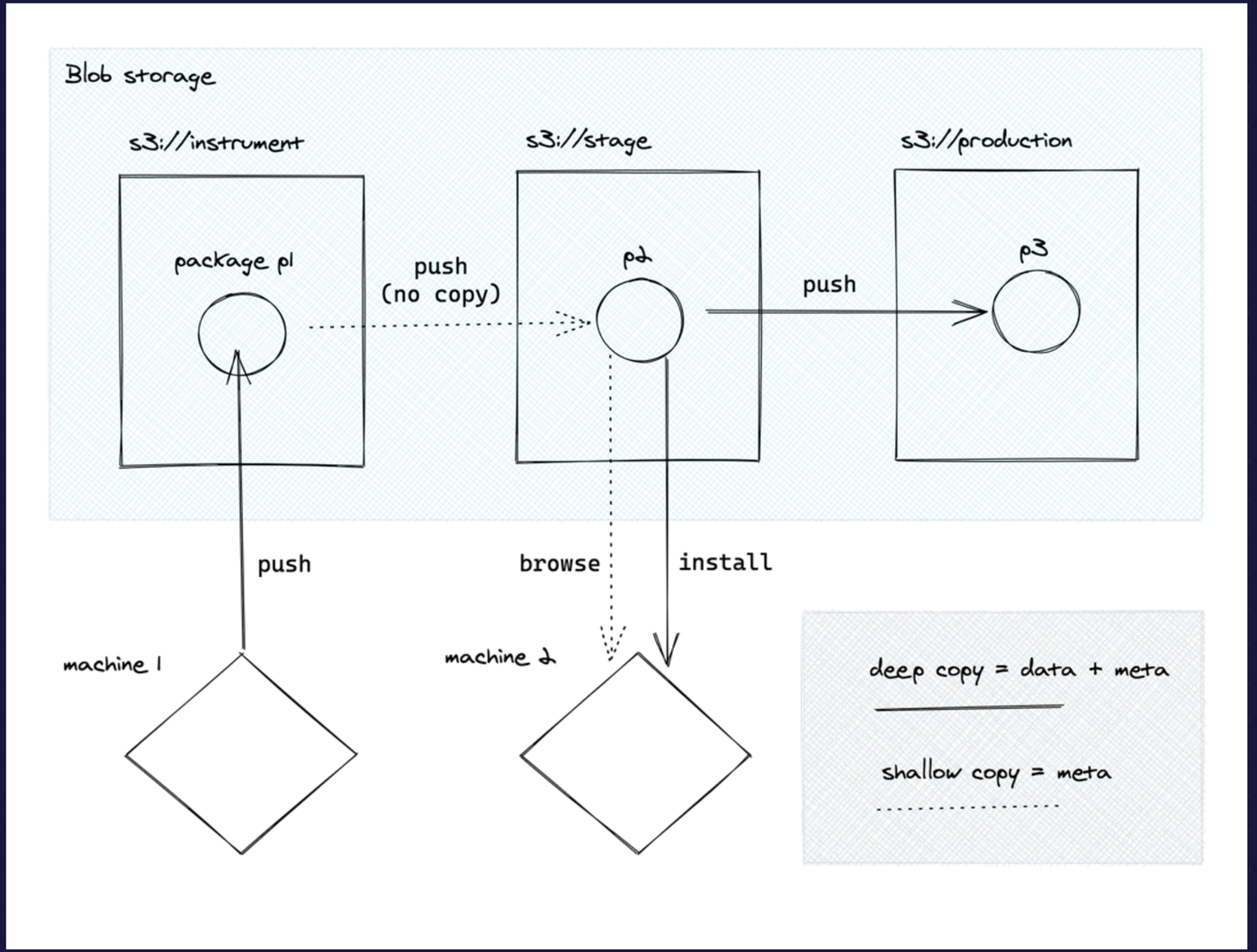
Package experiments into immutable building blocks

- Reproducible
- Discoverable
- Trusted

Blob storage at the core

- Optimal cost:performance trade-off
- No structure? No problem
- Avoid: databases, NASs
- Bring compute to data
- Multi-cloud

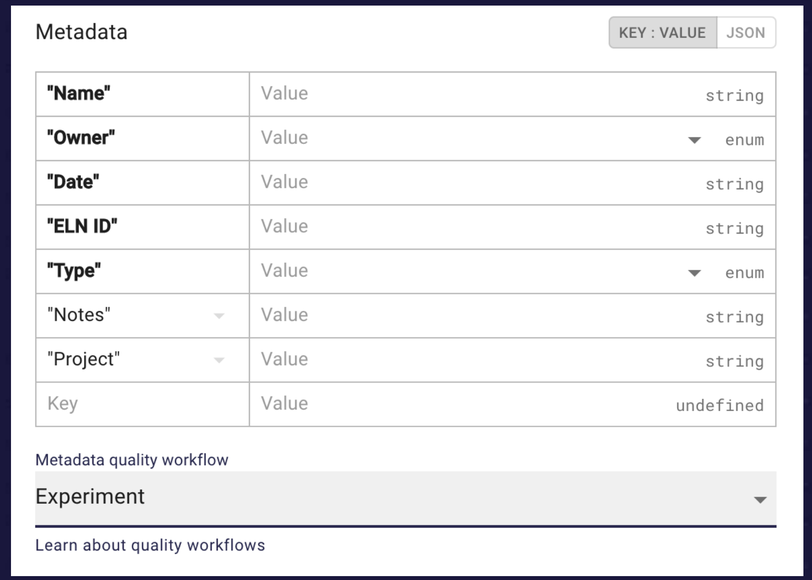
Define a data lifecycle



Gate your lifecycle with quality checks

- Unstructured data: PIL, etc.
- Semi-structure data: JSON Schemas
- Structured data: https://greatexpectations.io/

```
{
  "$id": "https://example.com/geographical-location.schema.json",
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Longitude and Latitude Values",
  "description": "A geographical coordinate.",
  "required": [ "latitude", "longitude" ],
  "type": "object",
  "properties": {
    "latitude": {
      "type": "number",
      "minimum": -90,
      "maximum": 90
    },
    "longitude": {
      "type": "number",
      "minimum": -180,
      "maximum": 180
    }
  }
}
```



For analytics, prefer schema-on-read

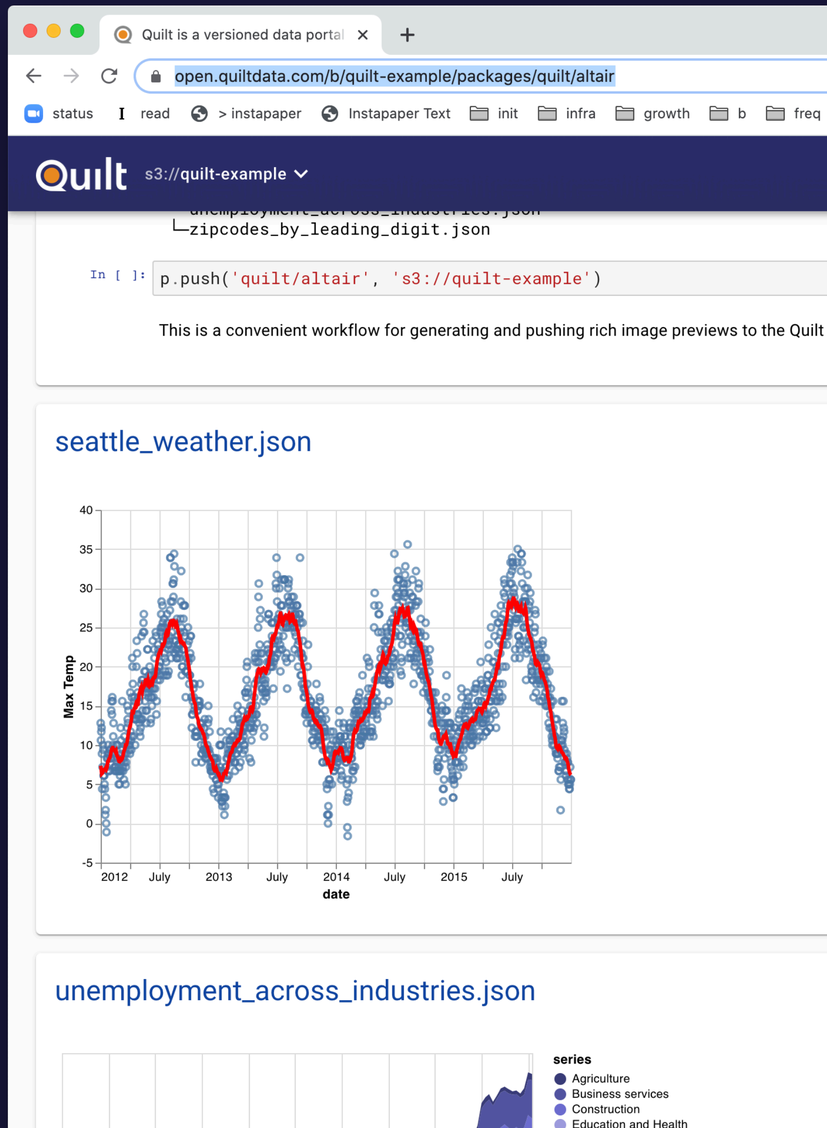
- Presto DB is tolerant of missing values, missing columns
- Can drop columns, change column names
- Read directly from S3

```
SELECT * FROM manifests_allencell
WHERE substr(logical_key, -5)='.tiff'
AND json_extract_scalar(meta, '$.user_meta.nucmembsegmentationalgorithmversion') LIKE
'1.3%'

AND json_array_contains(json_extract(meta, '$.user_meta.cellindex'), '5')
```

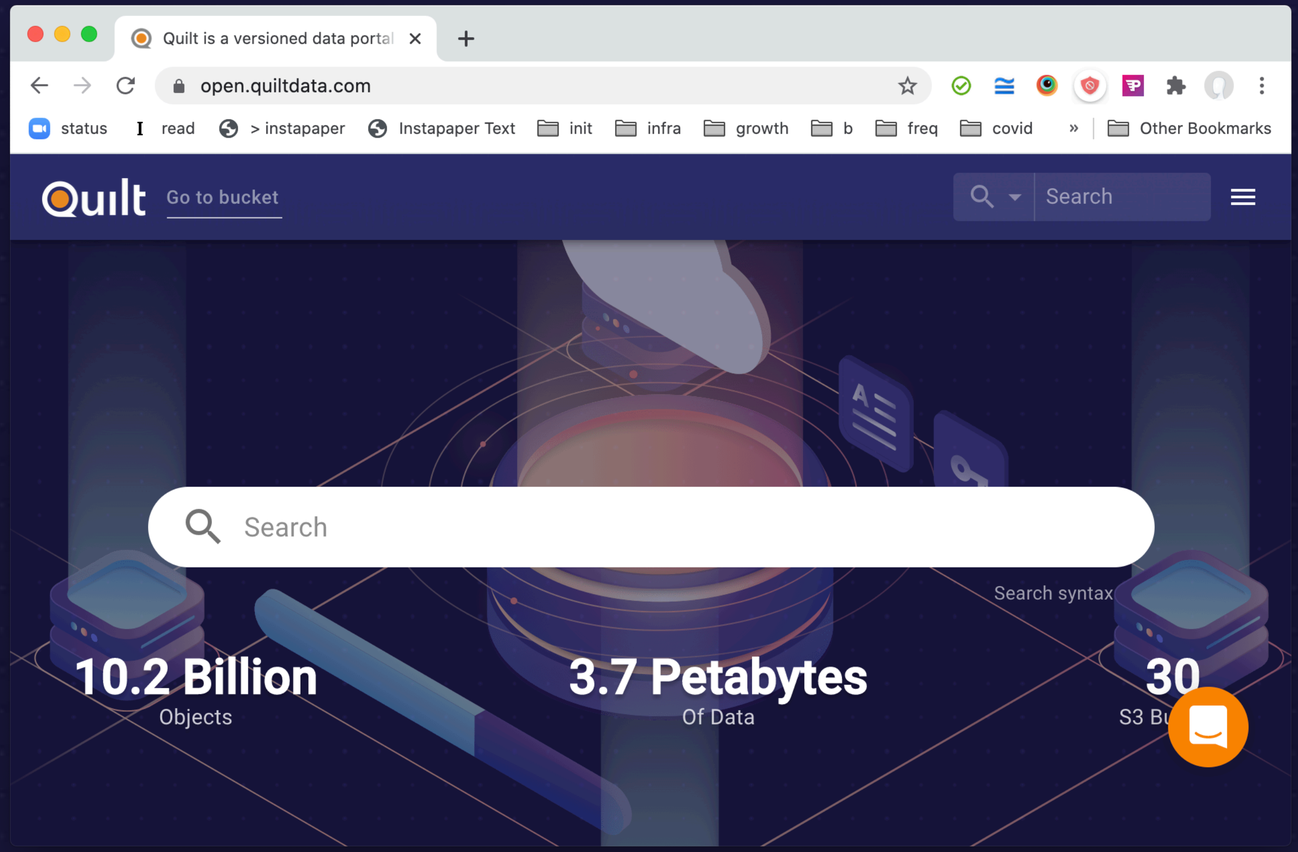
Embed documentation with data

- Data without context quickly becomes meaningless
- Example of “live docs” in S3



Index metadata for discovery

- ElasticSearch for fuzzy discovery
- Avoid fixed object schemas, prefer strings
- Control costs by indexing reasonable subset of data, key file types only, UltraWarm
- https://open.quiltdata.com/



Immutability means speed, good sleep

- Reproducibility across time, machines, and collaborators means
  - You are audit-ready
  - You can build on the work of other team members
  - Debugging is quicker
- Reproducibility means quicker, more correct results, with smaller teams
- Enable object versioning
- Versioning data and models for rapid experimentation in machine learning

model := script(code, environment, data)

Role-based access, auditing

- Break projects out as self-contained, regional buckets (times three)
- Use IAM roles to enforce access
- Bring compute to data
- Use a HIPAA-eligible data store, enable encryption at rest
- Avoid bucket policies
- Enable CloudTrail to audit every access (trail in separate account)

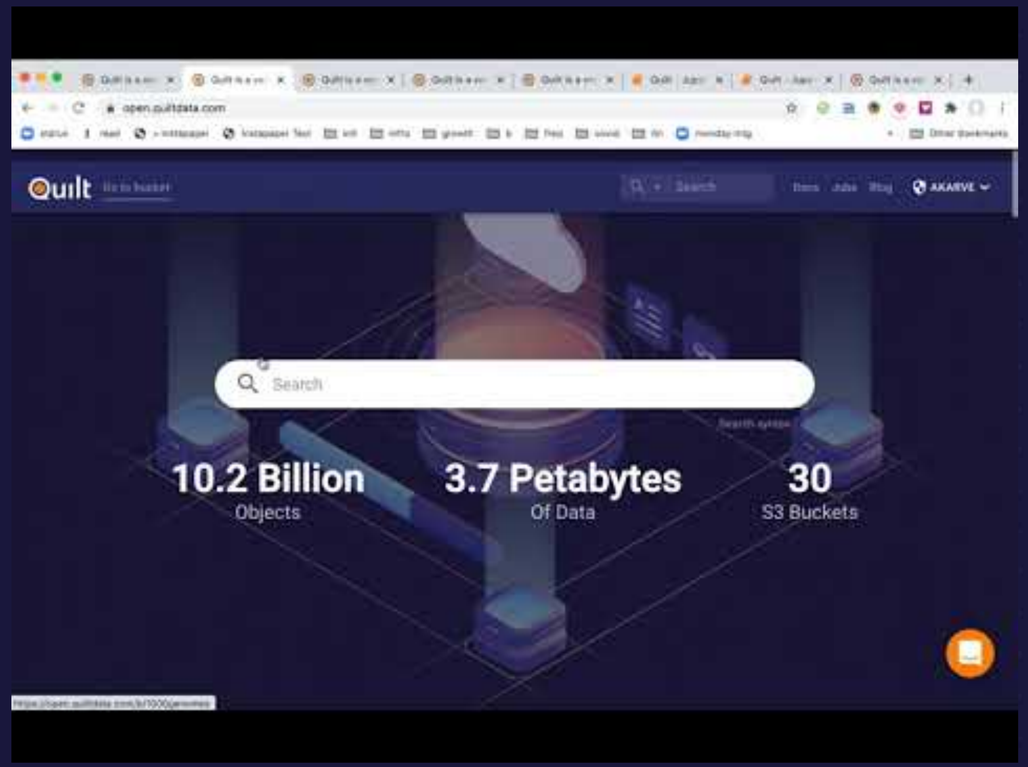
Customer ROI (return on investment)

- Gives your team a 3-month head start on our data lake and data management initiatives
- Regularly frees up one day a week for data architects and data engineers
- A team of 4 can do the work of 5, without a data engineer
- Every Quilt search that prevents a repeated experiment saves $1K to $20K USD in reagent and container costs
- Model velocity and correctness are higher with Quilt

Learn more

- Quilt overview video (4 min.)
- Versioning data and models for rapid experimentation in machine learning
- Principles of lazy data documentation — and how to get your team onboard
- https://github.com/quiltdata/quilt

- aneesh@quiltdata.io
- quiltdata.com



Quilt accelerates data-driven teams

- Prevents thousands of dollars in waste and repetition
- Mature data sets from swampy collections to trusted packages
- Discover models faster

Quilt is a self-organizing data lake

- Flexibility of data lake plus trust of data warehouse
- Search, query, visualize, verify
- Usable by all personas, technical and non-technical
- Usher data from swampy collections to trusted data sets

AWS showcases Quilt

We were very frustrated by these enormous datasets that collaborators couldn't quite seem to get at easily. Using Quilt, we've been able to solve that challenge.

Rick Horwitz, Executive Director, Allen Institute for Cell Science

Read More

aws partner network

Advanced
Technology
Partner

What makes Quilt unique

	Petabyte scale	Versioning	Heterogeneous datasets	Analysis tools	Data Types	Usability	Private cloud
Quilt	Yes	Yes	Yes	Athena, EMR, K8s, all of AWS compute	Structured Semistructured Unstructured	Entire company	Yes
Snowflake	Yes	Table	No	SQL	Structured Semistructured	Developers	No
Databricks Delta Lake	Yes	Multi-table	Yes	SQL, Spark	Structured	Developers	Yes
AWS Lake Formation	Yes	No	No	Glue-compatible compute	Structured Semistructured	Developers	Yes
Box	No	Single file	Yes	None	Semistructured Unstructured	Non-technical users	No

Security

- Quilt runs in your Virtual Private Cloud
- All data lives in your S3 buckets
- All compute runs on your infrastructure
- Configurable for VPN-only access
- Supports SSO via Okta, Google, OneLogin

DRAFT

Supplement: User stories for bench science

1. As a bench scientist, I would like to store and annotate experiments, so that my colleagues can build upon my results in the future
2. As a bench scientist, I would like for my experiment annotations to be checked for quality against a known schema, so that our team can trust its results
3. As a bench scientist, I would like to find all experiments that match a natural language query, so that I can more quickly plan future experiment iterations and unblock collaborators
4. As a bench scientist, I would like to automate away the process of creating decks and summaries, so that I can focus on running new experiments and new hypotheses

DRAFT

Supplement: User stories for data science

1. As a data scientist, I would like to make simple changes to my scripts that take advantage of Quilt packages, so that we can gain the advantages of Quilt with minimal effort
2. As a data scientist, I would like to use AWS power tools to search and query across all of Inscripta's experiments, so that I can better mine and model our experiments
3. As a data scientist, I would like to version all experiments and outputs, so that I can retrace the steps of our colleagues with high confidence
4. As a data scientist, I would like to ensure that the metadata for every new experiment is screened for data quality, so that the final analysis is highly trusted