Structuring Data for ML/AI Success in Biotech

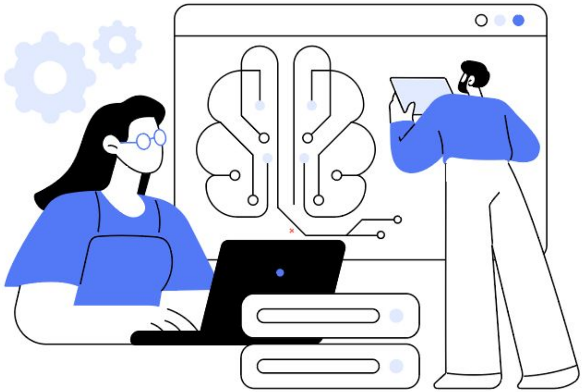
Bernard Lee, Product Manager

LabKey®
PARTNERS IN SCIENCE

Introduction

Objective:

To equip you with knowledge to prepare and perform data structuring processes for effective ML/AI application in biotech, emphasizing the critical role of data quality and organization



About LabKey

More than just another software vendor.

- Started in 2003, we offer a range of scalable solutions to academia, pharma/biotech and government research.
- Managing scientific data is our expertise. Our clients benefit from our decades of combined experience.
- We include support from our Account Managers to ensure your migration is successful and your goals are achieved.
- Your feedback and guidance are critical in shaping the solutions and the features we deliver.

Introduction

Why does AI/ML matter for biotechs?

- Cost, time, and ease of developing a new biologic application are high
- R&D failure rates are high
- Data science products can streamline and accelerate operations and learning

Introduction

Use of ML/AI in Biotech:

- Classifying and predicting protein structures, and performing molecular design using artificial neural networks
- Predicting the location of protein-encoding genes from sequence databases and the DNA sequence
- Microarrays expression pattern identification through classification and irrelevant data reduction
- Predicting binding site identification from biological component data
- Phylogenetic tree reconstruction to include the genomic comparison

Each involve different data but the same principles will drive success.

Introduction

What constitutes success for data science products?

- Scientists spending most of their time using their specialized domain knowledge
- Having clear metrics demonstrating time savings from some modeling product

Key Challenges in Biotech Data

Unstructured, inconsistent, and not interoperable data:

- Structured data is anything that conforms to a set of rules and conventions
- Consistency in data gathering, processing and storing
- Interoperability is ability of software to exchange and make use of information

Structure + Consistency + Interoperability

Key Challenges in Biotech Data

Unstructured, inconsistent, and not interoperable data:

Observations of scientists and their habits:

- Scientists will accomplish what they need with their data, falling back on spreadsheet operations
- Different teams, projects, and people will structure differently what is essentially the same data
- Storing data how and where it's convenient

Key Challenges in Biotech Data

Volume, Diversity and Complexity of Data:

- *Volume* - The right amount of the right data
  - Quantity of data impacts your ability manage, clean, process, and store it
  - Many data science products require much more data than you likely have

- *Diversity* - How many types of data do you collect? Is it data diversity or is it inconsistency?

- *Complexity* - Data that requires heavy processing, transformation or conversion to be structured and integrated with other sources.

Key Challenges in Biotech Data

Data Type Integration:

Integrating results with sample data, experiment conditions and other contextual metadata creates the landscape to explore scientific questions.

- Often curated manually
- Needed for analysis and reporting when dealing with multiple sources and formats.
- Needed for any data products like ML/AI

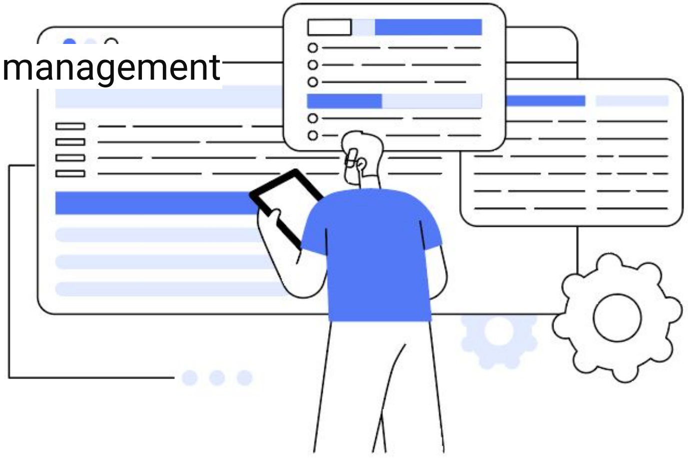
Example:

- Bioreactors - transactional data combined with drawn samples to be associated with assays characterizing them. Must rectify/align identifiers and time stamps

Key Challenges in Biotech Data

Institutional Support and Structure:

- Most scientists are not fundamentally prepared to surmount these challenges
- Effective management of the data lifecycle requires a suitable infrastructure
- Data engineering groups are expensive, require data management tools, and impact scientist



Fundamentals & Best Practices

Formulate the problem:

What knowledge and tools would advance your institutional aims?

Domain Expertise:

Incorporate domain knowledge in data structuring, automation, and data science products

- Selection of relevant features and understanding potential biases in the data collection process
- Don’t lose sight of the actual context. Include your scientists to make sure what you get is meaningful.

Fundamentals & Best Practices

Consider the whole Data Lifecycle:

- How will your data be captured, transformed, processed, analyzed, and stored to end up with useable data?

- Clearly identify what will happen at each step/stage then select the right tools so you don't fall back on variable processes

- Emphasis on supporting data analysis:
  - Transformations
  - Annotations
  - Calculations
  - Visualizations

Fundamentals & Best Practices

Structure and standardize your data: Data quality

- Adopting common data models
- Appropriate and consistent data types - 4.03 or $4.03, or even 4 dollars 3 cents
- Consistent naming
- Unique IDs for entities
- Avoid free-form text when possible - conform to standard, known values
- Validation rules

Integrate & Align Your Data:

- Convert all data to appropriate data types (dates, integers, reals, booleans, etc.)
- Key alignment
- Missing value consistency: null vs. blank vs. “N/A” vs. sentinel value

Importance of metadata:

Provide the context and conditions for key measures.

- What metadata might you need to understand the answer to the questions you tested?
  - Example - Purification condition prediction
  - If you are missing key metadata it can render your conclusions incorrect
  - Data volume and expertise will not compensate for incomplete context

What form do you want your data in for ML?

- No universal format. Transforms required - scalars, vectors, tensors
- Ideally you have a query mechanism - e.g. SQL
- Put it somewhere Python interactable. Lots of mechanisms built into Python for ML

Fundamentals & Best Practices

Automation:

Automate components of the data lifecycle to collect higher volumes of clean data

Use Automation to:

- Collect higher volumes of data without human intervention
- Start with the raw data where possible
- Ensure consistency of collected data
- Process and analyze consistently
- Store data where expected
- Save time, improve efficiency
- Reduce human error

Fundamentals & Best Practices

Institutional Support

- Hire data architects and data engineers - develop the data a architecture - comprehensive infrastructure to deal with this data lifecycle
- Deliver value to scientific staff effectively
- Use data management tools intended to govern the data lifecycle
- Work with institutional leadership to build a data driven culture

How does LabKey help?

LabKey®

Our software helps you organize, integrate, track, explore and analyze data.

SDMS
Integrate, process and analyze scientific data

Biologics LIMS
Data management for accelerating antibody discovery operations

Sample Manager
Track and manage the full life-cycle of lab samples