Applying Advanced Data Analytics to a Cell Growth Challenge in Commercial Biologics Manufacturing

Yiming Peng, PhD

Director, Nonclinical Biostatistics

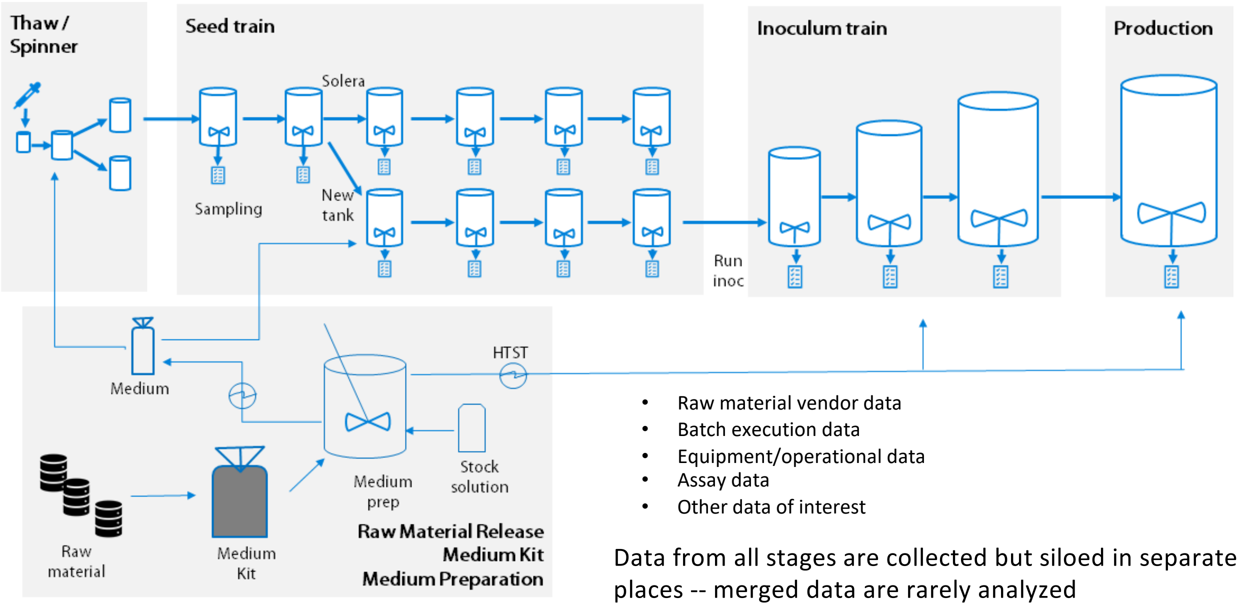
Genentech, A Member of the Roche Group

NCB Conference, 2021

Outline

- Problem Statement
- Advanced Data Analytics (ADA) Approach
- Study Results
- Lessons Learned

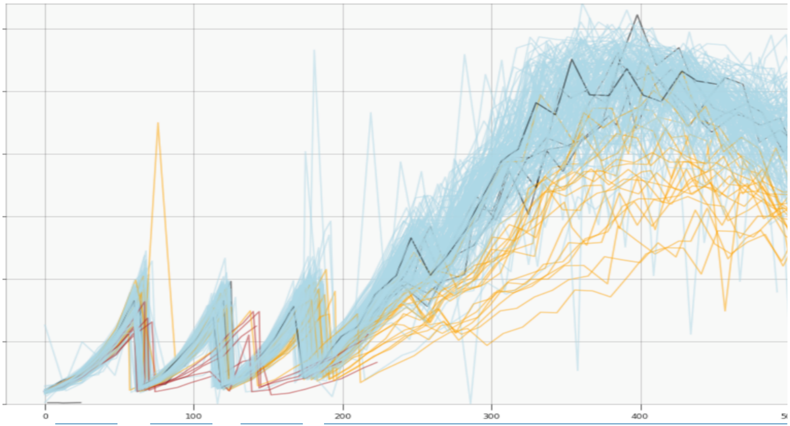
Background – A CHO Cell Culture Batch Process



- Raw material vendor data
- Batch execution data
- Equipment/operational data
- Assay data
- Other data of interest

Data from all stages are collected but siloed in separate places -- merged data are rarely analyzed

Problem – Cell culture experienced a poor growth with no root cause identified



Inoculum train **Production**

Project Goals

Capabilities

Determine whether Roche could benefit from the organizational and data analytics approaches established by a global consulting firm

Learn and enhance our capability to perform future advanced data analysis

Data

Connect siloed data sources from all stages: raw material to production culture

Analytics

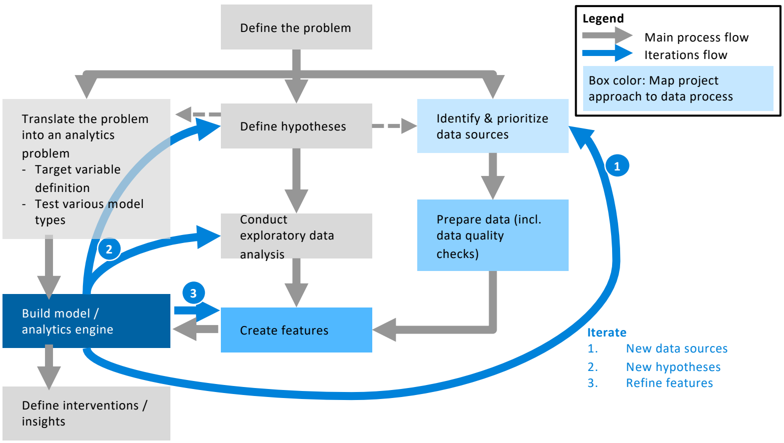
Identify improvement opportunities for the cell growth issue

Outline

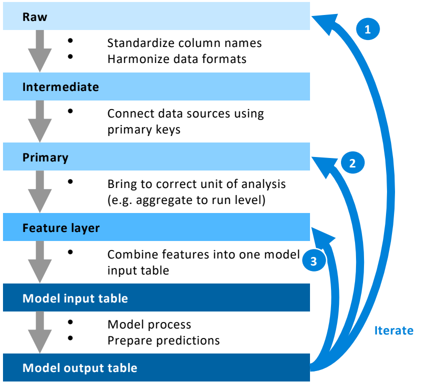
- Problem Statement
- Advanced Data Analytics (ADA) Approach
- Study Results
- Lessons Learned

Analytics problem-solving requires translating a business problem into an analytics problem and solving it through an iterative process of hypotheses generation, exploratory analysis, feature creation and modelling

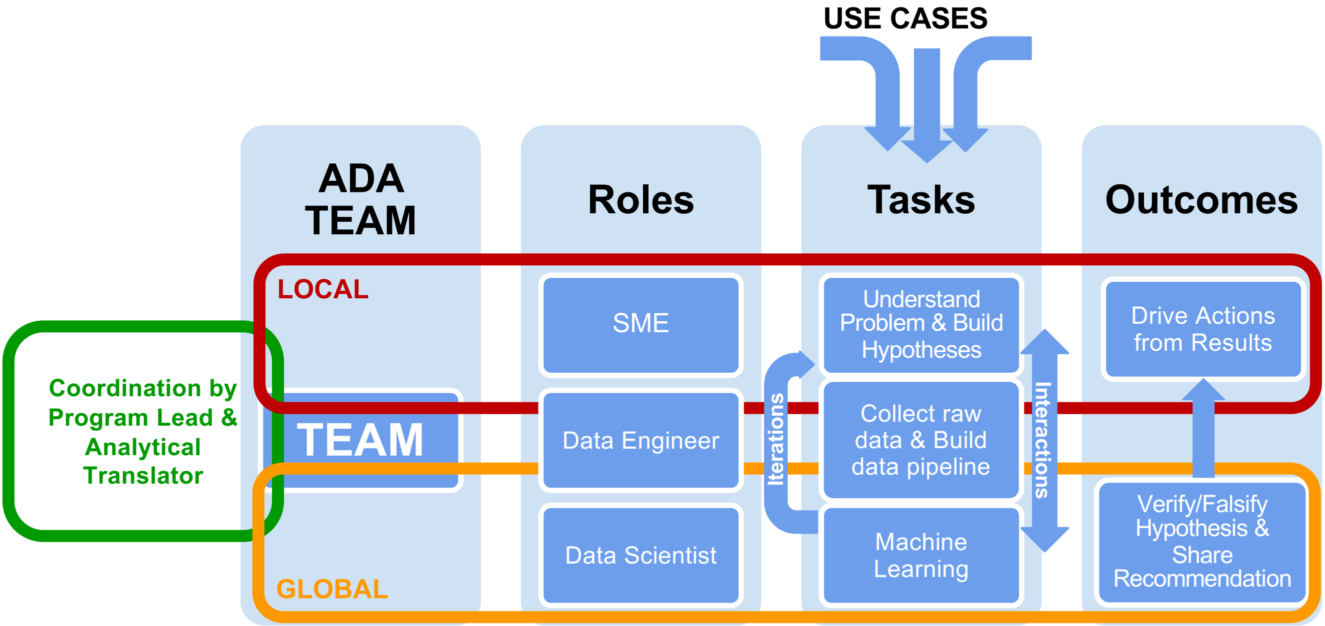
Project approach



Data transformation process



Advanced Data Analytics Process

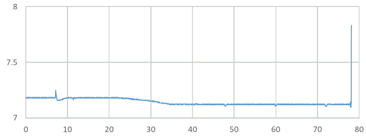


DE=Data Engineering
DS=Data Science
SME= Subject Matter Expert

Feature engineering transforms raw data into features that better represent the underlying problem and thus improves the machine learning models

Raw & Intermediate

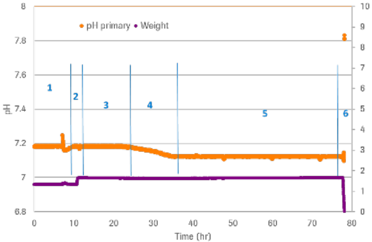
- An example online pH data in one inoculum train



Ingest and clean up the raw pH data

Primary

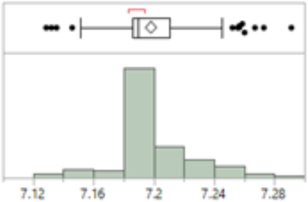
- Separate pH trend into six different phases - 1: medium batch; 2: inoculation; 3: pH at the top of deadband; 4: pH transition from top to bottom of deadband; 5: pH at the bottom of deadband; 6: transfer out



Investigate pH and fermenter weight trends together

Feature

- For each pH phase, explore the pH min, max, average, as well as the duration and its proportion to overall duration
- For culturing phase, explore pH excursions: spike, change due to online/offline adjustment, etc.



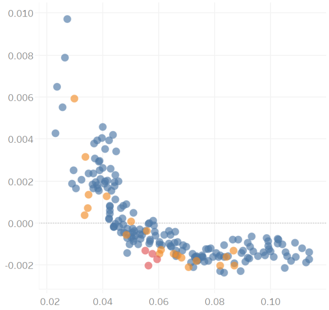
Phase 3 max pH for all runs showed large variation

Modelling Approach

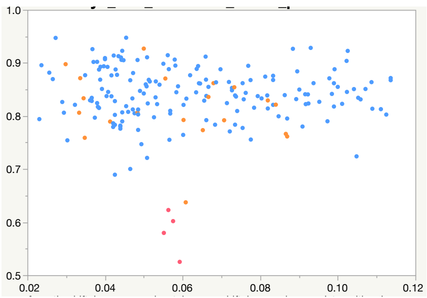
- Model preprocessing: Missing data imputation, Variance Inflation Factor (VIF)
- Model techniques: Random Forest, eXtreme Gradient Boosting (XGB), LASSO, etc.
- Model interpretation: SHapley Additive exPlanations (SHAP), Variable Importance, Plot the data

Be careful with SHAP and general model based interpretation, especially when the model fitting is not ideal

SHAP/Model says negative correlation



Actual data doesn’t show any negative correlation



ALWAYS check the actual data!

Outline

- Problem Statement
- Advanced Data Analytics (ADA) Approach
- Study Results
- Lessons Learned

Overview: Progress over 16 weeks

Roche

Data

11

Data sources ingested. 7 data sources linked

800 GB

of data processed

6,000

Pages digitized from pdf

Analytics

100+

Hypotheses identified for causes of growth issues and translated into >800 features

500+

Graphs generated for exploratory data analysis

13

Iterations on the machine learning models

Capabilities

40+

Hours of knowledge sharing sessions with technical and SME team

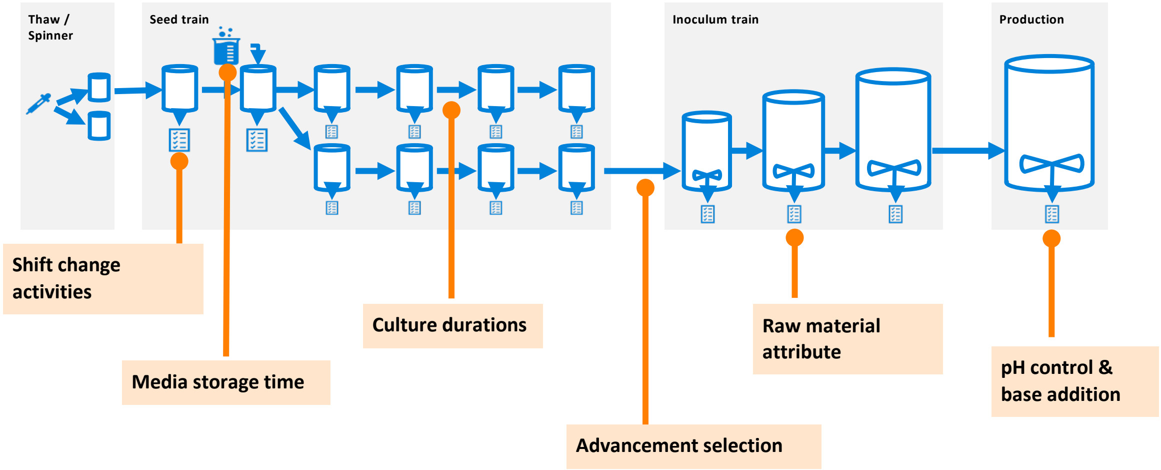
10+

Hours of translator training / lunch & learn sessions

1

Tech environment created in Roche GCP (Google Cloud Platform: Cloud Storage, Dataproc, BigQuery); Github, Python/R, JIRA, Confluence

Highlighted insights and actions to cell culture process



Outline

- Problem Statement
- Advanced Data Analytics (ADA) Approach
- Study Results
- Lessons Learned

Lessons Learned (1/3) – Organization

- SME input is the key for successful data analysis
  - SMEs include manufacturing, development, assay, raw material, quality - ADA team leaders need broad networks
  - Data analysis is driven by both DS and SME

- Cross functional team is a must
  - Highly specialized skills from various functions
  - Dedicated Resource and Clear R&R
  - Passion is great but don’t poach, respect functional SME
  - Translator manages the expectation - Chaos if everyone is doing what they want to do

- ADA is an iterative process of knowledge discovery
  - Requires multiple sprints with SME inputs
  - Downtime between sprints to reflect and think
  - Sprints fit well for data engineering but may not be amenable to data science

Lessons Learned (2/3) – Strategy

- Focus on actionable changes within the license range for commercial manufacturing

- Opportunities likely lie in "less-controlled" process parameters and raw material attributes
  - Critical Process Parameters are well controlled with small variation around their target setting

- How can development use this knowledge to improve the next molecule?

- Working with External collaborators,
  - Must understand exactly what's being done – Statistician is needed
  - Challenge when it doesn't makes sense, and improve accordingly
  - Be aware of vendor's canned analysis
    - SHAP (explainable AI) does not solve all the problem
    - Having a hammer doesn't mean everything is a nail

Lessons Learned (3/3) – Statistics

- Prediction requires a stable process - special cause variation is not predictable (Shewhart)

- Runs are highly correlated within campaign,
  - Most statistical techniques require independence
  - Using campaign as the experimental unit significantly reduces sample size
  - Cross validation overestimates accuracy => split the data by campaign

- Cloud Computing allows exhaustive search, e.g. all 2 or 3-way interactions of 100s of features
  - Due to effects hierarchy unlikely to discover 2-way interactions
  - Reporting that all 2-way interactions are investigated is important to SMEs

- Data are on different scales
  - CART is invariant to monotonic transformations

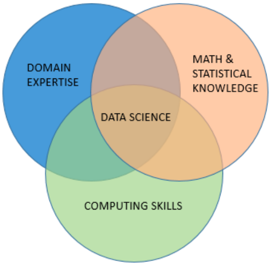
- Black box prediction algorithms are hard to interpret
  - Interpretability is required for manufacturing trouble shooting
  - Balance interpretability vs. performance

Statistics and Data Science

- What's Changed?
  - Implementations of K-Nearest Neighbors (Fix & Hodges 1951), CART (Breiman 1984), random forests (Tin Kam Ho 1995) and many other "machine learning algorithms" are so easy in R or various cloud technologies

- What hasn't Changed?
  - Assessment of process stability (SPC) as a requirement for prediction
  - Assessment of bias due to missing data, influential data points (i.e., outliers), correlated features
  - Selection of methods, e.g. with appropriate scale invariance properties
  - Understanding effect of lack of independence between runs on prediction and evaluation of accuracy
  - Is correlation causation - need experiment (DoE)

- What's Better for Staffing?
  - A statistician, a computer scientist, and a SME vs. 3 data scientists?
  - Should we be training statisticians to program, or to work effectively in cross functional groups?



Doing now what patients need next