Brushing off Old Data

Dr Bob Sochon,
GSK Consumer Healthcare
What I’m talking about today

Brushing off Old Data - GSK - MATLAB Expo 2019
Manufacturing Overview

Challenges & Opportunities

- Regulated industry
  - In many markets Sensodyne is regulated akin to a medicine
  - Change and experimentation is difficult

- Batch processes
  - Many isolated data sources
  - Never exactly the same day – day
  - Many interconnected steps

Can we learn from our history to make better products?
Big Data

What does this mean?
What have we got?

5 Years of Data
5 Similar Mixers
10 Formulations

Every Process Variable

10,000 Batches

Batch & Operations Information

Online Monitoring

Analytical Results

Big Data
60 GB of files

Big Results?
What can we do with all this data?

Univariate Statistics
- Long term averages, trending, Production Support

Multi-Variate Statistics
- PCA
- PLS

Machine Learning
- Classifications
- Predictions

Process Improvement and Understanding

Improved Manufacturing Efficiency & Reduced Cycle Time

Improved R&D → Factory Product Transfer

Brushing off Old Data - GSK - MATLAB Expo 2019
Why MATLAB?

- Ability to link different data files & formats together
- Quickly iterate code
- Format / library independence
- Commercially supported / maintained
- Easy consulting support
- Build GUI to enable easy observation of data

Quick & easy code development cycles
Results

What did we do?
Process Workflow

Cleaning & Context
Phase Separation
Data investigation
Results

Typical Workflow
Process Workflow

Cleaning & Context

Phase Separation

Data investigation

Results
Data Visualisation

Excel

Process Historian

Operations Database

Analytical Archive

MATLAB

Brushing off Old Data - GSK - MATLAB Expo 2019
Process Workflow

- Cleaning & Context
- Phase Separation
- Data investigation
- Results
Typical Batch Trace
Overall Batch
Adding Silica

Brushing off Old Data - GSK - MATLAB Expo 2019
Batch Finishing

Brushing off Old Data - GSK - MATLAB Expo 2019
Improvments Over Time

Can now **automatically** pick phases out of data and start to ask simple questions:

i.e:

1) Does process time improve over time?
2) Is processing time mixer dependant?
3) Do our process improvements work?
Process Workflow

- Cleaning & Context
- Phase Separation
- Data investigation

Brushing off Old Data - GSK - MATLAB Expo 2019
Linear Correlations

Too much Data to easily see!

Can see relationships in the data

Needs further understanding
Can link final batch properties to the batch to each phase of the data

Not all mixers perform the same
  - Confirms our operator reports
Process Workflow

Cleaning & Context

Phase Separation

Data investigation

Results
Multivariate Understanding
Brushing off Old Data - GSK - MATLAB Expo 2019

Cleaning & Context

Phase Separation

Data investigation

Results

Process Workflow
What have we done?

- We had / have a lot of data that is very siloed, in many different systems

- We finally have a way to match all Batch, Process & Analytical information together.
  - Can only get better as systems progresses in GSK

- We can reliably deduce information about phasing and performance of the mixers, without human intervention.
  - Reduce time and error processing, easier conclusions.
  - Can do this across formulations, mixers (& years)
  - **We can now ask for conclusions in process data** – i.e. does “this” effect “that”?  
  - Every growing as our factories make more batches!
What have we learnt?

– Not all our variability is random
  – For example different mixers seem more important than different processing times
  – More inputs will likely make better models, but never going to be perfect

– Big Data is not a panacea! – It will not immediately solve all our problems

– We still have to do experiments to generate understanding
  – Machine Learning / Multivariate Statistics can’t replace science.

– Allows us to guide experiments to examine impacts of input variables.
However: There are no SIMPLE easy correlations in the data

1. We might not have enough data:
   Is there enough data here? – we only have 200 – 300 data points per batch

2. It might not vary enough to make clear versions of truth
   Because of GMP / Factory nature of data i.e. mixer speeds seem mostly similar between batches

3. We might not be measuring the right thing
   Our test methods don’t exactly detect what we’re looking for. Viscosity vs Rheology for example