We create chemistry

Discovering hidden relationships in production data

Elie Maricau - BASF Antwerp

Context: the BASF Antwerpen production site

The BASF Antwerpen site is the **second largest production site of the BASF Group**. It's 55 production plants mainly consist of large scale continuous processes and produce commodity chemicals.

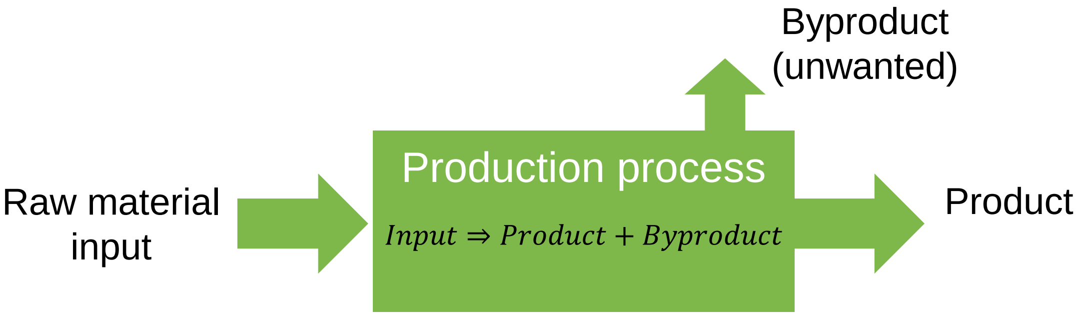
It's size provides sufficient critical mass for sustaining site-central expertise teams related to manufacuring and it's directly supporting functions. The highly integrated site (product streams, utilities, logistics) in combination with the presence of third parties provides a unique set of challenges

BASF
We create chemistry

Context: the problem

Project context

- Continuous production process
- High raw material cost
- Production of unwanted byproducts reduces process efficiency
- Process efficiency varies over time



$$Yield = \frac{Product}{Product + Byproduct}$$

Illustrative picture of continuous production plant in BASF Antwerp

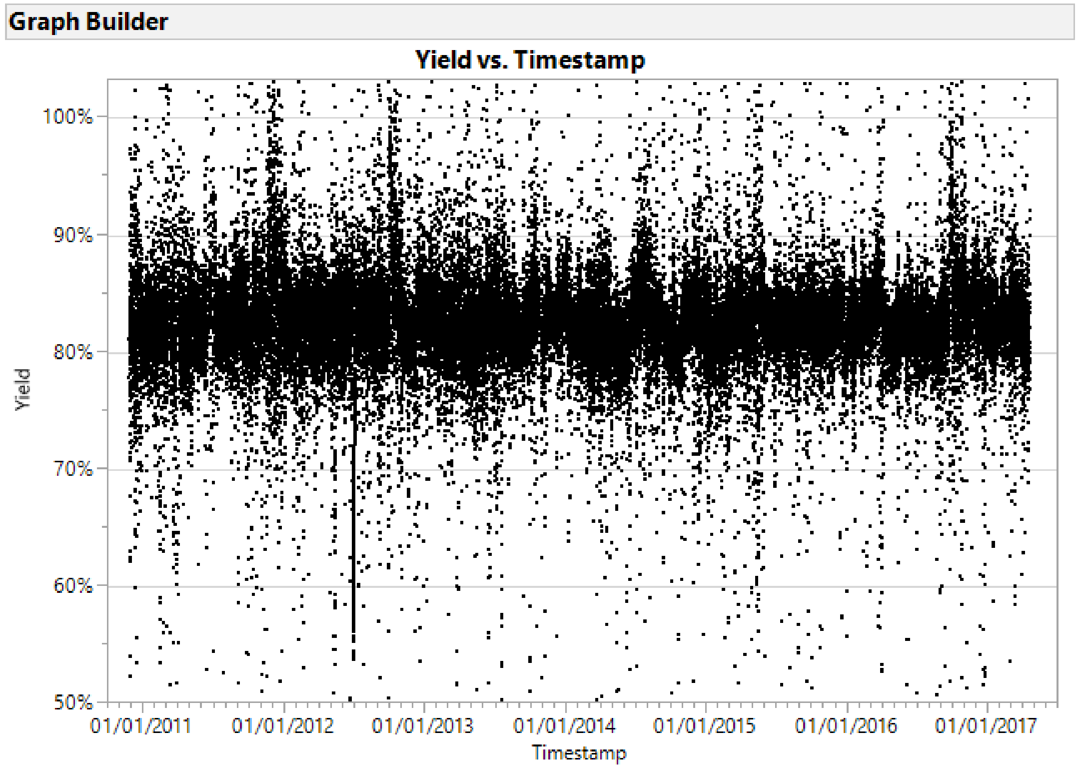
What data do we have?

- PIMS system: sensor data, stored in a big database with time series (vectors)
- LIMS system: lab data, stored in a separate database with sample times and lab values
- Operator logbook: manual text entries logging specific actions with a timestamp

Context: the problem in numbers

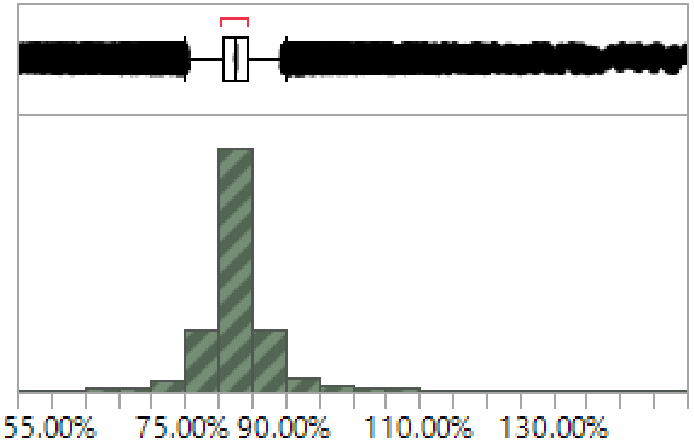
Project goal

- Improve average production yield – improvement potential unknown at start of the project



Distributions

Yield



Summary Statistics

Mean	0.8272471
Std Dev	0.0731346
Std Err Mean	0.000312
Upper 95% Mean	0.8278586
Lower 95% Mean	0.8266355
N	54943

BASF
We create chemistry

Let’s solve the problem – attempt 1

- Lots of data available
  - 5 years data (hour values)
  - 250+ sensors
- Use statistical algorithms and data analytics techniques to identify key variables for the yield

Yield = f(X1, X2, X3, …)

Select X1, X2, X3,… from a set of 250+ variables

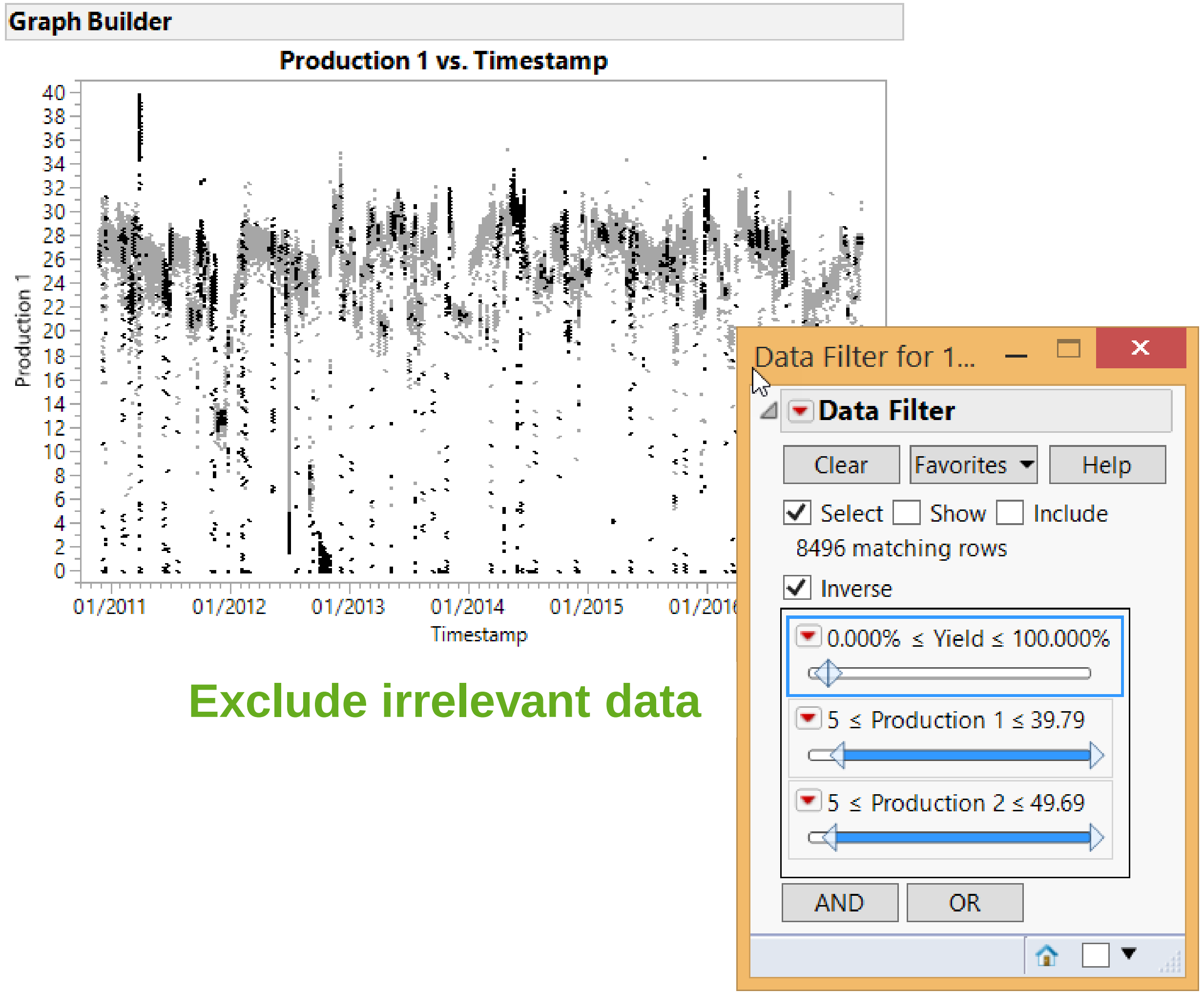
BASF
We create chemistry

Discovering key process variables
Attempt 1 – data crunching

- Collect 5 years data for all measurements related to that part of the production process (online sensor + offline lab data)
  ➔ over 250 variables

- Exclude irrelevant data: no or very low production output, yield<0% or yield>100%

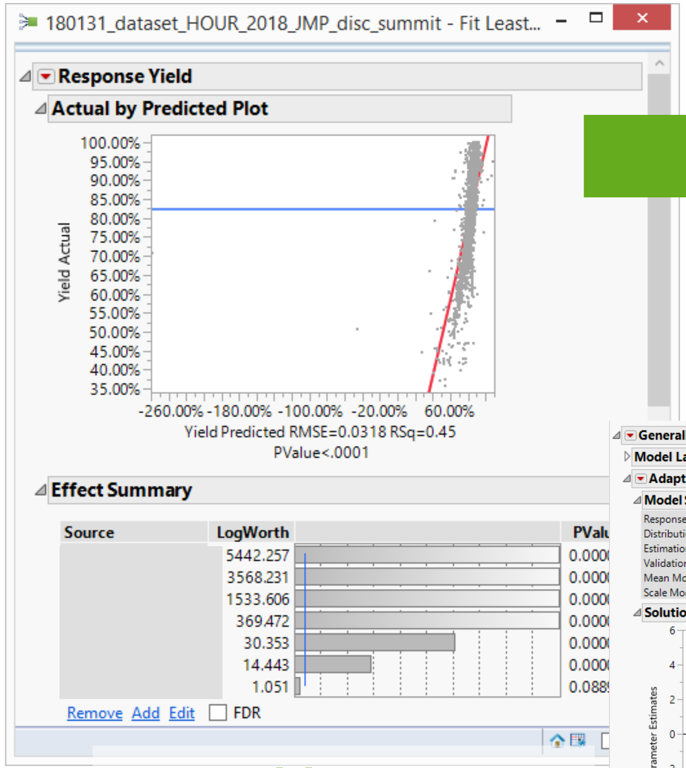
- Find root causes for yield variation: try various statistical models
  1. Stepwize OLS
  2. Partial Least Squares
  3. Generalized regression Enet (JMP pro)
  4. Bootstrap forest (JMP pro)



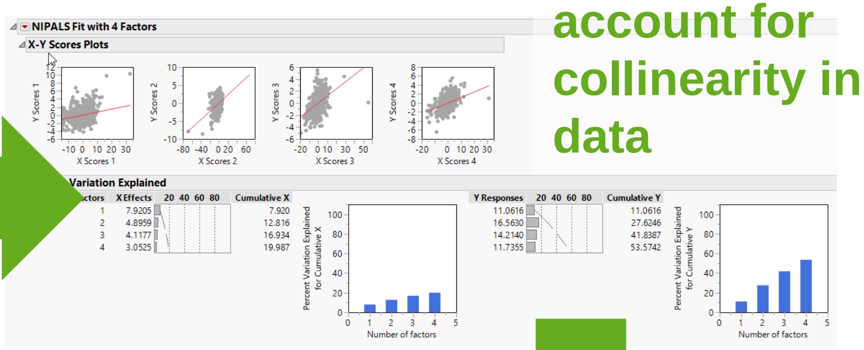
Exclude irrelevant data

BASF
We create chemistry

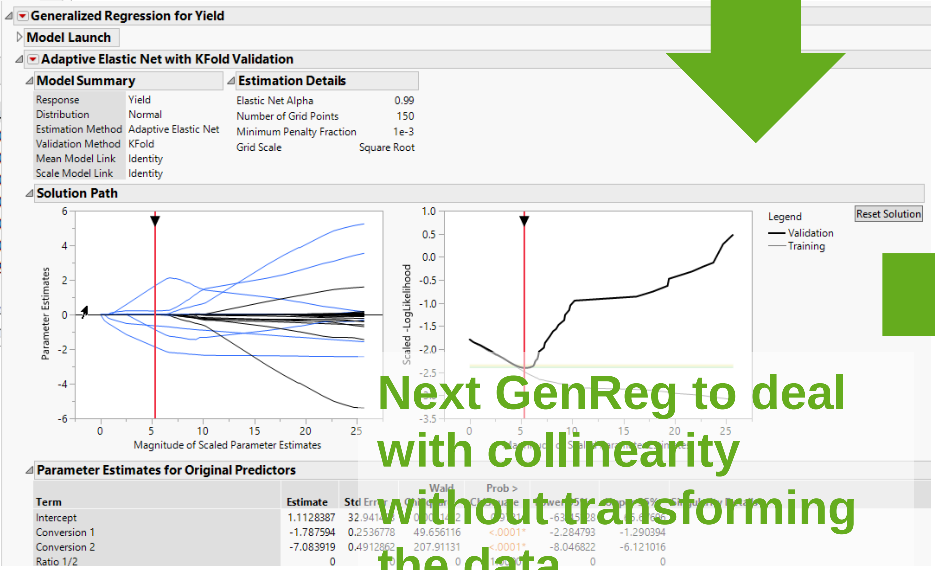
Discovering key process variables
Attempt 1 – data crunching



Start with OLS



Then PLS to account for collinearity in the data



Next GenReg to deal with collinearity without transforming the data

Finally a random forest to account for interactions and non-linearities

BASF
We create chemistry

Discovering key process variables
Attempt 1 – data crunching

180131_dataset_HOUR_2018_JMP_disc_summit - Model Comparison -...

Model Comparison

Predictors

Measures of Fit for Yield

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
OLS Yield	Fit Least Squares		0.4516	0.0318	0.0176	47561
PLS Yield	Partial Least Squares		0.2762	0.0348	0.0239	2648
BF Yield	Bootstrap Forest		0.6853	0.0241	0.0137	47562
GenReg ENET Yield	Fit Generalized Adaptive Elastic Net		0.3101	0.0357	0.0222	47561

OLS	PLS	GenReg	RF
Conversion 2	**Conversion 2**	**Conversion 2**	**Conversion 2**
Production 2	Total production	Conversion 1	Total production
Flow_1	Conversion 1	Production 2	Ratio 1/2
Level_1	Level_1	Feed 2	Conversion 1
Pressure_1	Temp_1	Quality_1	Level_2

- None of the models performs great (especially the RASE/RMSE is too large compared to the target) – note in the actual study training/validation and test data has been used
- Key process variables in each model are different (expect for “conversion 2”)
- Some of the parameters cannot be explained from a expert point of view (creates skeptism)
- Although there is some predictive power, none of the models is good enough to optimize the process (what are the ideal process settings?)

Let’s solve the problem – attempt 2

- Ask the subject matter expert: what do they think X1, X2, etc. is?

Yield = f(X1, X2, X3, …)

Select X1, X2, X3,… from a set of preselected (SME input) variables

BASF
We create chemistry

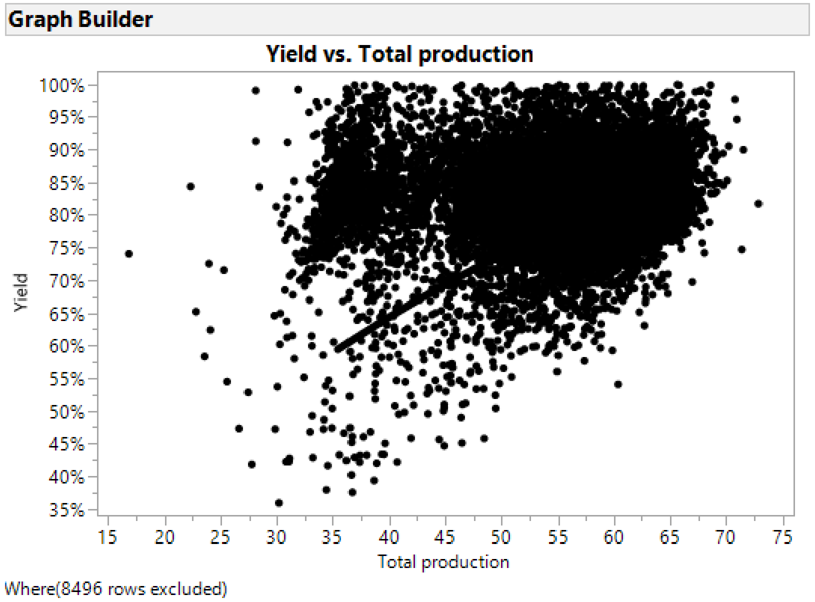
Discovering key process variables
Attempt 2 – ask the process expert

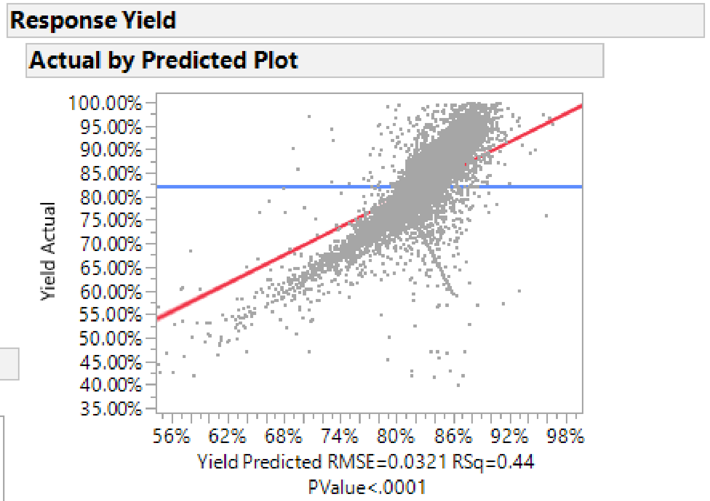
- Extensive interviews with plant management, plant operators and technology experts
  - Shortlist of suspected key process variables
  - Three categories
    - Measurement noise
    - Production (production planning, hard to change)
    - Other process settings

BASF
We create chemistry

Discovering key process variables
Attempt 2 – ask the process expert

- In an OLS model, a subset of expert defined parameters are relevant.
- Performance of the model is on par with the data crunching OLS model, which is still not good enough
- Parameters suspected to have the biggest impact (total production) is not as relevant as suspected





Effect Summary

Source	LogWorth	PValue
Conversion 2	5253.518	0.00000
Production 2	2134.664	0.00000
Feed 2	1835.876	0.00000
Feed 1	27.497	0.00000
Production 1	8.943	0.00000
Conversion 1	2.827	0.00149
Ratio 1/2	2.771	0.00169

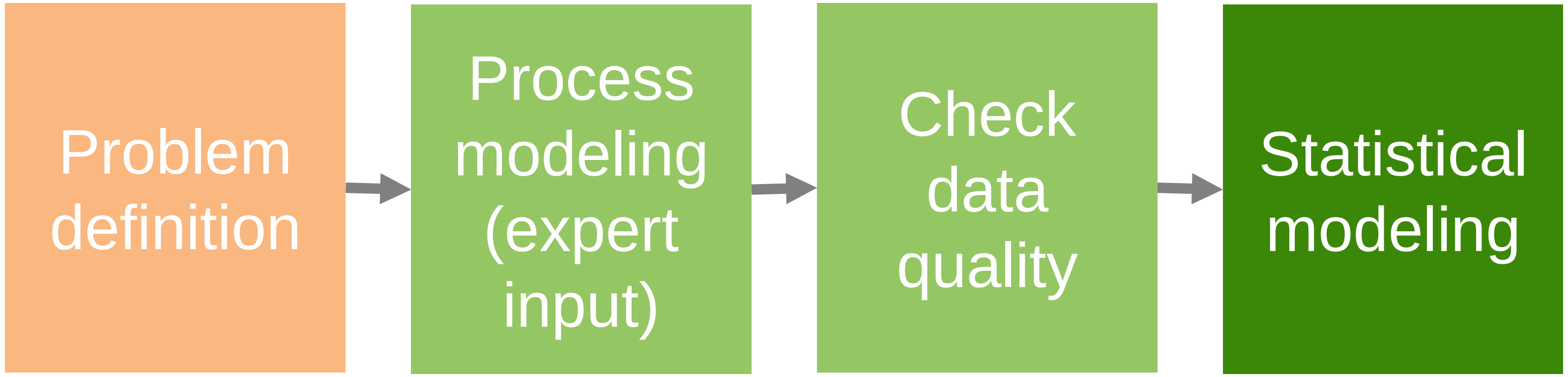
BASF
We create chemistry

Let's solve the problem – attempt 3

- Improve the data quality.

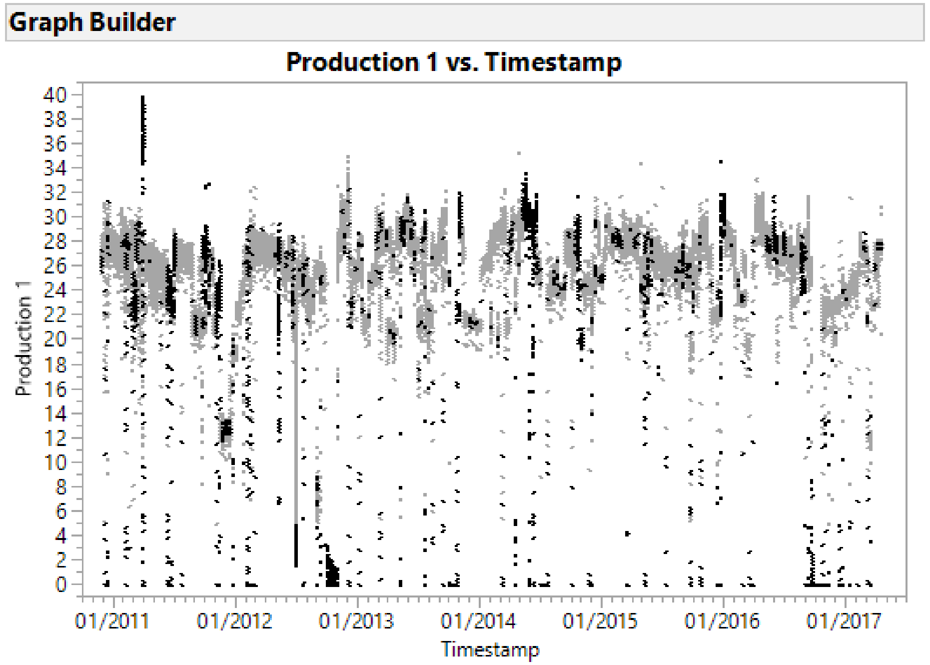
Yield = f(X1, X2, X3, …)

Select X1, X2, X3,… from a set of preselected (SME input) variables



Discovering key process variables
Attempt 3 - data quality evaluation

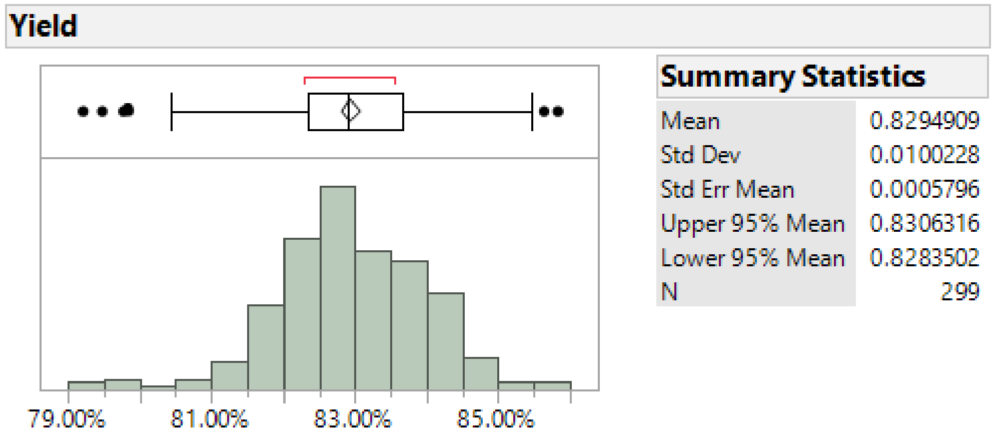
- Step 1: only look at periods with “normal” operating regimes (untill know, our data cleaning was limited to this)

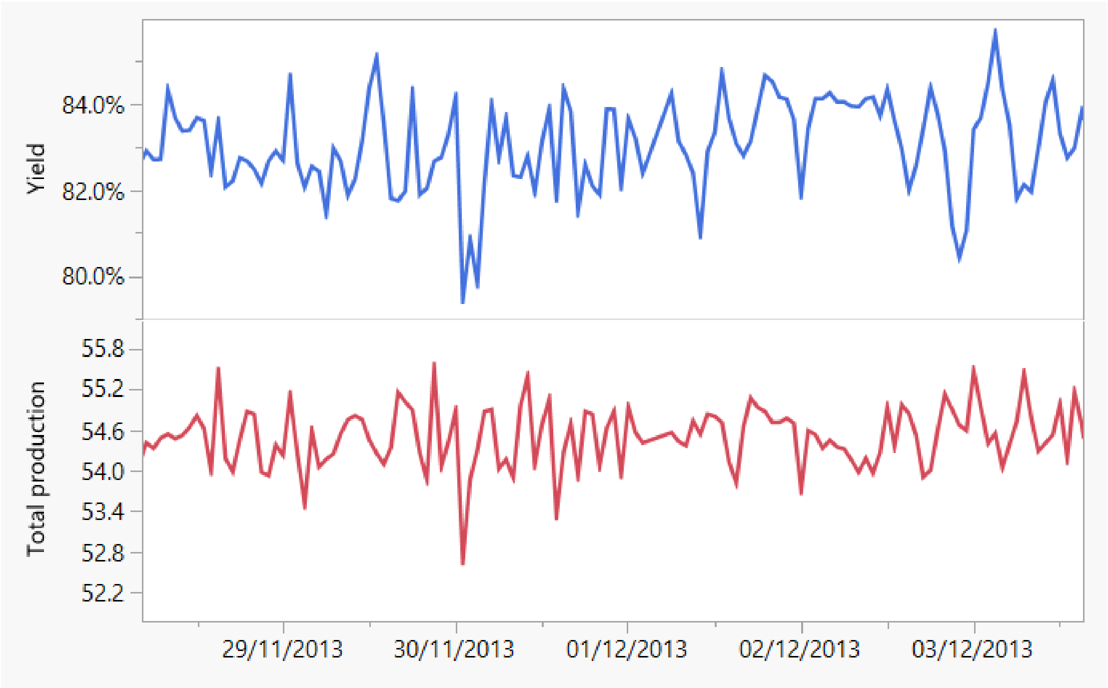


Discovering key process variables
Attempt 3 - data quality evaluation

- Step 2: measurement noise
  - Measurement system analysis: 1% variation (stddev – 15% of overall variation) in yield measurement due to variations in flow measurement

➔ Solution: look at daily (24H) averages (or medians) instead of hourly values (reducing the measurement error by approximately 5)



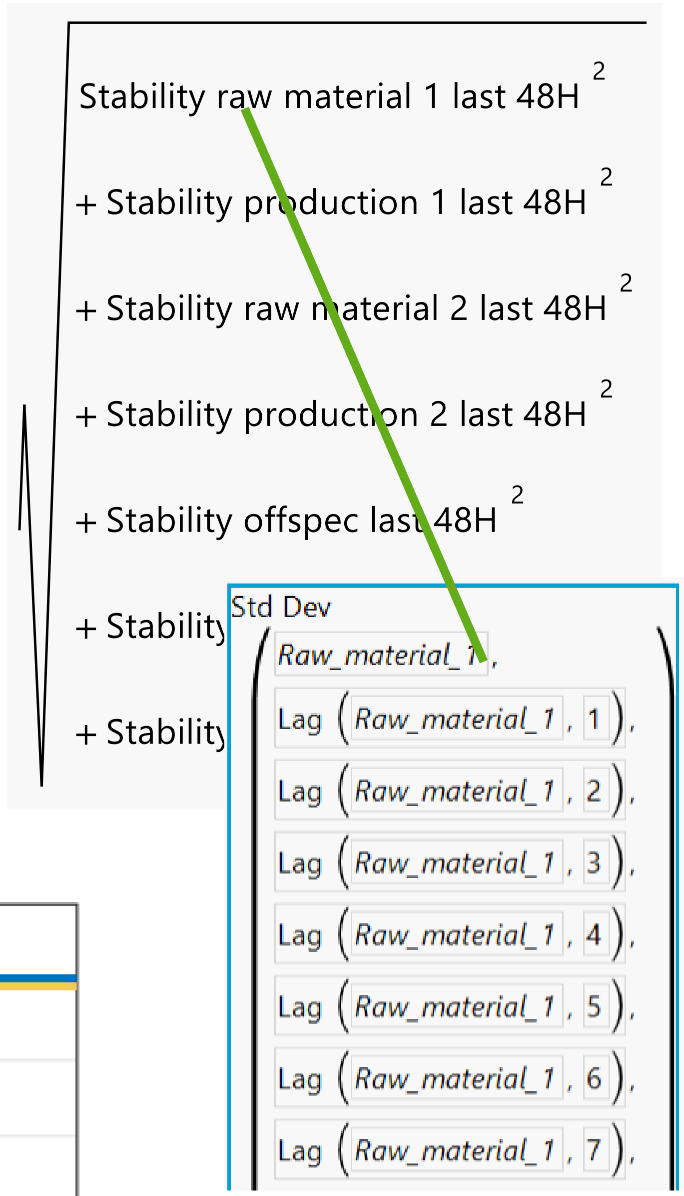


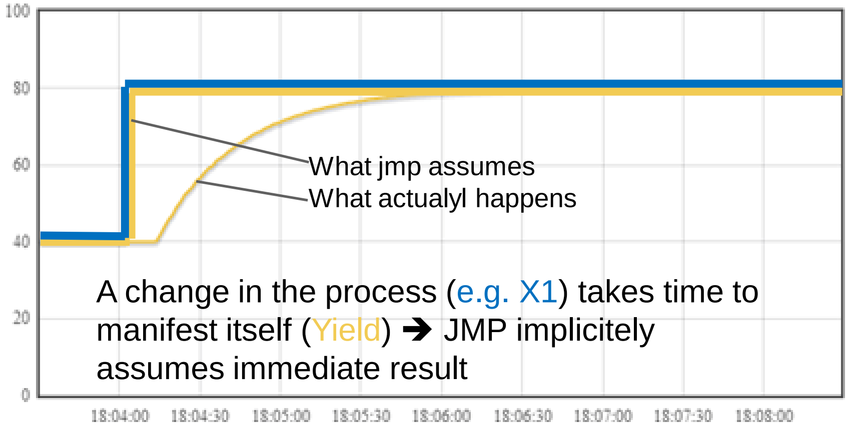
BASF
We create chemistry

Discovering key process variables
Attempt 3 – data quality evaluation

- Step 3: dynamic effects
  - Dynamic effects: after changing the process, it takes up to 48hours to get to a new steady stade condition (and often another change is made within that time ➔ seldom at steady state)

➔ Solution: formula column to identify moments where the process is stable for at least 48H (look at overall 48H stddev of all major production flows)

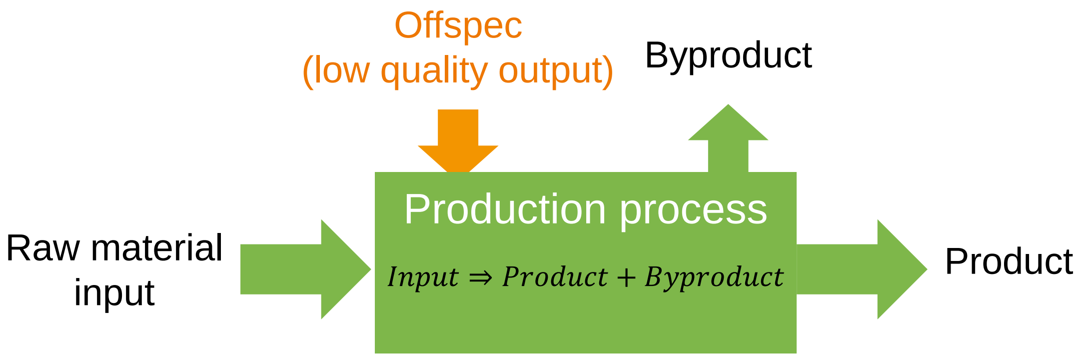




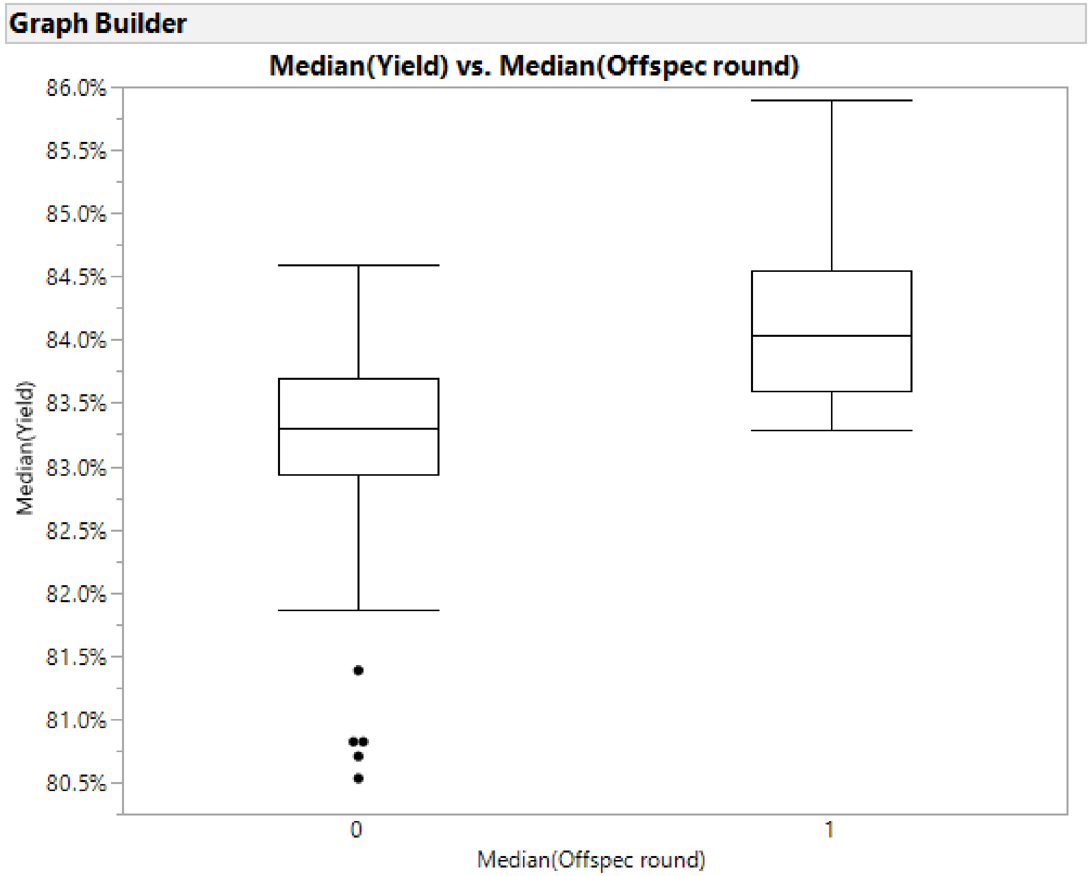
BASF
We create chemistry

Discovering key process variables
Attempt 3 – data quality evaluation

- Step4: offspec intake
  - Intake of offspec product affects yield calculations (artificial increase of output wrt what is expected from amount of input product)
  - Note: offspec composition unknown ➔ impact on yield cannot be quantified



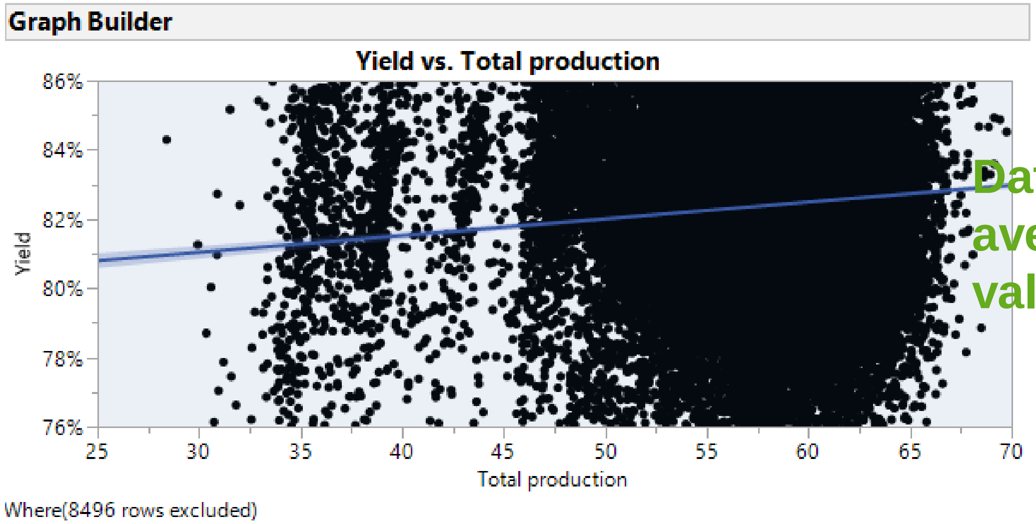
$$Yield = \frac{Product + Offspec}{Product + Byproduct}$$



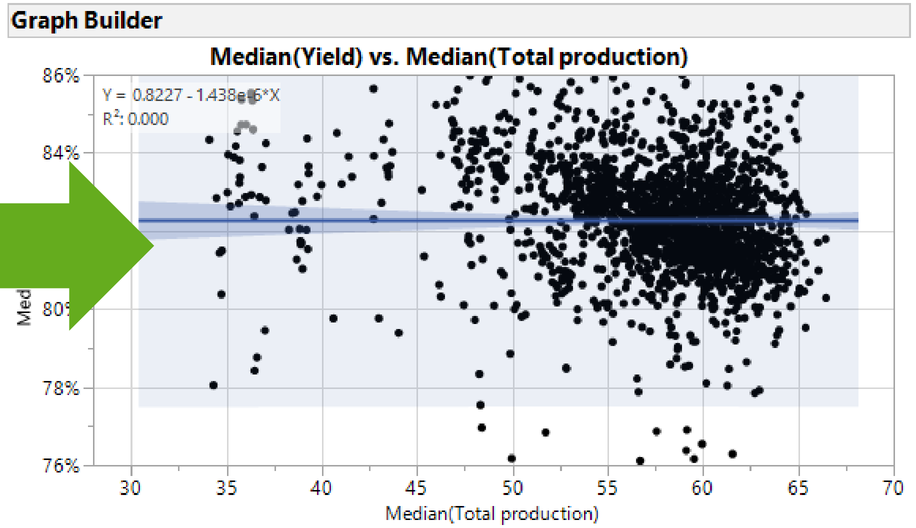
Where((Median(Offspec round) = 0, 1) and (Name("Median(Total production)") >= 45 & Name("Median(Total production)") <= 55)...)

We create chemistry

Discovering key process variables
Attempt 3 - data quality evaluation (summary)



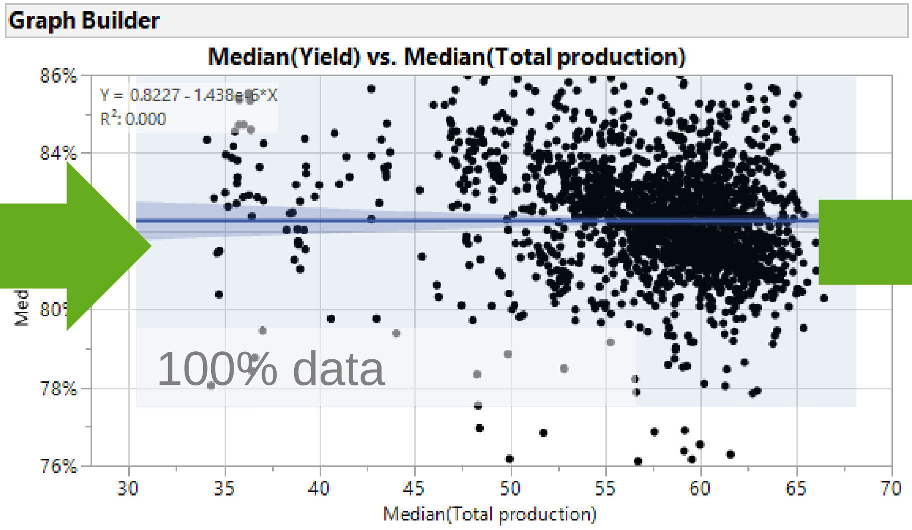
Data: 24H average values



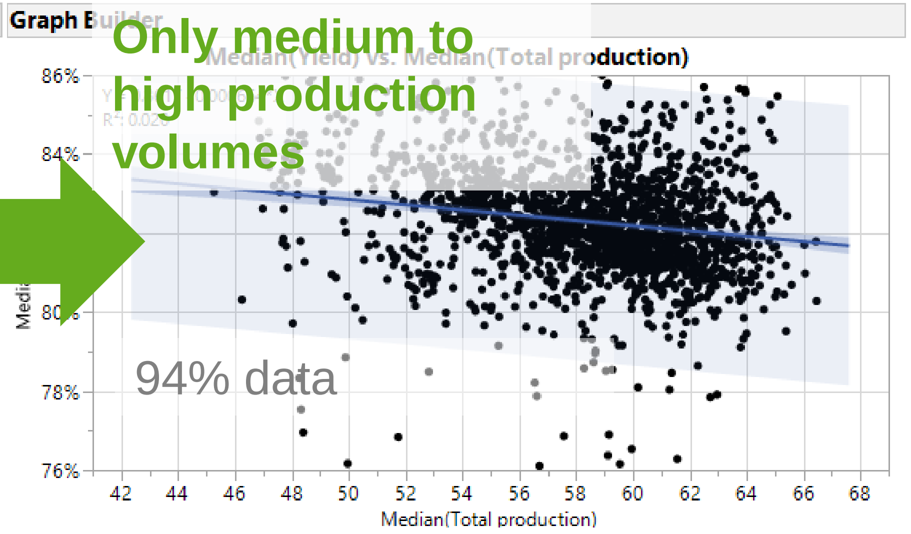
BASF
We create chemistry

Discovering key process variables
Attempt 3 - data quality evaluation (summary)

Data: 24H average values



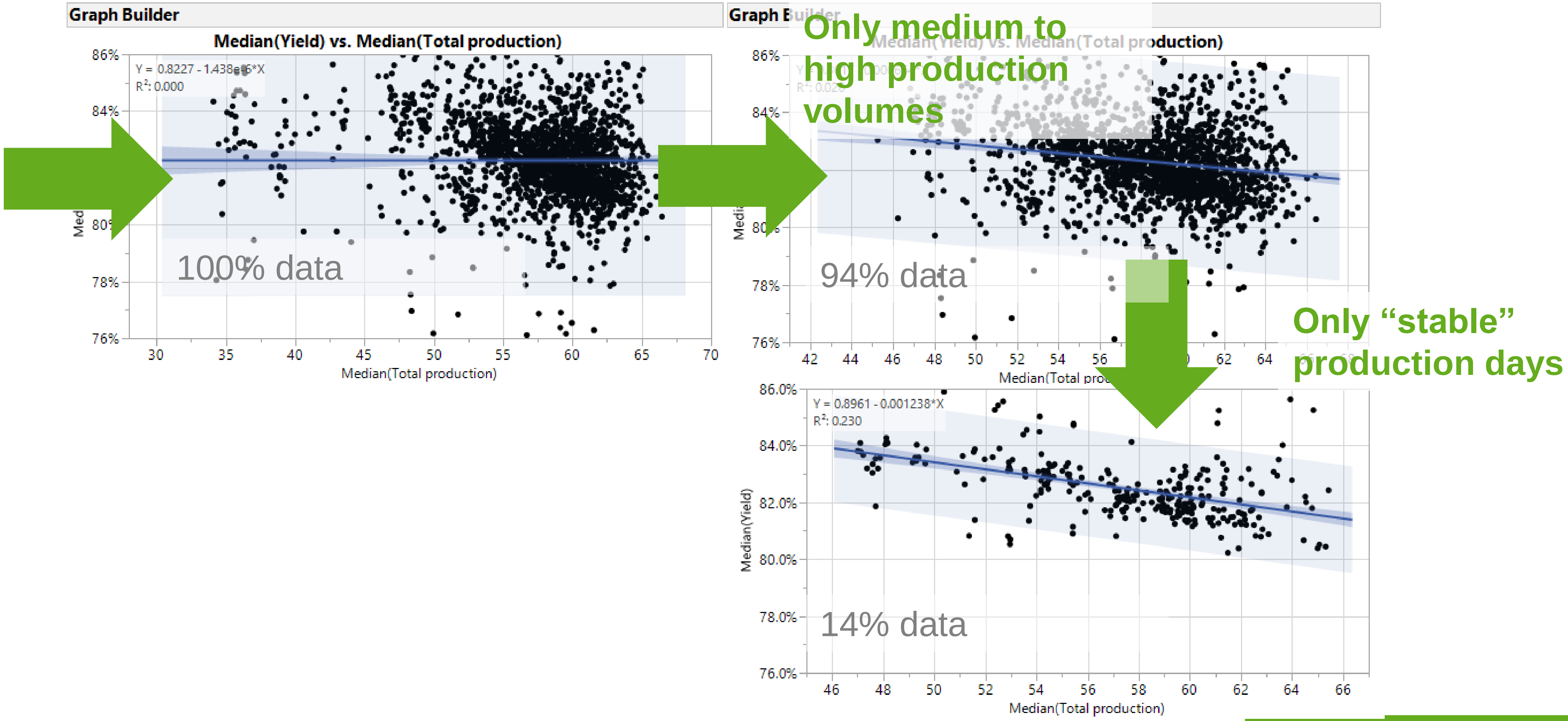
Only medium to high production volumes



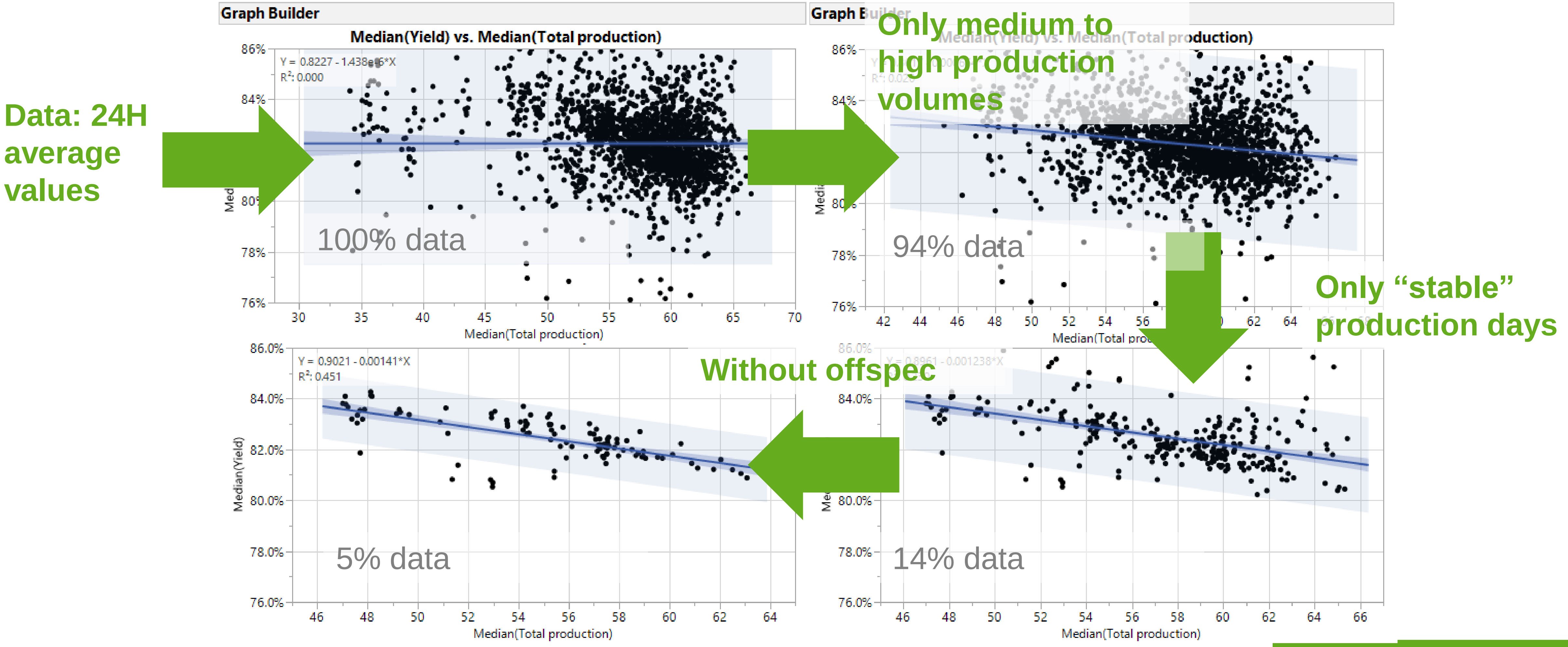
BASF We create chemistry

Discovering key process variables
Attempt 3 - data quality evaluation (summary)

Data: 24H average values



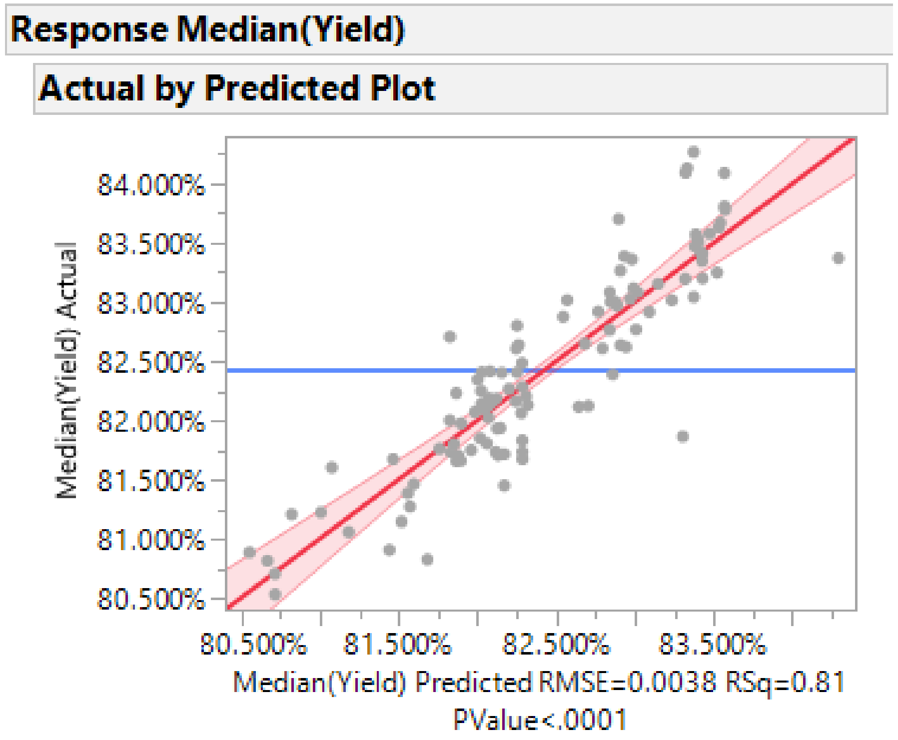
Discovering key process variables
Attempt 3 - data quality evaluation (summary)

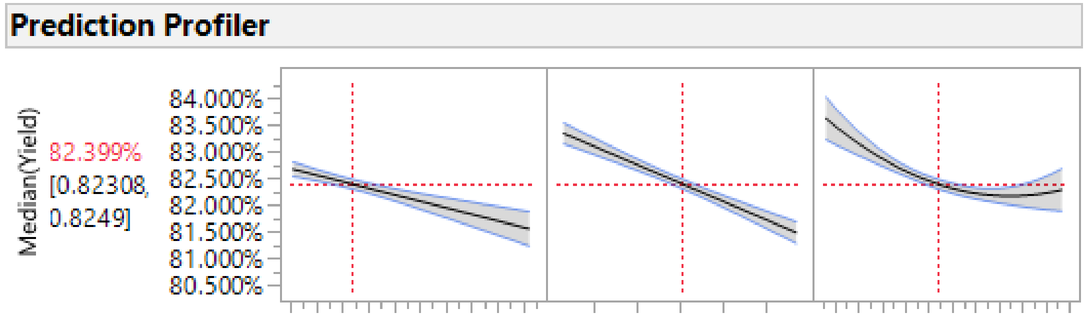


- Only 5% of the data is high quality data!

Discovering key process variables
Attempt 3 – data quality evaluation

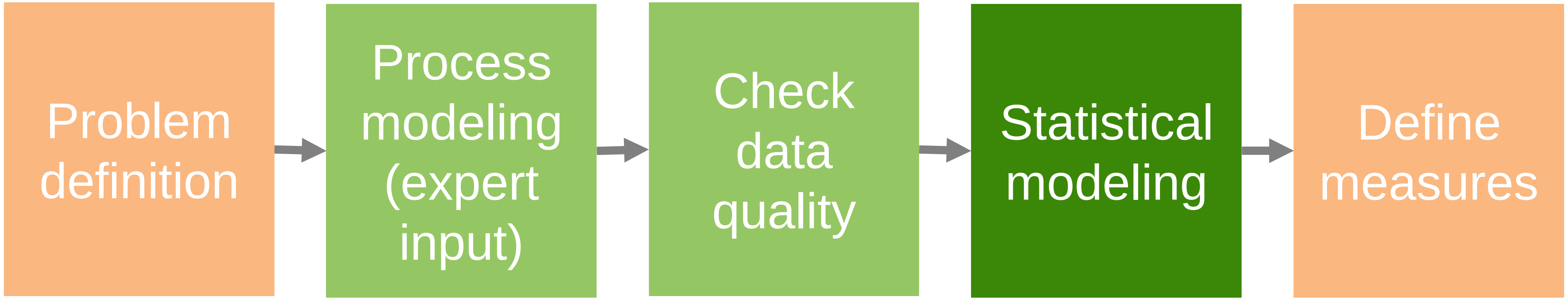
- OLS model on high quality data
- Key process variables can be explained by process expert
- Impact of key process variables can be accurately estimated (including an interaction and a non linear effect)
- Result
  - Optimal settings for yield (at a certain load)
  - Prediction of expected yield (an detection of deviations)





Knowledge based action

- Now we converted data into knowledge ➔ let's act on that knowledge



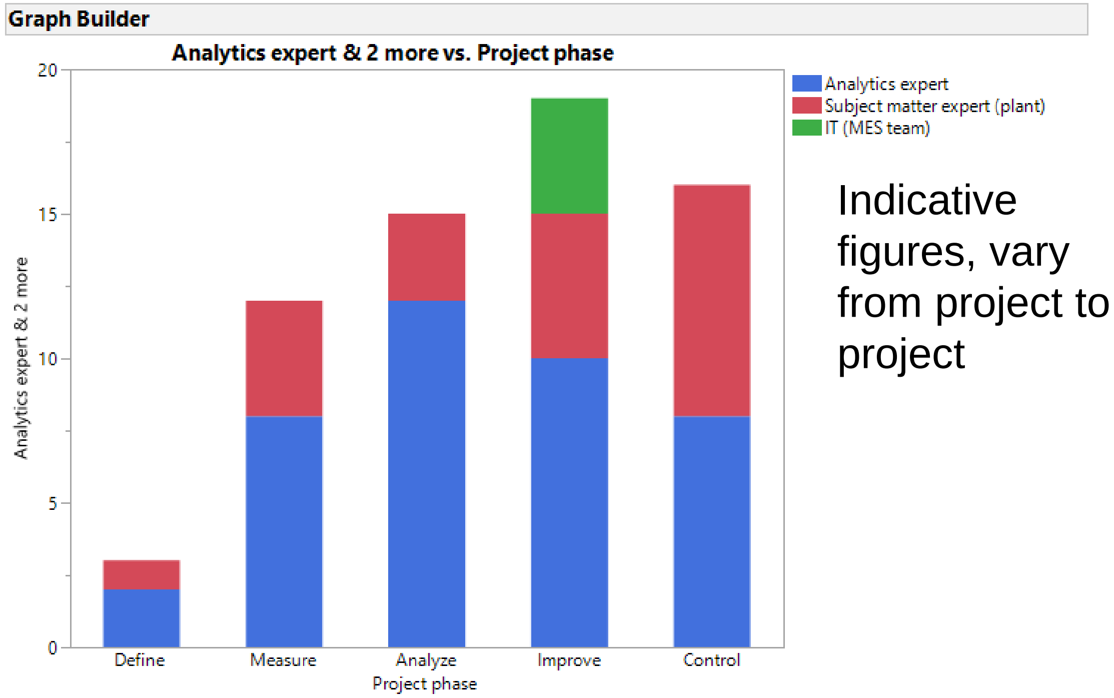
Knowledge based action

- Implementation of a dashboard in the control room
- Targets show where the process should be (for maximal yield)
- The actual yield WITH indication of reliability of that value is displayed
- A corrected yield value based on other process parameters (model)

BASF
We create chemistry

Its not only about technology…

- To be succesful, the analytics part (doing the datamining and modeling) is only 25% of the work!
- Communication (in all directions) and change are major succes factors
- A project lead (in our case the data scientist) must oversee the project from end to end (clear problem definition ➔ sustainable benefits)
- A LSS DMAIC project workflow (with a good amount of advanced analytics sauce in the measure and analyze phase) is a best practice



Indicative figures, vary from project to project

CONTROL

BASF
We create chemistry

Key take aways

- Succes (creating value) =
  - technology (JMP) +
  - data science +
  - expert input +
  - thorough data cleaning +
  - project based approach

- Handling time series data (which is typical for process industry) requires some specific approaches
  - Data preprocessing: measurement noise, dynamic effects, …
  - Modeling: colinearity, autocorrelation between consecutive data points, …

We create chemistry