Regression Strategies for Large Datasets

2017 AIChE Spring Meeting

Big Data Analytics & Smart Manufacturing

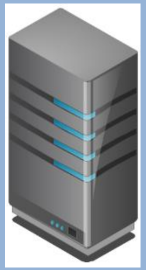
James C Cross III

March 28, 2017

Scope of Presentation

Goal: Leverage data to improve plant operations.

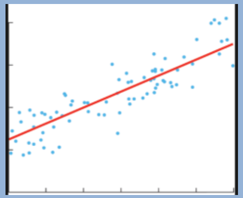
Plant Data



Predictors & Responses

t_op_d	online	batch_h	flow_kpi	w_mks	econ_a
1.67E-01	1.00E+00	4.00E+00	2.97E+04	3.01E+04	7.09E+02
3.33E-01	1.00E+00	8.00E+00	3.06E+04	2.97E+04	7.09E+02
5.00E-01	1.00E+00	1.20E+01	3.06E+04	3.01E+04	7.09E+02
6.67E-01	1.00E+00	1.60E+01	3.06E+04	3.03E+04	7.10E+02
8.33E-01	1.00E+00	2.00E+01	3.07E+04	3.09E+04	7.10E+02
1.00E+00	1.00E+00	2.40E+01	3.08E+04	3.09E+04	7.10E+02
1.17E+00	1.00E+00	2.80E+01	3.08E+04	3.13E+04	7.08E+02
1.33E+00	1.00E+00	3.20E+01	3.09E+04	3.13E+04	7.08E+02
1.50E+00	1.00E+00	3.60E+01	3.08E+04	3.07E+04	7.09E+02
1.67E+00	1.00E+00	4.00E+01	3.08E+04	3.10E+04	7.08E+02
1.83E+00	1.00E+00	4.40E+01	3.06E+04	3.07E+04	7.08E+02
2.00E+00	1.00E+00	4.80E+01	3.06E+04	3.06E+04	7.07E+02

Descriptive Model



Process Optimization

This work:

- Data created using process simulation
- Imposed variability

- P: 5 primary control setpts, catalyst age
- R: production rate, profit, catalyst age rate

- Feature engineering (3rd deg poly)
- MATLAB fitlm

- Brute force search
- Process simulation

and … what to do when the data collection is larger than machine memory? ← the new part

Regression Analysis

In this work the modeling of data is via conventional multivariate regression.

n : # predictors
m : # examples
y : responses

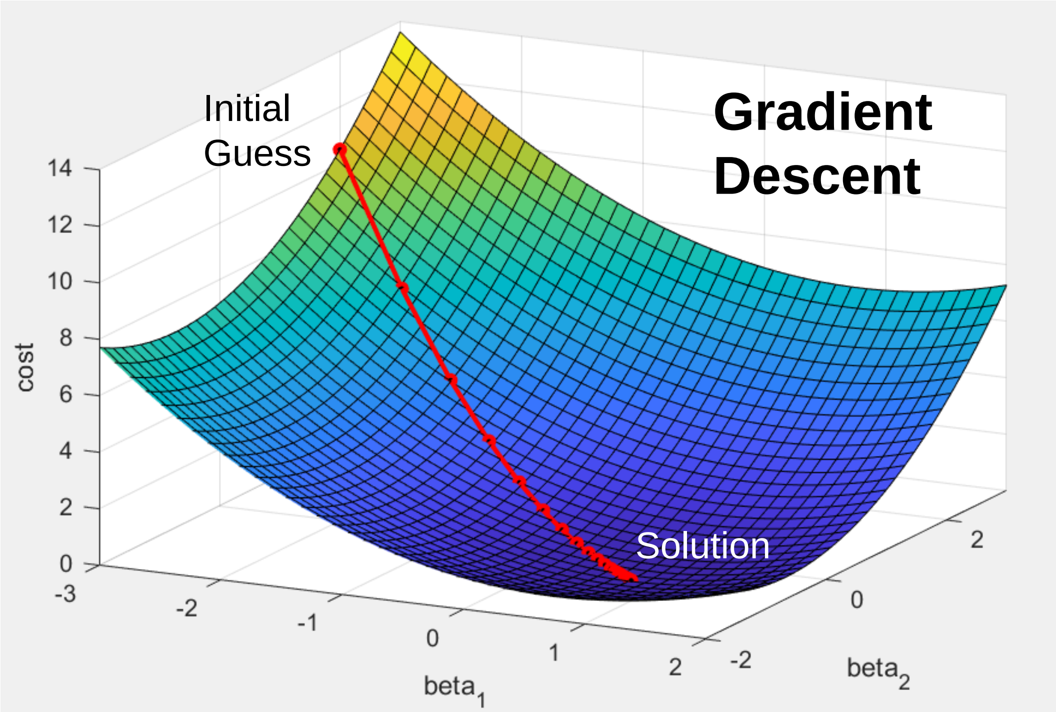
Data matrix: $X_{ij} = [x_j^{(i)}]$

Estimator: $$\hat{y}(x,\beta) = \sum_{k=1}^{n} \beta_k \cdot x_k$$

Cost function: $$J(x,\beta) = \sum_{i=1}^{m} (\hat{y}(x,\beta) - y_i)^2$$

Normal Equations: $$X^T X \beta = A^T y$$

Alternatively, can use an iterative method:



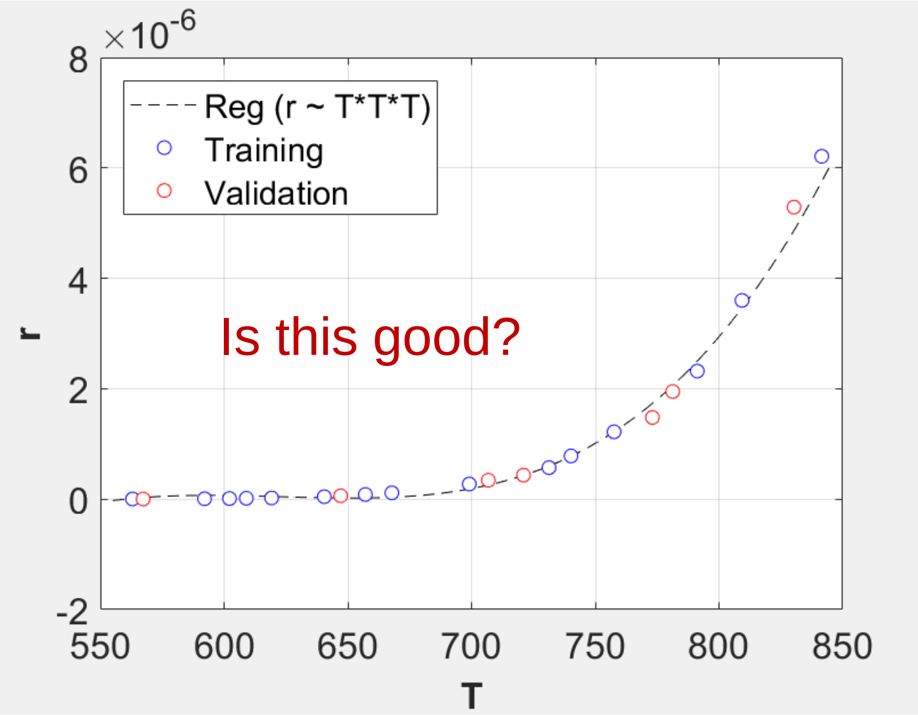
There are many packaged codes for regression, e.g. the *fitlm* function in MATLAB (used here).

Regression Example

Reaction rate data – feature engineering is applied to achieve a reasonable fit.

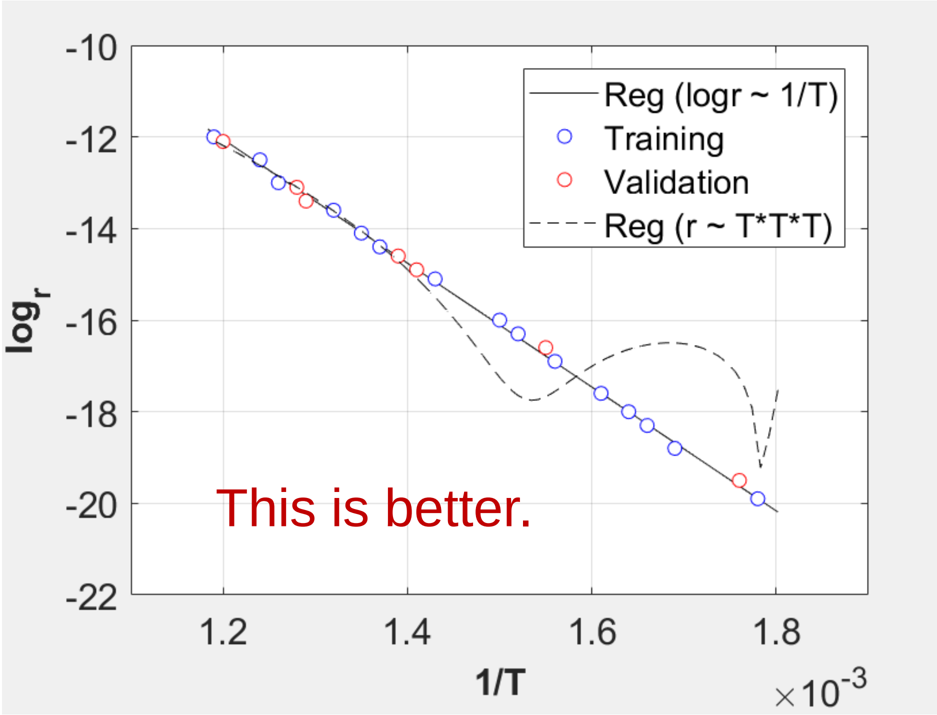
$$\hat{y}(T,\beta) = \beta_0 + \beta_1 \cdot T + \beta_2 \cdot T^2 + \beta_3 \cdot T^3$$

$$\hat{w}(x,\lambda) = \lambda_0 + \lambda_1 \cdot x$$



$$w = log(y)$$

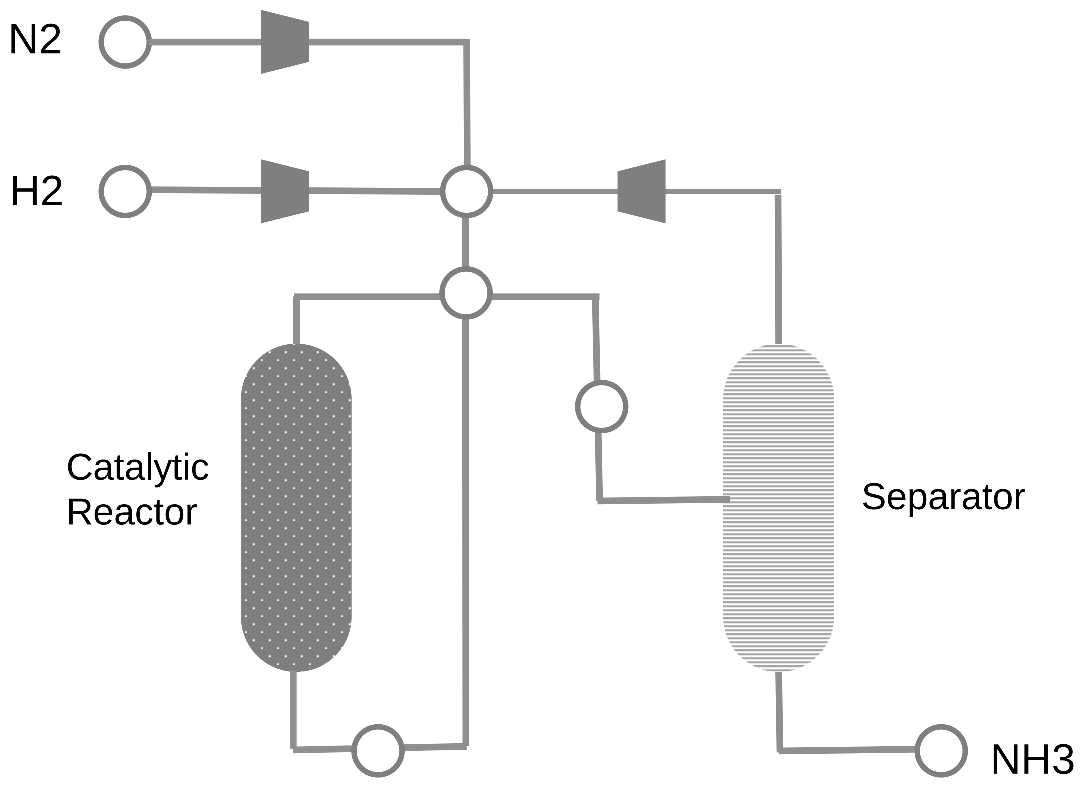
$$x = \frac{1}{T}$$



Modeling informed by understanding outperforms a blind approach.

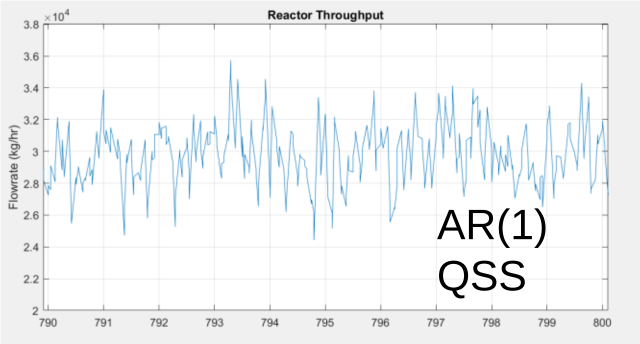
Case Study – Process Plant

The case study consists of an ammonia production plant with 5 control setpoints.



Process Setpoint	Symbol	Nominal Value
Input flowrate	w_{inp}	30,000 kg/hr
Input temp	T_{inp}	710 K
Sys pressure	P_{sys}	20 MPa
Separator temp	T_{sep}	257 K
H/N ratio	η	3.0

Imposed Variability:



A process simulation code was written to generate plant data.

Process Data Record

The process record consist of setpoints and other variables – recorded every 30s, for 6 years.

One year of data:

287 columns

telap	online	catlot	winp	tinp		profit	fkf_avg	cm_lcc	costs_lcc	profit_lcc
3.4722222E-04	1.0000000E+00	1.0000000E+00	1.0198234E+00	9.9893354E-01		8.7907014E+00	9.4039952E-01	1.1551346E+00	1.2694504E+01	7.6355668E+00
6.9444444E-04	1.0000000E+00	1.0000000E+00	1.0325181E+00	9.9854434E-01		7.7449733E+00	9.4039925E-01	1.0761656E+00	1.3035123E+01	6.6688077E+00
1.0416667E-03	1.0000000E+00	1.0000000E+00	1.0335325E+00	9.9869734E-01		8.5996233E+00	9.4039898E-01	1.0995458E+00	1.2625944E+01	7.5000774E+00
1.3888889E-03	1.0000000E+00	1.0000000E+00	1.0115028E+00	9.9794601E-01		9.3880535E+00	9.4039870E-01	1.1367564E+00	1.2018200E+01	8.2512971E+00
1.7361111E-03	1.0000000E+00	1.0000000E+00	1.0088727E+00	9.9702911E-01		8.8154230E+00	9.4039842E-01	1.1220836E+00	1.2473741E+01	7.6933394E+00
3.6499861E+02	1.0000000E+00	1.0000000E+00	1.0549119E+00	1.0053685E+00		9.2109537E+00	9.2756658E-01	1.4003094E+00	1.3631070E+01	7.8106442E+00
3.6499896E+02	1.0000000E+00	1.0000000E+00	1.0424903E+00	1.0047468E+00		9.9245822E+00	9.2756623E-01	1.4358232E+00	1.3084357E+01	8.4887590E+00
3.6499931E+02	1.0000000E+00	1.0000000E+00	1.0123936E+00	1.0032611E+00		1.0401597E+01	9.2756587E-01	1.4628611E+00	1.2382733E+01	8.9387362E+00
3.6499965E+02	1.0000000E+00	1.0000000E+00	1.0253715E+00	1.0033272E+00		1.0086297E+01	9.2756549E-01	1.5404428E+00	1.3470816E+01	8.5458542E+00
3.6500000E+02	1.0000000E+00	1.0000000E+00	1.0066723E+00	1.0030778E+00		1.0118471E+01	9.2756512E-01	1.5000454E+00	1.2785329E+01	8.6184255E+00

1,051,200 rows

(2x60x24x365)

4.7 GB

The size of the total dataset = ~28 GB, more than 2x my laptop RAM (12 GB).

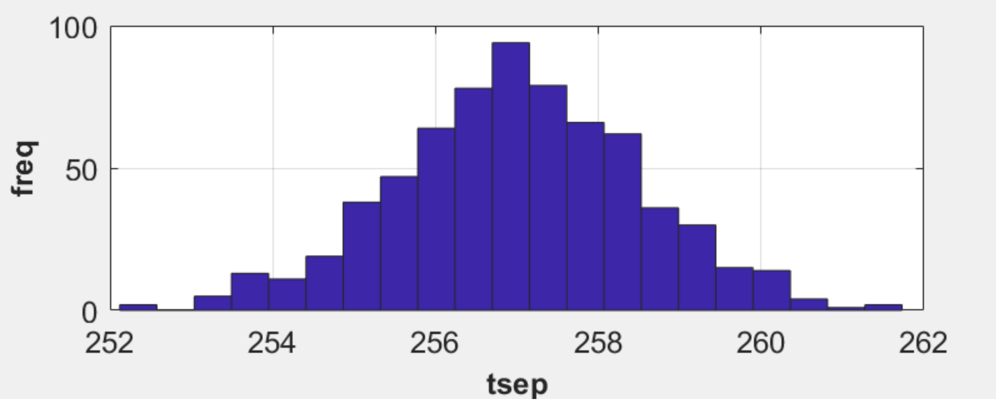
Process Variability

Controlled variables were modeled as AR(1) processes.

Ranges for controlled variables:

Parameter	Units	Minimum	Mean	Maximum
winp	kgmol/h	21360	30000	39000
tinp	Kelvin	667	710	758
psys	atm	183	200	221
tsep	Kelvin	250	257	263
hnrat	---	2.72	3.00	3.30

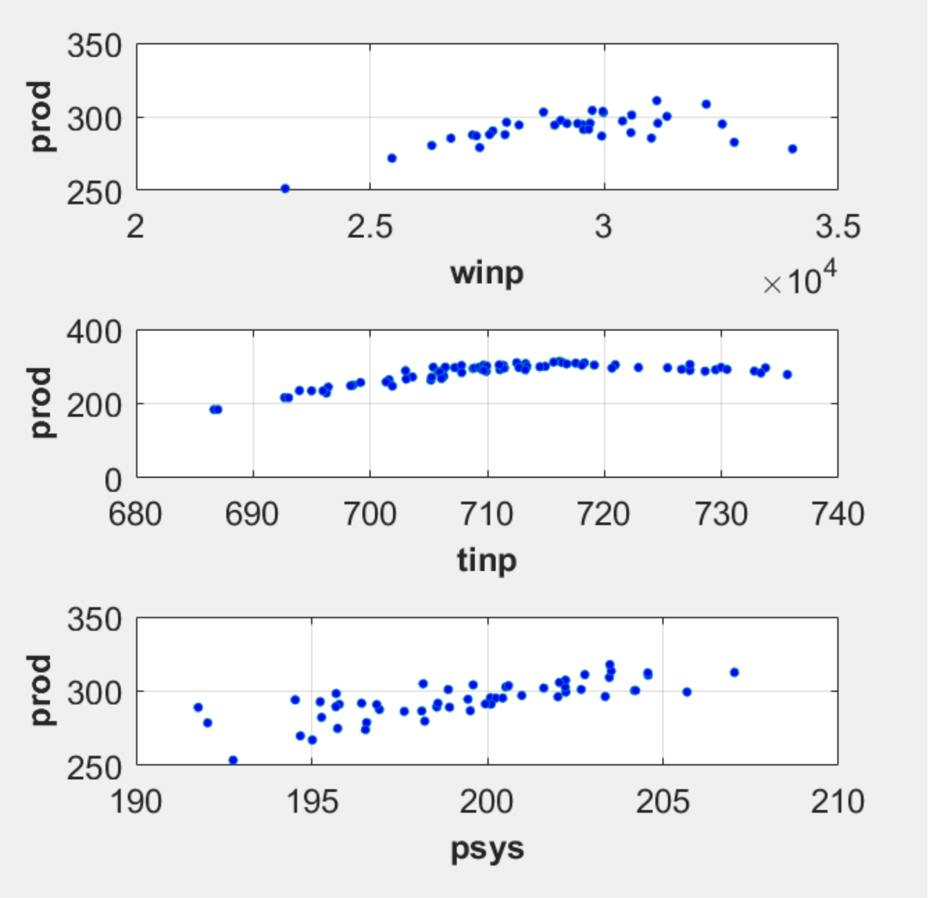
These are later used to define the optimization search region.

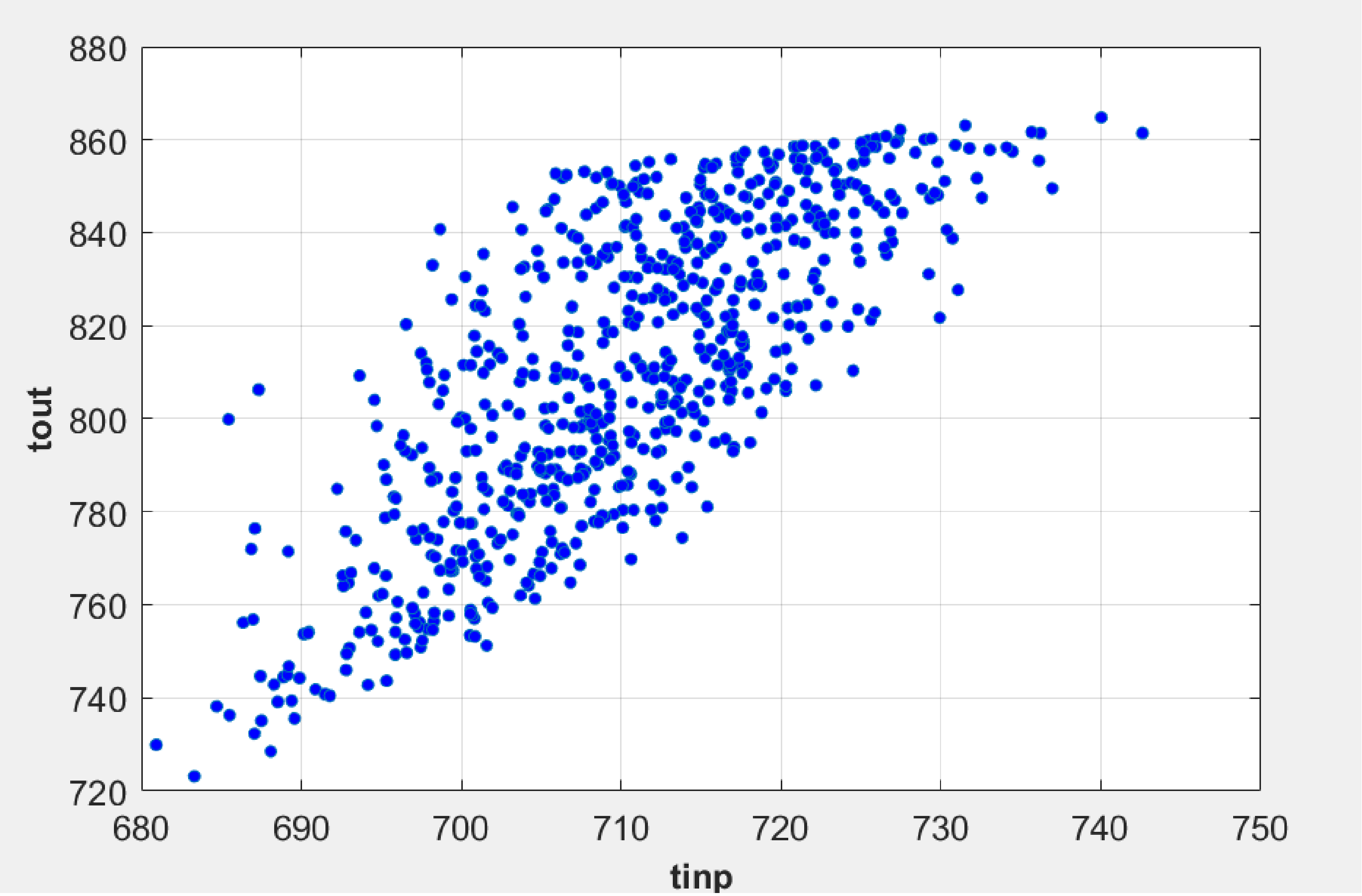


If the process conditions didn't vary, there wouldn't be anything to model!

Setpoints and Select Responses

A sensitivity scatter around the mean confirms some intuitions.

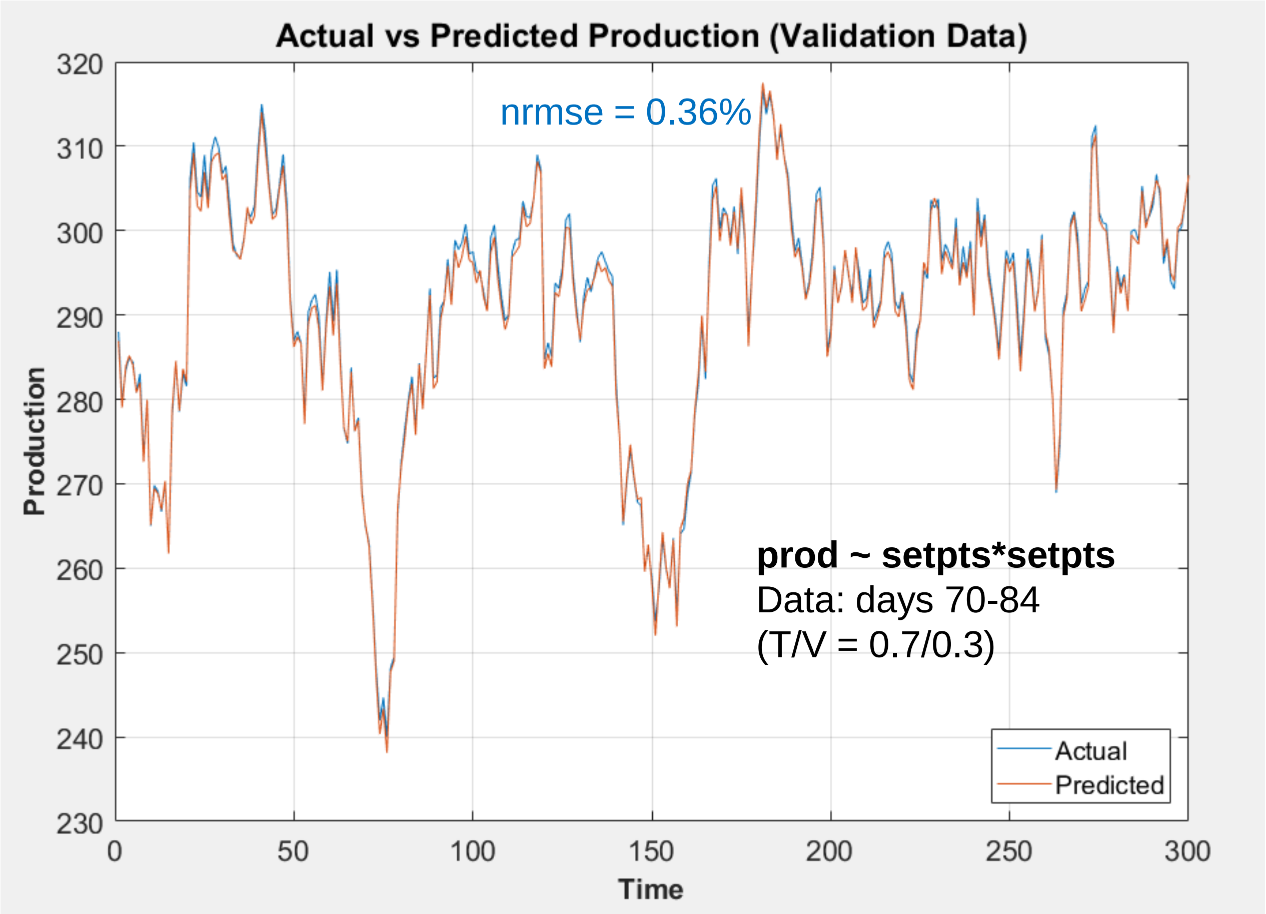




… but it is difficult to infer interdependencies due to the high number of dimensions.

Modeling Production Rate

Regression on production rate was performed using the plant setpoints as predictors.



For comparison, a simpler model:
prod ~ setpts was also trained.

Predictors	Total # predictors	NRMSE (Valid'n)	R^2 (Train)
setpts	5	2.17%	0.837
setpts*setpts	20	0.36%	0.996

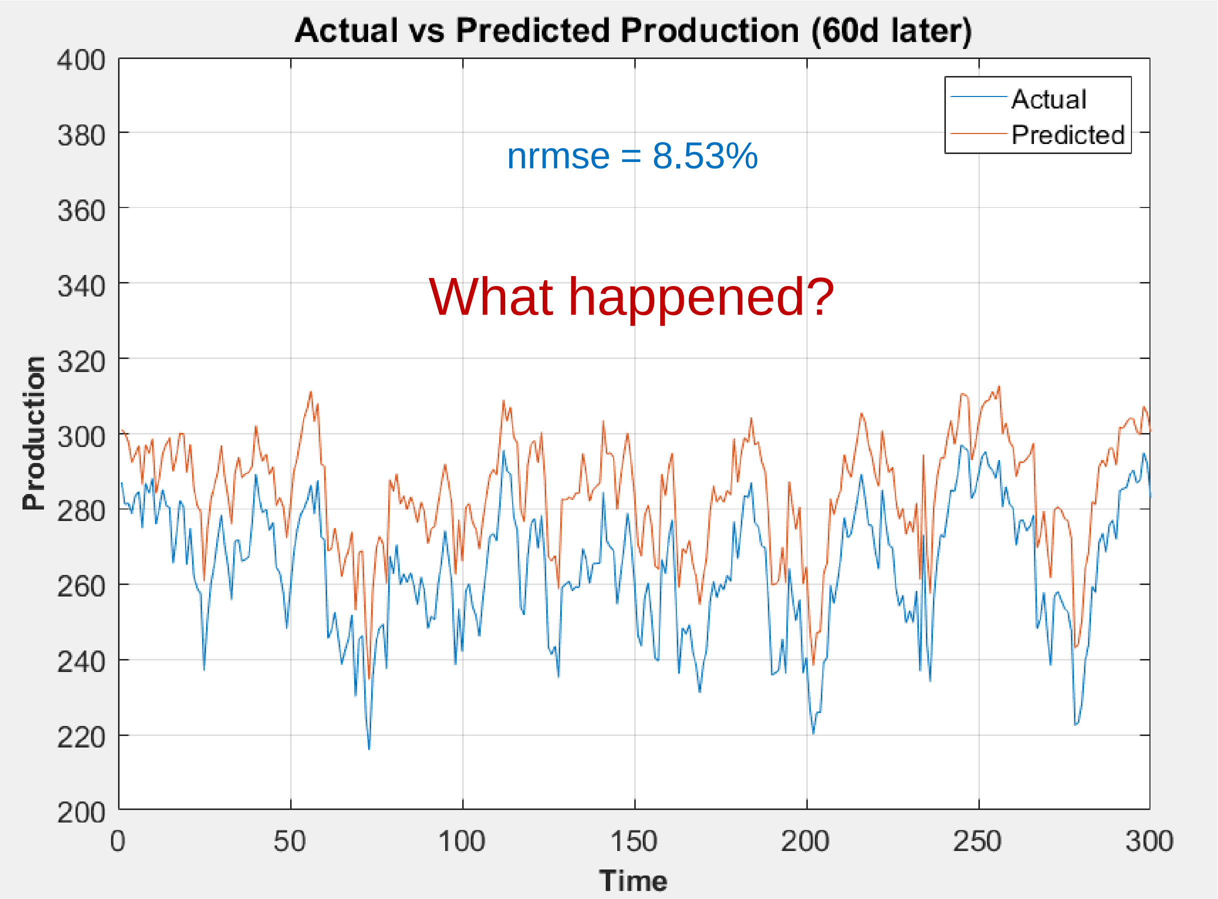
For this limited duration (300 hrs) of data (180 MB), a highly predictive model can be built.

Model Generalization

When applied to data from a different time interval, the fit is not as good.

60
days
later

same
model

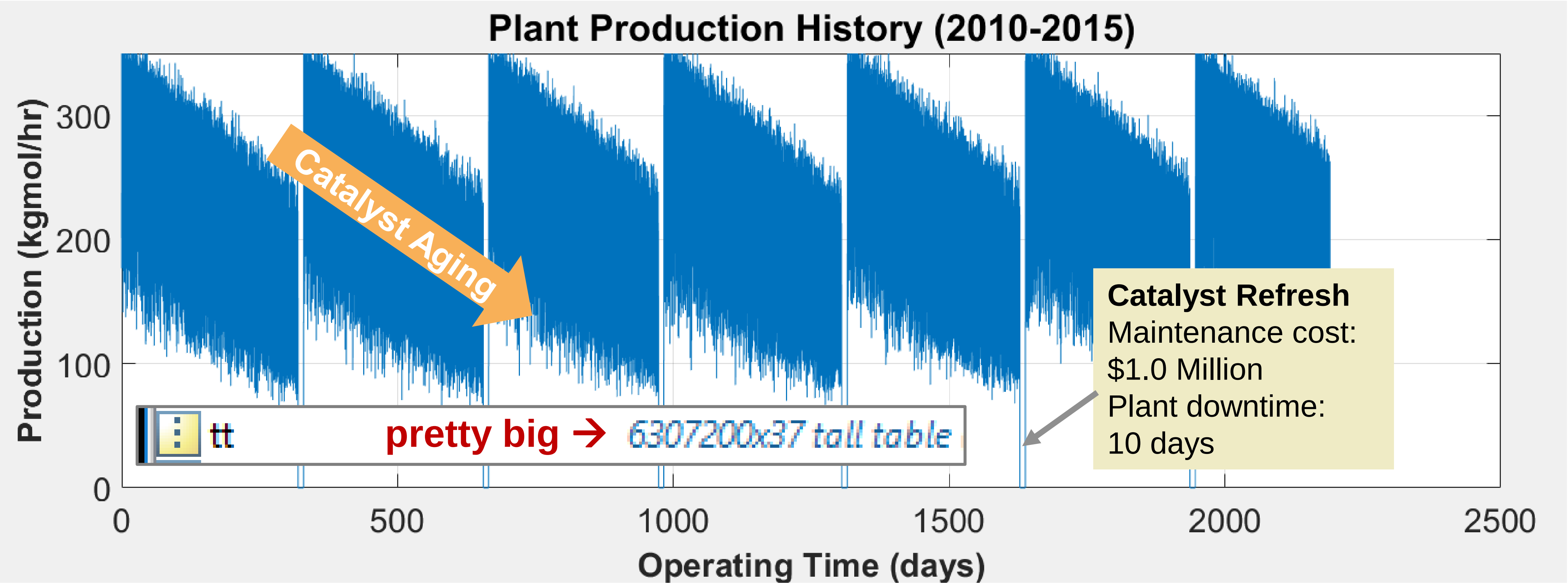


This model does not generalize well

→ Need to take a closer look at the trend of production with time.

Production History

To ensure adequate generalization, the training set must represent the entire data collection.



Production depends on catalyst age state → need to add a new predictor.

Forward Analysis Plan

Take stock of where we are, and where we're going …

- Dataset is fully generated (6 files, 28 GB total)
- Predictors have been defined: setpoints + catalyst age parameter
- Regression of production on for limited data (180 MB) completed with good results
- To capture catalyst aging effects, all of the data (28 GB) will be used → need big data tools
- Instead of modeling production, look at the bottom line: create a regression model of plant operating **profit**
- Once profit model is built, use it to compute the profit-maximizing setpoints as functions of catalyst age (remaining activity).
- Use those operating schedules in a process simulation, and compute the estimated profit – is it better than the status quo?

Is it possible to create a simple regression model of profit?

Big Data Computing Infrastructure

Special strategies are needed for handling data sets larger than machine memory.

Dataset mgmt:

Distributed computing:

- Windows users need to set up a virtual machine supporting LINUX
- Need IT help to set up a Hadoop cluster (on premise or cloud), and install Spark
- In this work I used a test cluster with 11 nodes

```
%% Specify the data files location
filename = 'hist_201*.csv';  % 6 files:  2010, ..., 2015
hdfspath = 'hdfs://hadoop01glmxt64:54255/datasets/plant_model/';
fileloc  = strcat(hdfspath,filename);
```

```
%% Create the datastore
ds = datastore(filelocs,'SelectedVariableNames', varnames);

%% Create a tall table
tt = tall(ds);

%% Remove selected data ranges
idx        = (tt.online == 0);
tt(idx,:) = [];
```

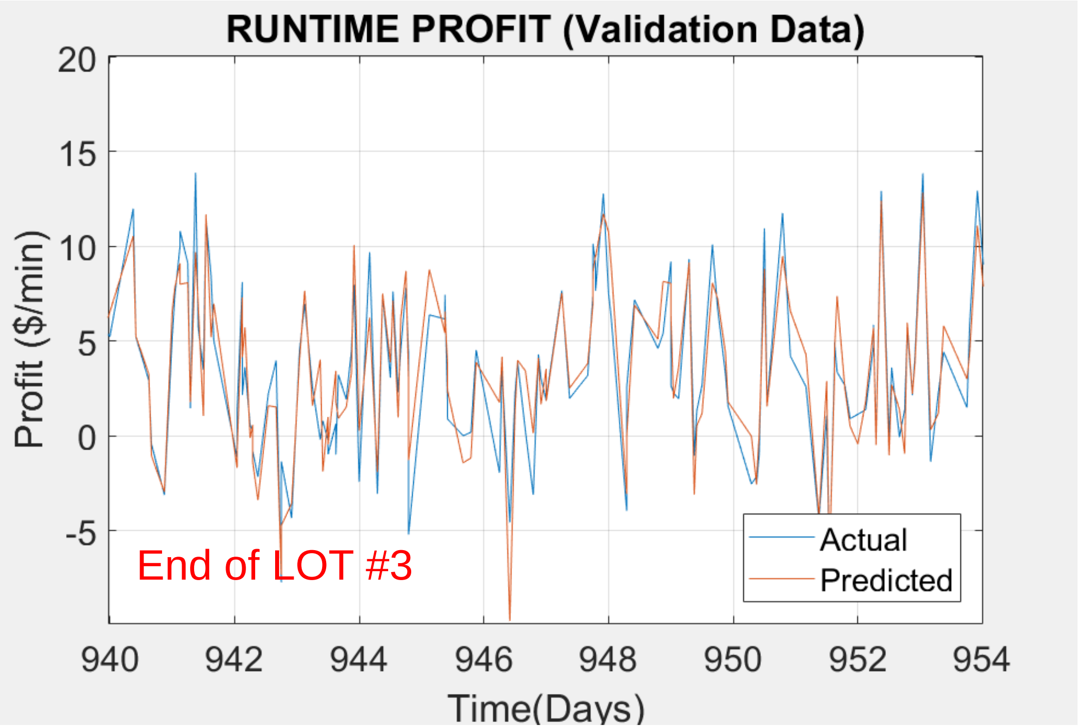
```
%% Build the model
model = fitlm(ttTrain,modelform);

%% Validate Model
yPred_valid  = predict(model,ttValid);
resid_valid  = yPred_valid - ttValid.profit;
rmse_valid   = gather(sqrt(mean((resid_valid).^2)));
```

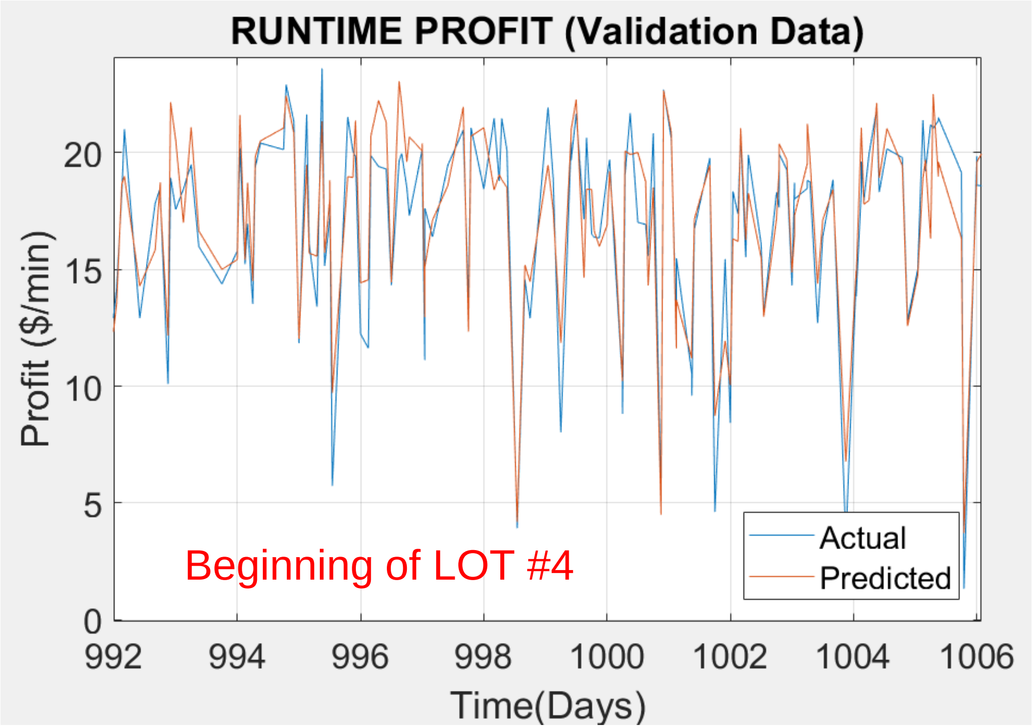
In this work, MATLAB data management tools and a Spark-enabled Hadoop cluster were used.

Modeling Profit

Operating profit is an extremely complicated function of plant setpoints, and other parameters.

prod ~ setpts+ * setpts+ * setpts+ setpts+ = (setpts + cat_age)



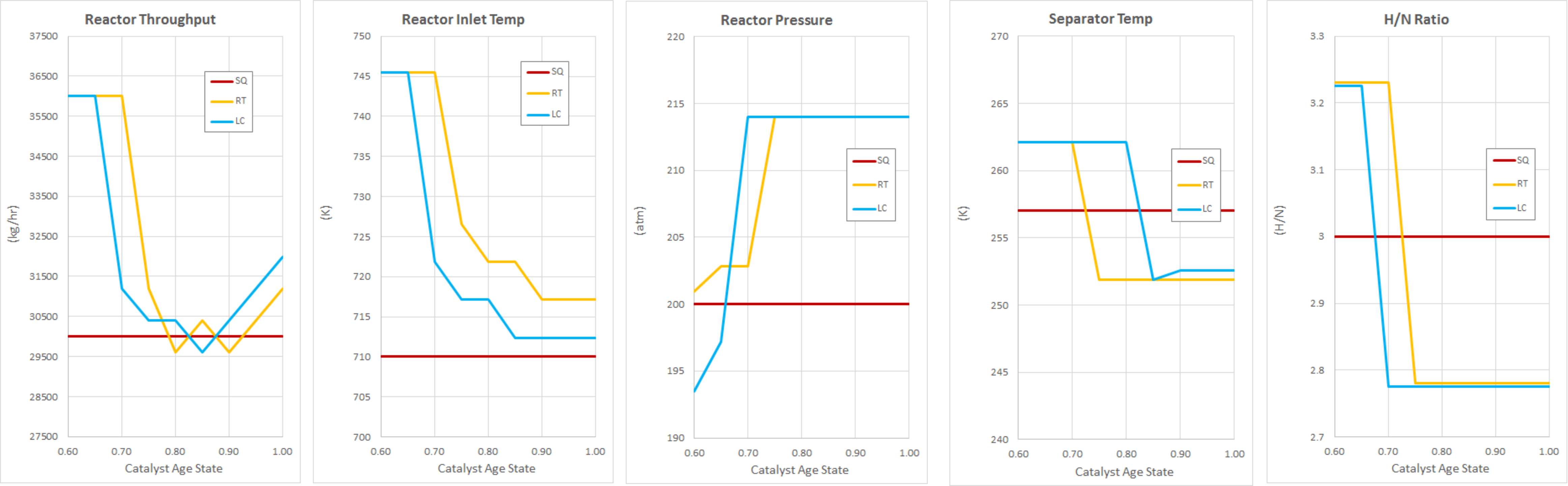


The regression model for profit can now be used to explore alternative operating strategies.

Operating Schedules

Schedules are computed by maximizing profit estimates over a range of catalyst age states.

Realtime Profit	Maximize runtime profit without accounting how conditions impact catalyst aging
Life Cycle Profit	Maximize profit including imputed catalyst deactivation & maintenance costs

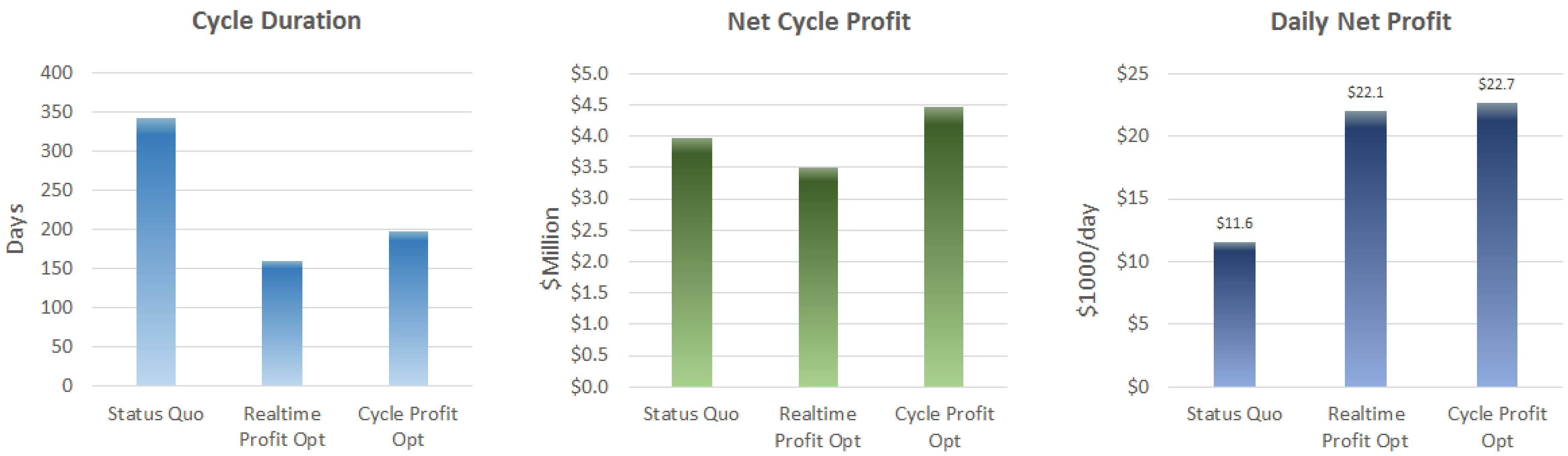


Plant simulation can now be used to estimate the financial performance of each strategy.

Operating Strategy Comparison

During simulation, an additional regression model for catalyst aging rate is used.

The simulation stops after a complete cycle → when the catalyst activity reaches 60% of its initial value.



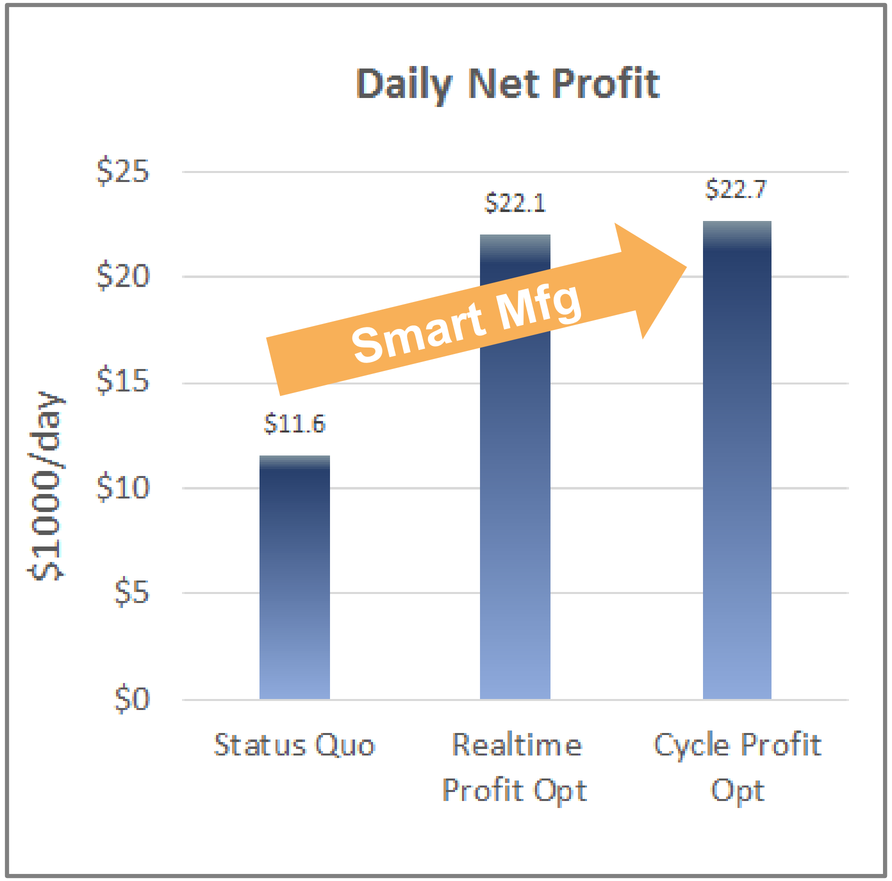
→ Analysis reveals opportunities for increasing daily profits by **1.9X** !

The results indicate that the status quo operating procedure is significantly non-optimal.

Summary

Regression modeling of plant data led to identification of more profitable operating strategies.

- Regression methods very mature
- Large datasets require newer tools
- Modeling provides valuable insight



Acknowledgments
Jason Ross, Lucio Cetto,Vick Chellappa-Smith, Heather Gorr, Chetan Rawal

Thank you for your attention.