From Data Science to Data Stories

Katya Vladislavleva, PhD, PDEng
CEO DataStories (Evolved Analytics Europe)

OUR TECHNOLOGY IS SHAPED BY THE REAL WORLD

Different industries are solving the same analytics problems

- Energy
- Advanced manufacturing
- Materials
- Finance
- (Digital) Health
- Consumer products
- Business operations

**If you can measure it,
you can understand it.**

**If you can understand it,
you can alter it.**

Katherine Neville

datastories

90% OF THE BUDGET GOES TO DATA COLLECTION

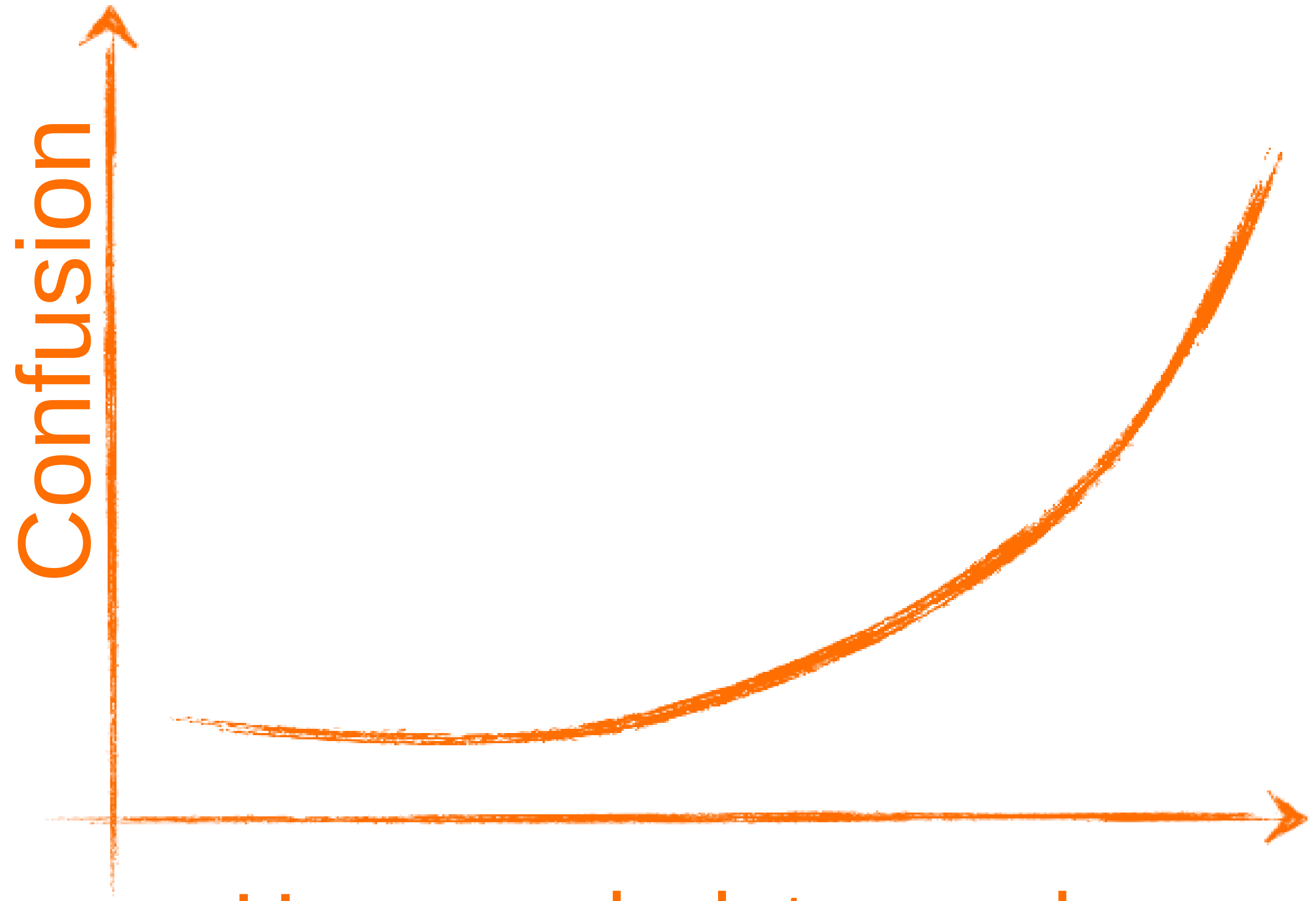
BANG YOUR HEAD
HERE

95% OF THE DATA

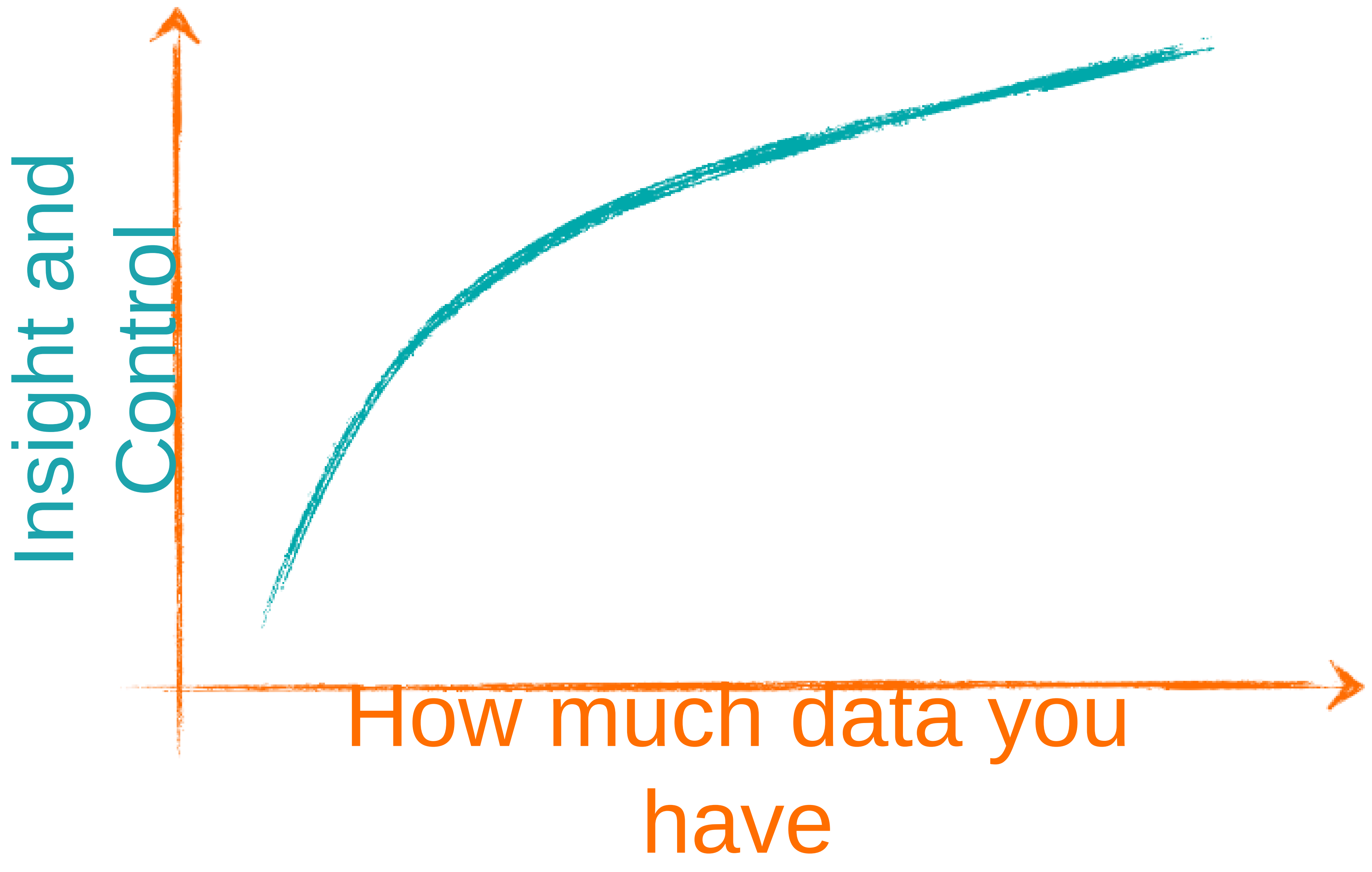
IS NOT UTILIZED

Source: IDC 2014

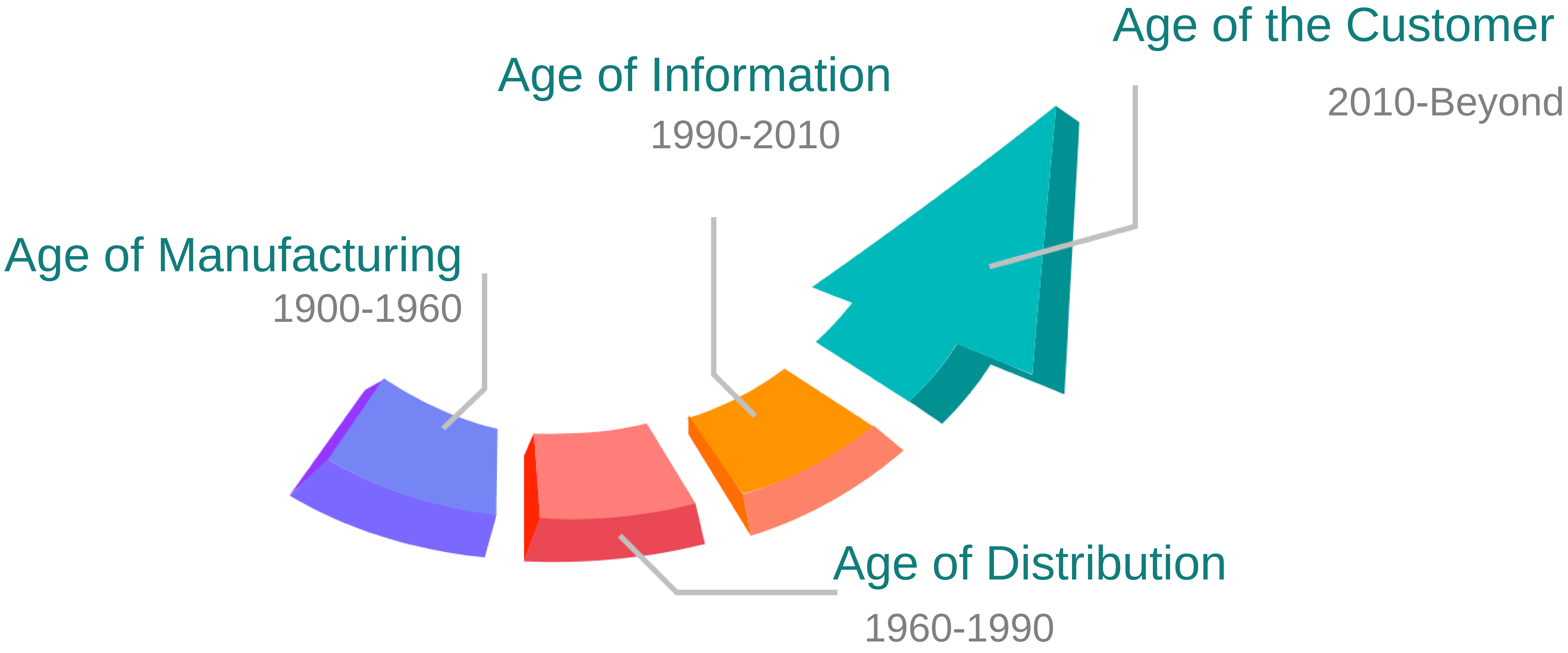
More data does not always imply more information



The real goal is understanding and control



The Age of the Customer is here



Source: « Competitive Strategy in The Age of the Customer » Forrester Report, 06/2011

Gartner

"...by 2016, 89% of companies expect to compete mostly on the basis of customer experience."

accenture

"CIOs attach more importance to developing consistent and relevant multi-channel experiences"

FORRESTER

"...densely collaborative space between the CIO's staff and the CMO's staff."

McKinsey&Company

"...Nearly half of respondents say their CEOs personally sponsor digital initiatives."

Good Advice from Forrester

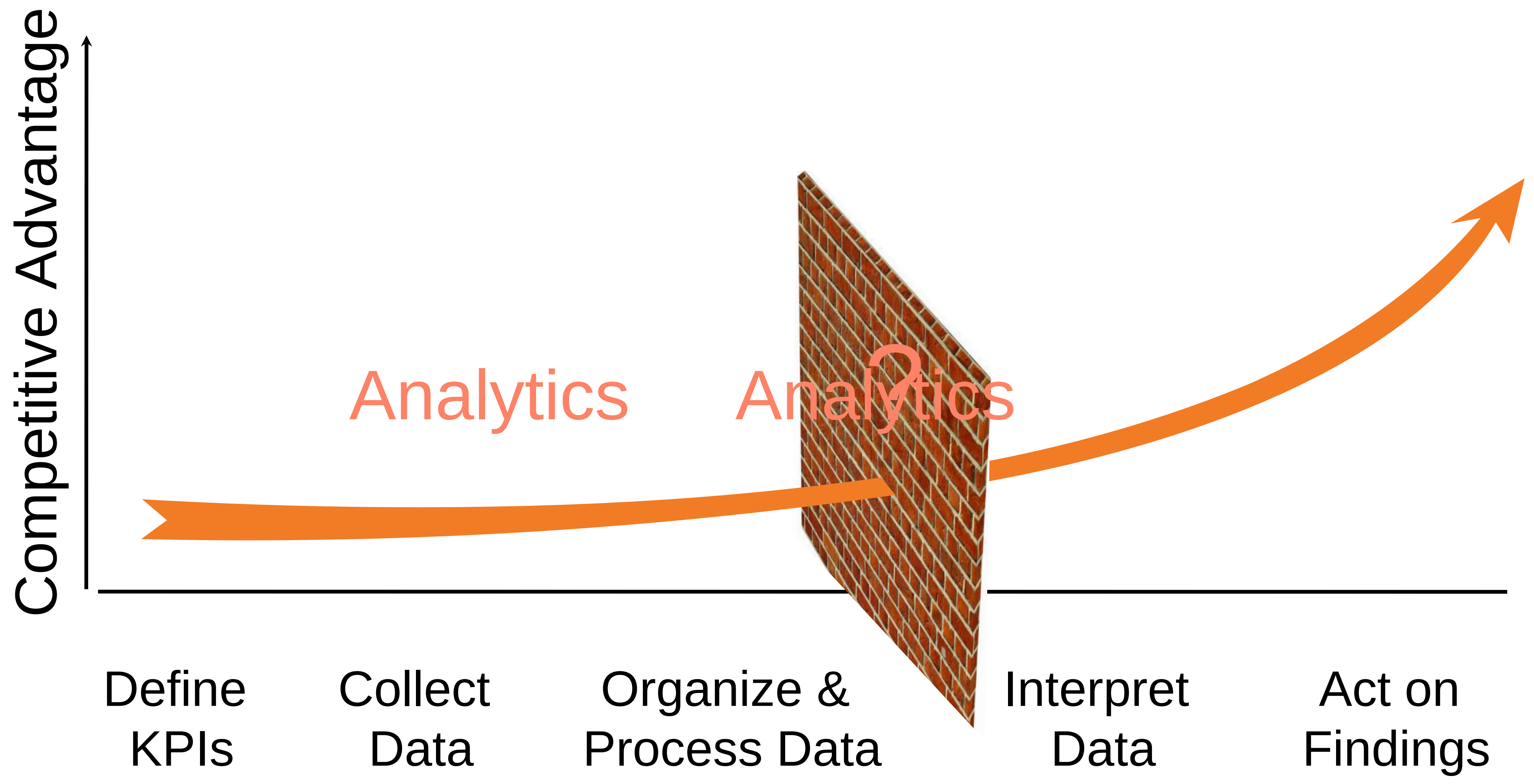
1. Attain Maximum Customer Intelligence with Data Analytics
2. Create customer experience excellence
3. Be there, be digital
4. Embrace the mobile

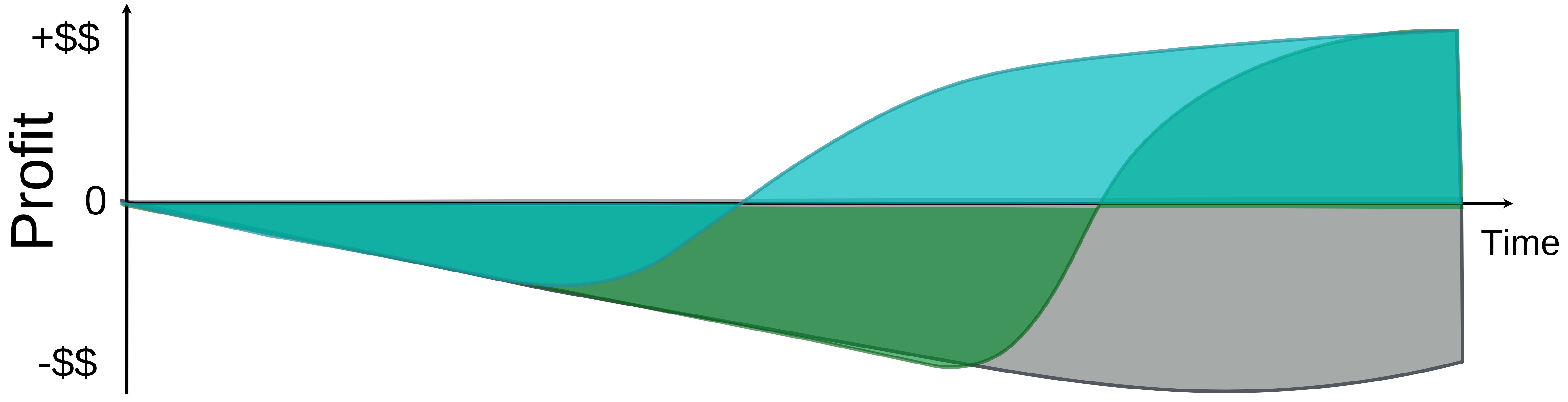
Source: « The Four Imperatives of Winning in The Age of the Customer » Forrester Report, 10/2015

MISMATCH BETWEEN THE VOLUME OF DATA AND OUR CAPACITY TO ANALYSE IT GROWS ALARMINGLY FAST

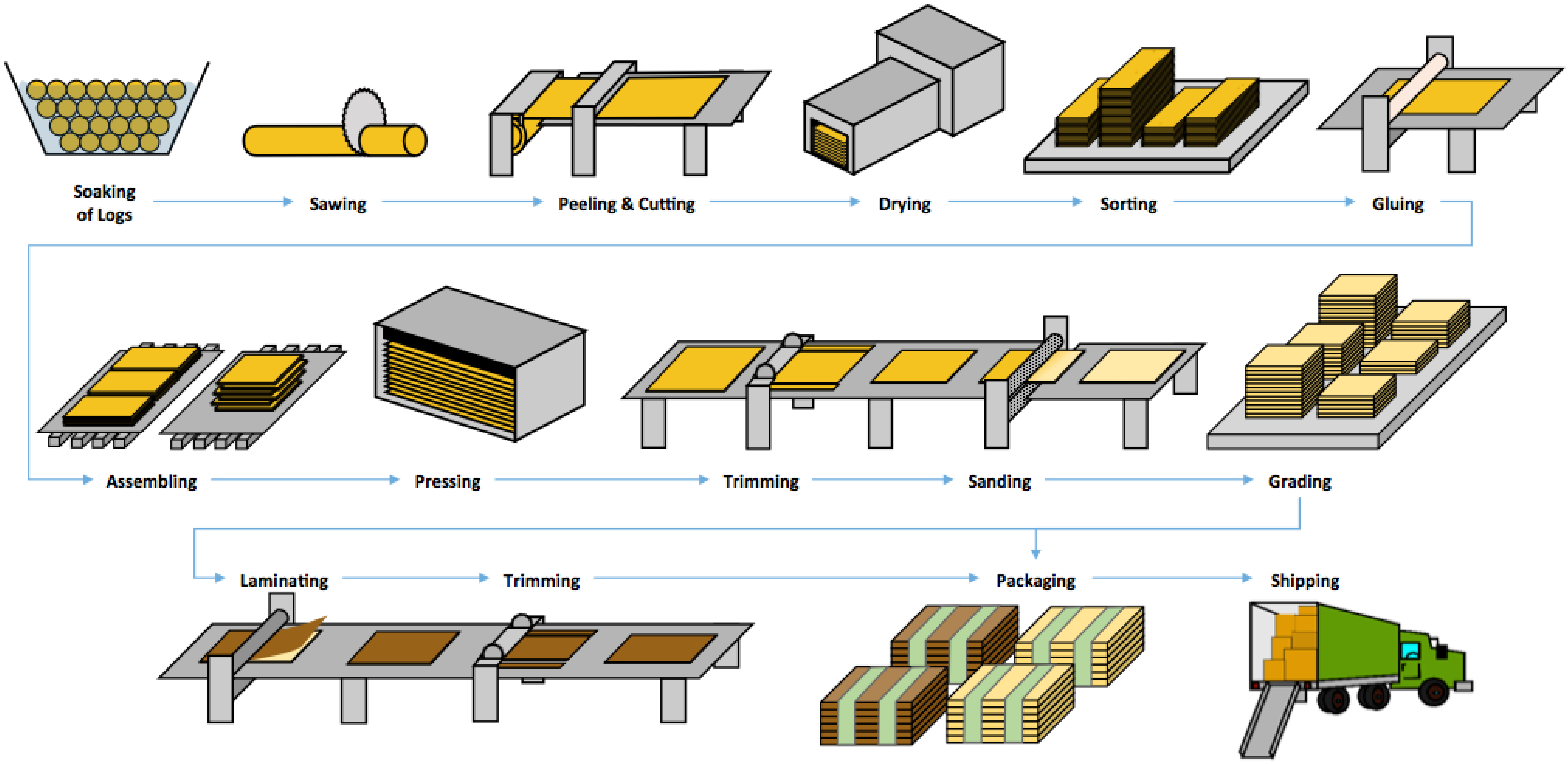
By 2018, the US only will face a shortage of up to 190,000 data scientists as well as 1.5 Million managers and analysts with enough proficiency in statistics to use big data effectively.

McKinsey Global Institute 2013

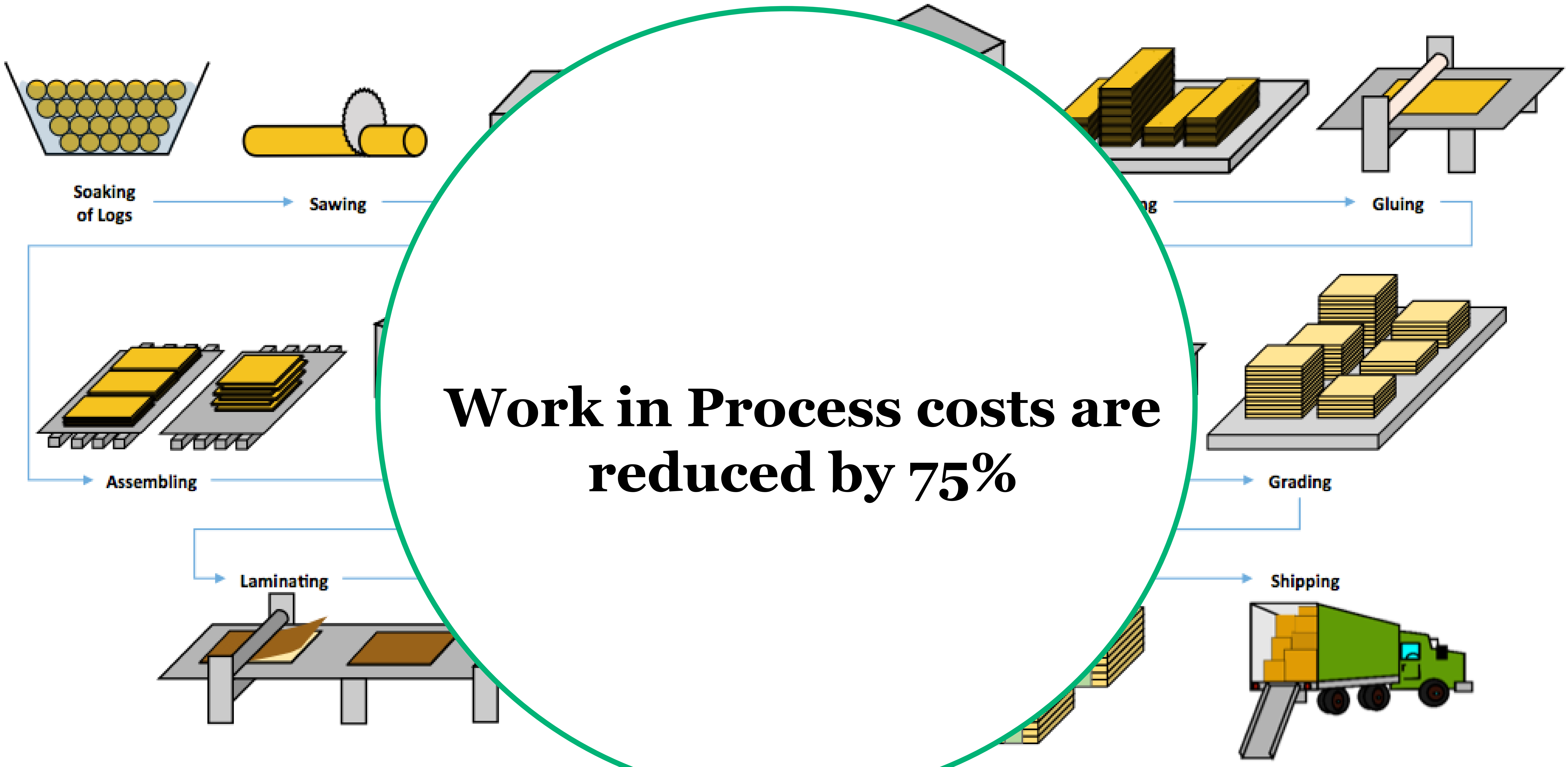

Competitive Advantage
Analytics
Analytics
?
Define KPIs
Collect Data
Organize & Process Data
Interpret Data
Act on Findings


Profit
+$$
0
-$$
Time

Change the Mind set

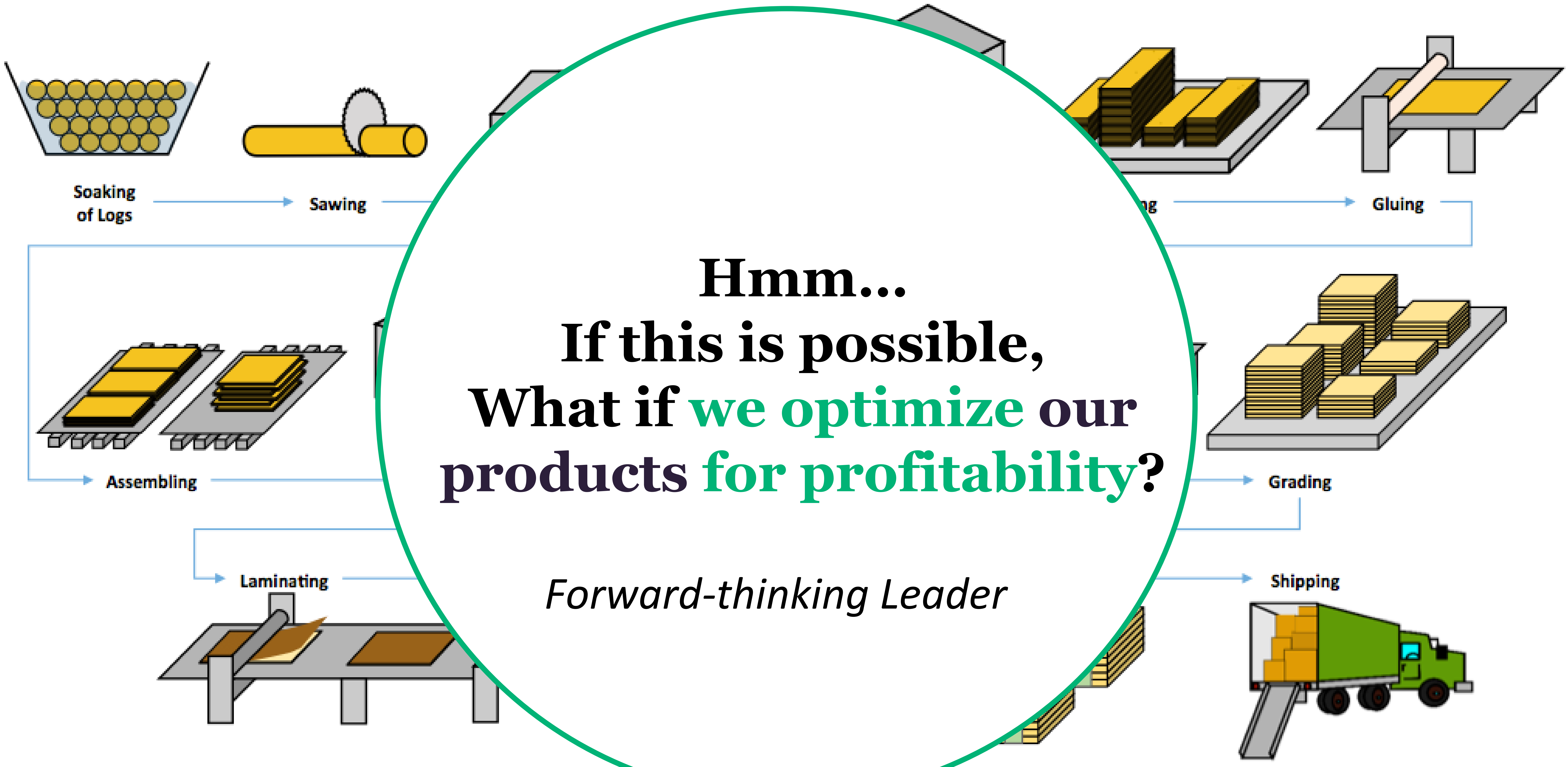


Change the Mind set



Work in Process costs are reduced by 75%

Change the Mind set



Hmm…
If this is possible,
What if we optimize our products for profitability?

Forward-thinking Leader

There are different kinds of big data

Big Data

Big Hype

Medium & Small Data

Big Mess

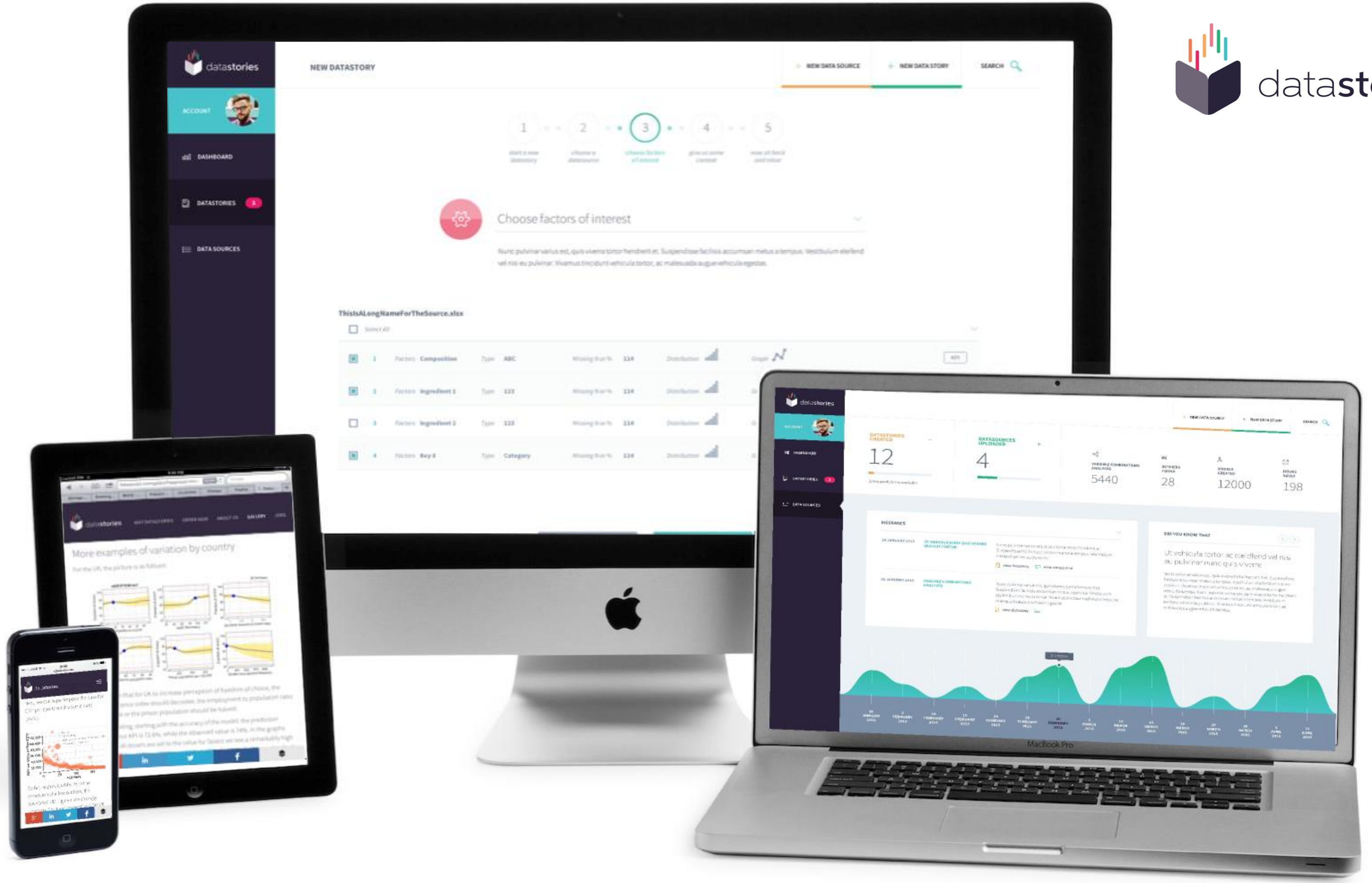

TOP 8
TRENDS FOR 2016
BIG DATA
http://www.tableau.com/sites/default/files/media/top8bigdatatrends2016_final_2.pdf
tableau

The number of options for *preparing* end users to discover all forms of data grows.

Self-service data preparation tools are exploding in popularity. This is in part due to the shift toward business-user-generated data discovery tools such as Tableau that reduce time to analyze data. Business users also want to be able to reduce the time and complexity of preparing data for analysis, something that is especially important in the world of big data when dealing with a variety of data types and formats. We've seen a host of innovation in this space from companies focused on end user data preparation for Big Data such as Alteryx, Trifacta, Paxata and Lavastorm while even seeing long established ETL leaders such as Informatica with their Rev product make heavy investments here.

Additional Reading:
Alteryx, Trifacta, Paxata, Lavastorm, Informatica

datastories



data_extended.csv

A1 deltaPressure

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	deltaPressure	OverheadPr	OverheadTe	refluxFlow	feedTemper	feedFlow1	feedFlow2	temperature	temperature	temperature	temperature	temperature	temperature	temperature	temperature	temperature	temperature	temperature	temperatu
2	4.66	383	-152	135	32.3	323	1070	28.4	30.4	30	31	39.1	80.2	59.2	60.6	60.4	60.6	144	16
3	4.64	383	-152	135	32.3	322	1030	28.1	30.2	29.8	30.8	38.6	80.4	59.1	60.6	60.3	60.6	142	16
4	4.66	383	-152	135	32.4	324	971	28	30.2	29.8	30.8	38.7	80.4	59.1	60.6	60.2	60.6	139	16
5	4.65	383	-152	135	32.2	323	957	28.2	30.4	30	30.9	38.9	80.2	59.2	60.7	60.4	60.7	139	16
6	4.67	383	-152	135	32.2	324	1060	28.2	30.3	29.9	30.8	38.8	79.8	59.1	60.6	60.2	60.6	137	16
7	4.64	383	-152	135	32	323	1180	28	30.1	29.7	30.6	38.6	80.2	58.9	60.4	60	60.4	136	16
8	4.68	383	-152	135	32.1	325	1090	27.9	30.1	29.7	30.6	38.7	80.2	58.9	60.3	60	60.3	137	16
9	4.71	383	-152	135	32.1	326	1080	27.8	30	29.6	30.5	38.6	80.4	58.7	60.2	59.8	60.2	137	16
10	4.71	383	-152	135	32	326	1000	28	30.1	29.8	30.7	38.9	80.4	58.8	60.2	59.9	60.2	143	16
11	4.7	383	-152	135	31.9	325	951	27.9	30	29.6	30.5	38.6	80.2	58.7	60.2	59.9	60.2	144	16
12	4.68	383	-152	135	31.9	327	944	27.7	29.9	29.5	30.4	38.2	80.1	58.8	60.2	59.9	60.2	141	16
13	4.7	383	-152	135	31.9	327	921	27.7	29.9	29.5	30.4	38.3	79.4	58.7	60.2	59.8	60.2	137	16
14	4.7	383	-152	135	32	325	870	27.9	30	29.7	30.6	38.6	79.2	58.8	60.3	59.9	60.2	139	16
15	4.71	383	-152	135	31.9	326	969	27.8	29.9	29.5	30.4	38.4	79.3	58.6	60.1	59.7	60.1	140	16
16	4.7	383	-152	135	31.9	325	1070	27.8	29.9	29.6	30.5	38.6	79.4	58.6	60	59.7	60	140	16
17	4.69	383	-152	135	32	325	1110	27.8	30	29.6	30.5	38.5	79.3	58.7	60.2	59.8	60.2	140	16
18	4.7	383	-152	135	31.9	327	1040	27.8	30	29.6	30.5	38.6	79.3	58.8	60.2	59.8	60.2	139	16
19	4.7	383	-152	135	31.8	325	1050	27.8	29.9	29.6	30.5	38.5	79.3	58.7	60.1	59.8	60.1	140	16
20	4.7	383	-152	135	32	324	924	27.8	30	29.6	30.5	38.6	79.4	58.7	60.1	59.8	60.2	141	16
21	4.72	383	-152	135	31.9	326	930	27.8	29.9	29.6	30.5	38.4	79.4	58.7	60.2	59.8	60.2	141	16
22	4.69	383	-152	135	31.9	325	923	27.8	29.9	29.5	30.4	38.4	79.4	58.7	60.2	59.8	60.1	140	16
23	4.71	383	-152	135	31.9	326	1020	27.7	29.9	29.5	30.4	38.4	79.3	58.6	60.1	59.8	60.1	140	16
24	4.72	383	-152	135	31.9	327	1110	27.8	29.9	29.6	30.5	38.6	79.2	58.7	60.1	59.7	60.1	140	16
25	4.7	383	-152	135	31.9	327	1150	27.8	29.9	29.5	30.4	38.5	79.1	58.6	60.1	59.7	60.1	141	16
26	4.71	383	-152	135	31.9	326	1050	27.8	29.9	29.5	30.4	38.4	78.9	58.6	60.1	59.7	60.1	140	16
27	4.72	383	-152	135	31.9	326	982	27.8	29.9	29.6	30.5	38.5	78.5	58.7	60.2	59.7	60.1	140	16
28	4.72	383	-152	135	31.8	327	890	27.8	29.9	29.6	30.5	38.4	77.8	58.7	60.2	59.8	60.2	140	16
29	4.74	383	-152	135	31.8	328	909	27.8	29.8	29.5	30.4	38.4	77.4	58.6	60.1	59.7	60.1	140	16
30	4.76	383	-152	135	31.8	325	1090	27.7	29.8	29.5	30.4	38.5	77.2	58.5	59.9	59.5	59.9	140	16
31	4.69	383	-152	135	31.8	325	839	27.8	29.9	29.5	30.5	38.5	77.1	58.7	60.1	59.8	60.1	143	16
32	4.71	383	-152	135	31.8	326	745	27.6	29.8	29.4	30.3	38.1	76.9	58.7	60.1	59.8	60.1	142	16
33	4.77	383	-152	135	31.8	328	895	27.6	29.8	29.4	30.3	38.3	76.7	58.6	60.1	59.7	60.1	139	16
34	4.73	383	-152	135	31.9	326	1050	27.9	30	29.6	30.5	38.7	77	58.7	60.1	59.8	60	142	16
35	4.71	383	-152	135	31.7	327	1130	27.7	29.8	29.4	30.3	38.2	77.2	58.6	60	59.7	60	141	16
36	4.72	383	-153	135	31.8	326	1060	27.4	29.6	29.2	30.3	38.3	77.2	58.6	60.1	59.7	60.1	141	16
37	4.72	383	-152	135	31.9	327	942	27.7	29.8	29.4	30.4	38.5	77.5	58.6	60.1	59.7	60.1	142	16
38	4.71	383	-152	135	31.8	326	768	27.7	29.8	29.4	30.4	38.3	78.8	58.6	60	59.7	60	142	16

datastories

NEW DATASTORY

+ NEW DATA SOURCE

+ NEW DATA STORY

LOGOUT

ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 1

ADMIN

1 choose a datasource

2 start a new datastory

3 choose factors of interest

4 give us some context

now sit back and relax!

data_extended.csv

See all DataSources

Please choose a datasource from the list on the right or type a name in the box above. If you do not have any data sources please upload one using the link below.

Or upload a new one

New DataSource +

If you do not have a datasource, please follow the link to upload a new one.

NEXT PLEASE

datastories

NEW DATASTORY

+ NEW DATA SOURCE

+ NEW DATA STORY

LOGOUT

ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 1

ADMIN

1 choose a datasource

2 start a new datastory

3 choose factors of interest

4 give us some context

now sit back and relax!

Predicting Propylene Output (extended dataset)

Nunc pulvinar varius est, quis viverra tortor hendrerit et. Suspendisse facilisis accumsan metus a tempus. Vestibulum eleifend vel nisi eu pulvinar. Vivamus tincidunt vehicula tortor, ac malesuada augue vehicula egestas.

TAKE ME BACK

NEXT PLEASE

datastories

ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 1

ADMIN

NEW DATASTORY

+ NEW DATA SOURCE

+ NEW DATA STORY

LOGOUT

1 choose a datasource

2 start a new datastory

3 choose factors of interest

4 give us some context

now sit back and relax!

Choose columns we may use and Select your Key Performance Metric

We will analyse your data and build predictive models for the Key Performance Indicator (KPI) you select in the table below. Please, make sure that all columns that we may use as potential predictors are selected in the table below. If there are columns that you do not want to see in the final models (e.g. if they are difficult to measure or control), please, exclude them from the list of options. We will consider all the columns allowed by you and distill a minimal list of necessary and sufficient columns which impact your KPI.

Select your KPI by clicking a KPI button corresponding to the metric of interest. At this point only one KPI at a time is allowed. Let us know at beta@datastories.com if modeling multiple KPIs using the same list of metrics is important for your application. We will make it happen!

TAKE ME BACK

NEXT PLEASE

data_extended.csv

Showing 1 to 25 of 73 data columns

1 2 3

25

datastories

ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 1

ADMIN

	#	Column	Type	Missing			Unique	Values	
☑	15	temperatureTray8	[a,b]	100% non-missing			158	[56.5, 57.5, 57...]	☐
☑	16	temperatureTray9	[a,b]	100% non-missing			169	[57.4, 56.2, 57...]	☐
☑	17	temperatureTray10	[a,b]	100% non-missing			166	[57.9, 57.8, 56...]	☐
☑	18	temperatureTray11	123	100% non-missing			21	[152.0, 148.0, 1..]	☐
☑	19	temperatureTray12	123	100% non-missing			5	[158.0, 160.0, 1..]	☐
☑	20	steamFlow	[a,b]	100% non-missing			158	[38.3, 38.1, 39...]	☐
☑	21	vaporFlow	[a,b]	100% non-missing			96	[181.0, 182.0, 1..]	☐
☑	22	bottomFlow	[a,b]	100% non-missing			352	[112.0, 111.0, 1..]	☐
☑	23	bottomTemperature	123	100% non-missing			24	[150.0, 151.0, 1..]	☐
☑	24	Propylene	[a,b]	100% non-missing			158	[0.27, 0.18, 0.2..]	☑
☑	25	Delta_temperatureTray1_feedTemperature	[a,b]	100% non-missing			162	[-4.5, -5.0, -4...]	☐

Showing 1 to 25 of 73 data columns

‹ 1 2 3 25 ›

datastories

NEW DATASTORY

NEW DATA SOURCE

NEW DATA STORY

LOGOUT

ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 1

ADMIN

1 choose a datasource

2 start a new datastory

3 choose factors of interest

4 give us some context

now sit back and relax!

Give us some context

As much as you can. Context-free solutions lead to context free results, and we want to make sure your relationship with us is an investment rather than a cost!

What is your application?

What is your critical business objective related to this data? Or not related at all?

datastories

NEW DATASTORY

NEW DATA SOURCE

NEW DATA STORY

LOGOUT

ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 1

ADMIN

1 choose a datasource

2 start a new datastory

3 choose factors of interest

4 give us some context

now sit back and relax!

Now sit back, relax and give us some time to create your DataStory!

You can WATCH OUR PROGRESS HERE. Our algorithms are very computationally intensive. Finishing all steps of the analysis would take us a minimum of 10 minutes.

You can earn many karma points from us if you send an email with suggestions on how your experience could be made smoother. Please, write us to beta@datastories .com. Katya, Robbe, Sean, Sasha are all checking this email and will respond asap.

You will be redirected to the OVERVIEW PAGE in 2 secs

datastories

DATASTORY REPORT

SHOW STORY THUMBNAILS

PREV SLIDES

NEXT SLIDES

LOGOUT

ACCOUNT 

DASHBOARD

DATA SOURCES

YOUR STORIES 2

ADMIN

1 Data Overview

2 Your KPI

3 Simple Relationships

4 Relationships to KPI

5 Predictive Models

6 What-If's

7 Conclusions

Quick Summary of your Data

1 2

HERE IS WHAT WE GOT FROM YOU

You uploaded data_extended.csv file with a filesize of 1.68 Mb on 20.06.2016

ROW COUNT	COLUMN COUNT	MOSTLY NUMERIC COLUMNS
2199	69	100%

SELECTED KEY PERFORMANCE INDICATOR	COLUMNS CONTAINING NUMBERS	COLUMNS CONTAINING TEXT
Propylene	69	0

javascript:

datastories

DATASTORY REPORT

SHOW STORY THUMBNAILS

PREV SLIDES

NEXT SLIDES

LOGOUT

ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 2

ADMIN

Very Simple Relationships

1 | 1

EXPLORE SIMPLE RELATIONSHIPS BASED ON CORRELATION AND MUTUAL INFORMATION

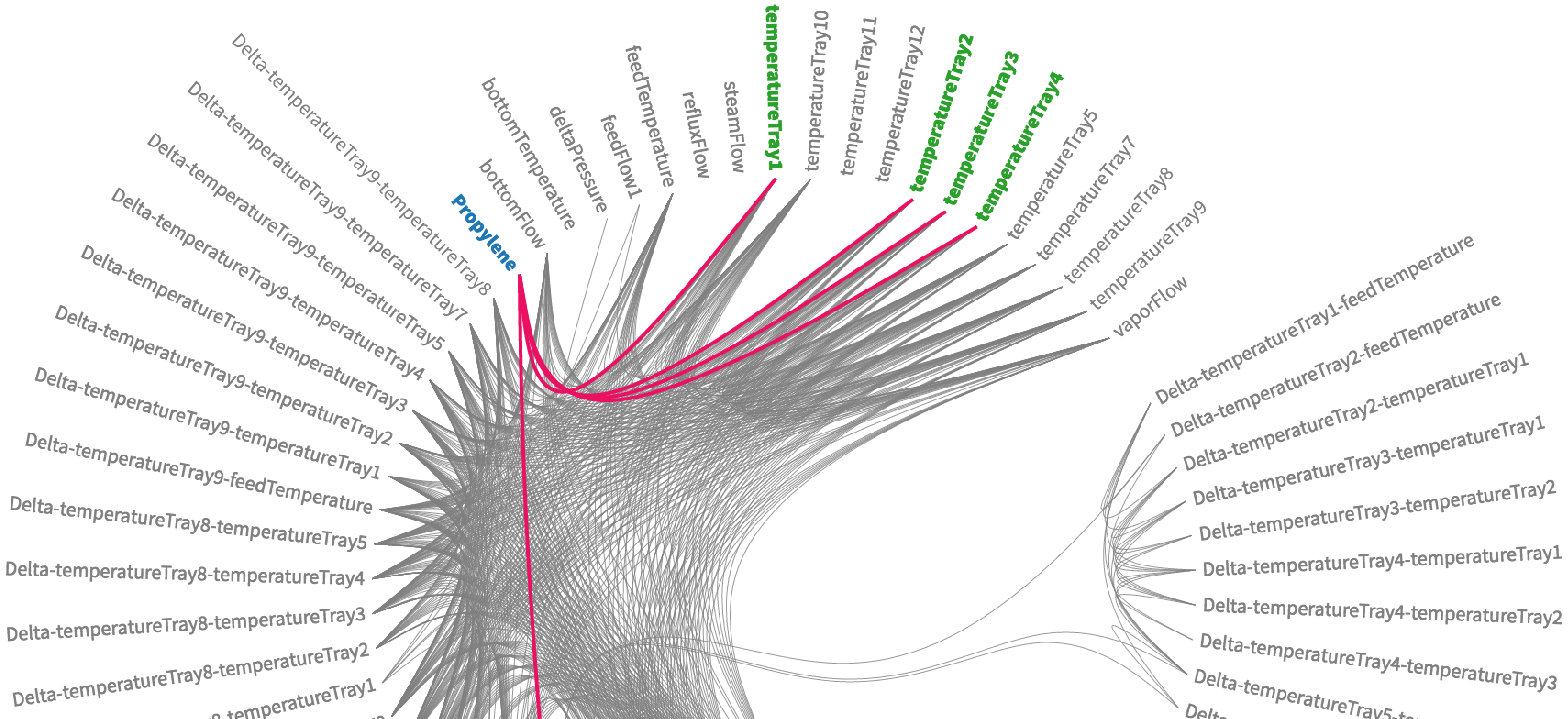
To drill down into how your KPI is connected to the metrics we first checked how all metrics are connected to each other. We found several tightly connected groups of your metrics. We thought you might want to know about inter relationships.

Below you can play with the first two sliders to see which columns would be connected to each other in terms of correlation or mutual information if you change the the thresholds. Now, when the first slider below is set to 20, we draw a line connecting columsn if their mutual information is greater than 20%. The more you move sliders to the right, the stricter your connection requirements would be, and only super-strong pair-wize relationships will be shown (if any).

Play with a mutual information threshold

Value = 90

Play with a correlation threshold



ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 2

ADMIN

Predictive Models

A SUMMARY OF PREDICTIVE MODELING RUNS

We had to create and challenge 25579 predictive models to deeply learn which metrics are necessary and sufficient to predict your KPI. A half of the computational effort was spent on meticulous cross-validations to make sure we avoid over-fitting and maximizing the predictive power of models given your data. At the end we have build a final ensemble of 100 models with a minimal number of metrics, which you can use to run interacive "what-if" scenarios.

The final ensemble has the following characteristics:

AVERAGE CROSS-VALIDATION CORRELATION ACCURACY	NUMBER OF METRICS:	NUMBER OF METRICS WE STARTED WITH:
96.8 %	3	68

datastories

DATASTORY REPORT

SHOW STORY THUMBNAILS

PREV SLIDES

NEXT SLIDES

LOGOUT

ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 2

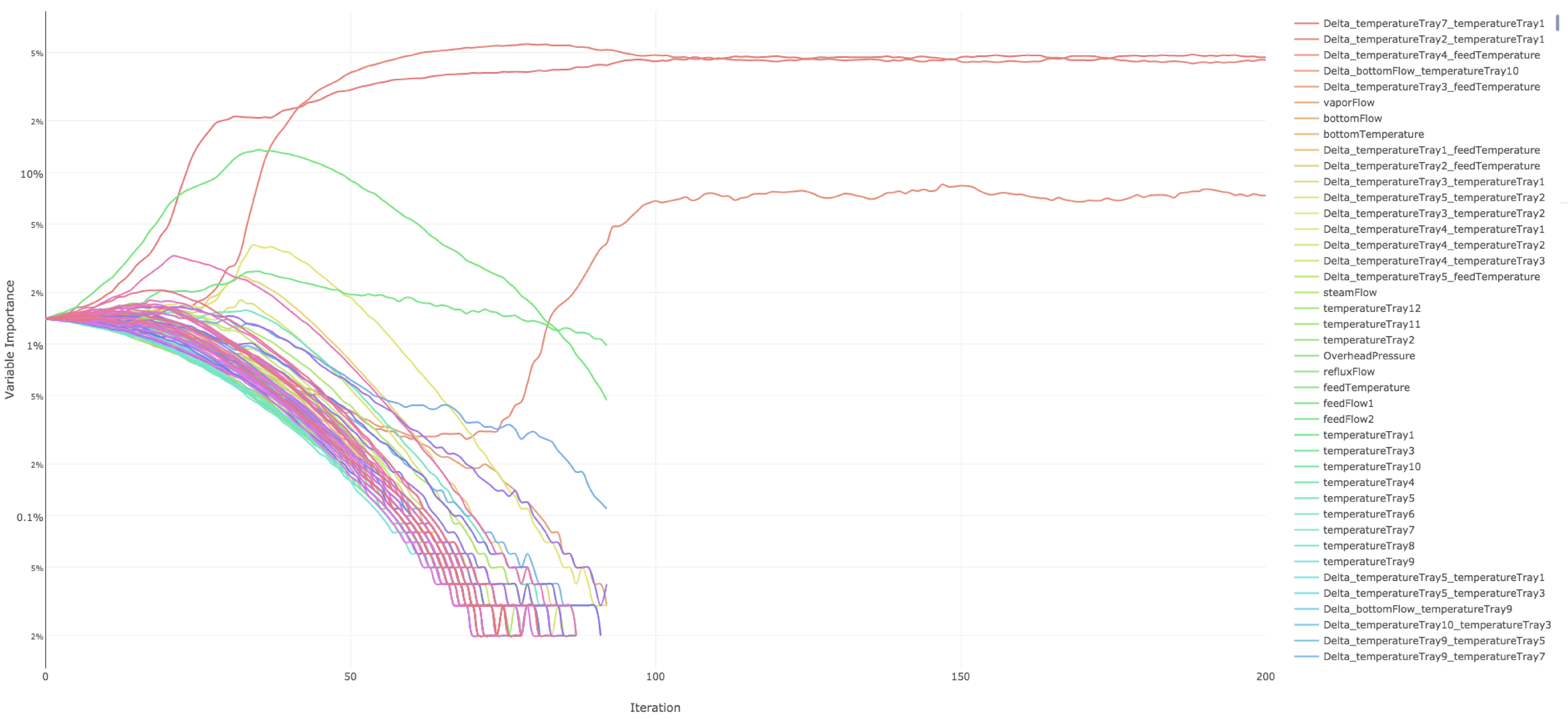
ADMIN

Predictive Models

A SUMMARY OF PREDICTIVE MODELING RUNS

1 1

Variable importance



datastories

DATASTORY REPORT

SHOW STORY THUMBNAILS

PREV SLIDES

NEXT SLIDES

LOGOUT

ACCOUNT

DASHBOARD

DATA SOURCES

YOUR STORIES 1

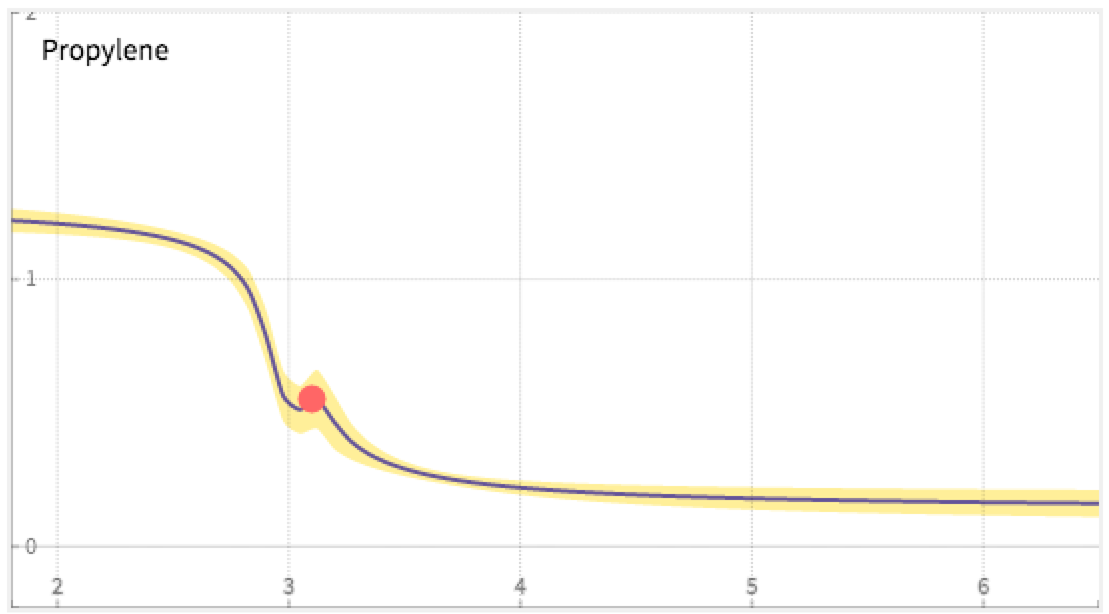
ADMIN

What-If's

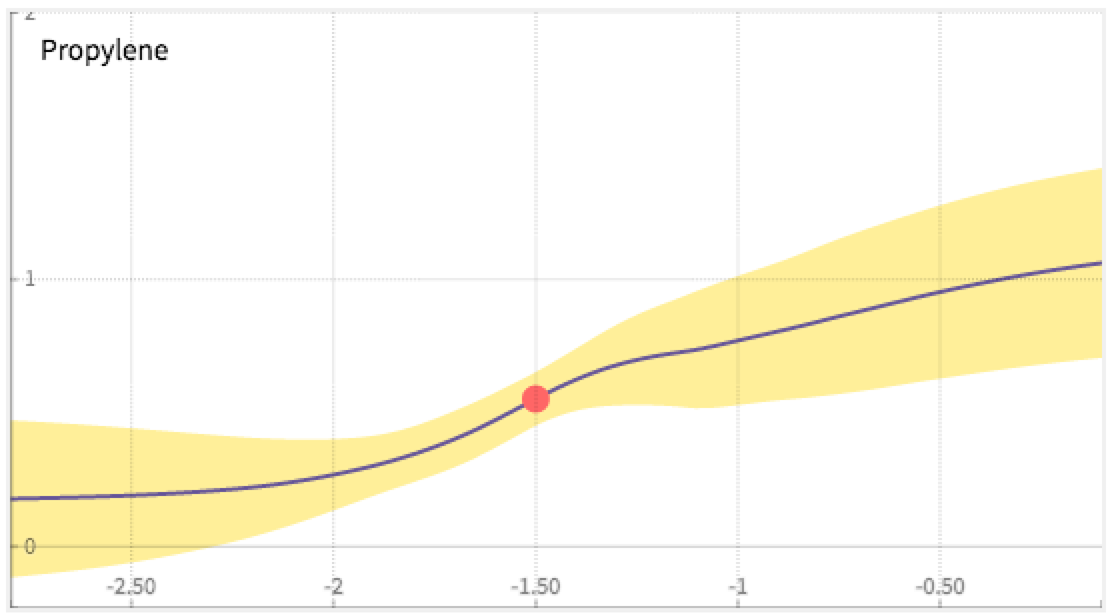
1 1

PLAY DIFFERENT WHAT-IF SCENARIOS BELOW TO SEE WHAT HAPPENS TO THE KPI. HOVER OVER THE GRAPHS FOR MORE INFO

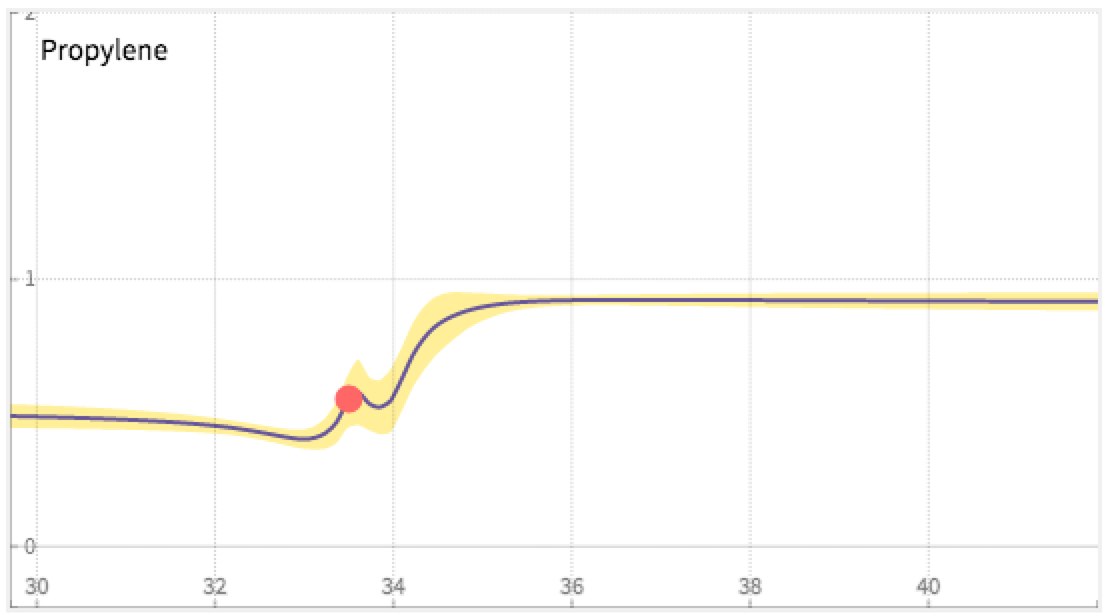
Propylene: 0.55



Delta_temperatureTray2_temperatureTray1: 3.10



Delta_temperatureTray4_feedTemperature: -1.50



Delta_temperatureTray7_temperatureTray1: 33.50

deltaPressure	OverheadPressure	OverheadTemperature	refluxFlow	feedTemperature	feedFlow1	feedFlow2	temperatureTray1	temperatureTray2	temperatureTray3	temperatureTray4
4.04	382.0	-152.0	135.0	44.6	269.0	1730.0	39.1	42.0	42.0	43.1
4.05	382.0	-152.0	135.0	44.4	264.0	253.0	38.4	41.4	41.5	42.9
4.05	382.0	-152.0	135.0	43.2	264.0	188.0	36.8	39.6	39.8	41.2
4.06	382.0	-152.0	135.0	44.0	264.0	440.0	39.2	42.1	42.0	43.1
4.06	382.0	-152.0	135.0	42.5	263.0	271.0	36.6	39.7	39.8	40.9
4.06	382.0	-152.0	135.0	44.0	269.0	554.0	38.4	41.3	41.3	42.5
4.06	382.0	-152.0	135.0	44.3	267.0	1040.0	38.4	41.4	41.5	42.8
4.06	382.0	-152.0	135.0	44.3	267.0	438.0	38.7	41.6	41.7	42.8

Showing 1 to 9 of 2,199 entries

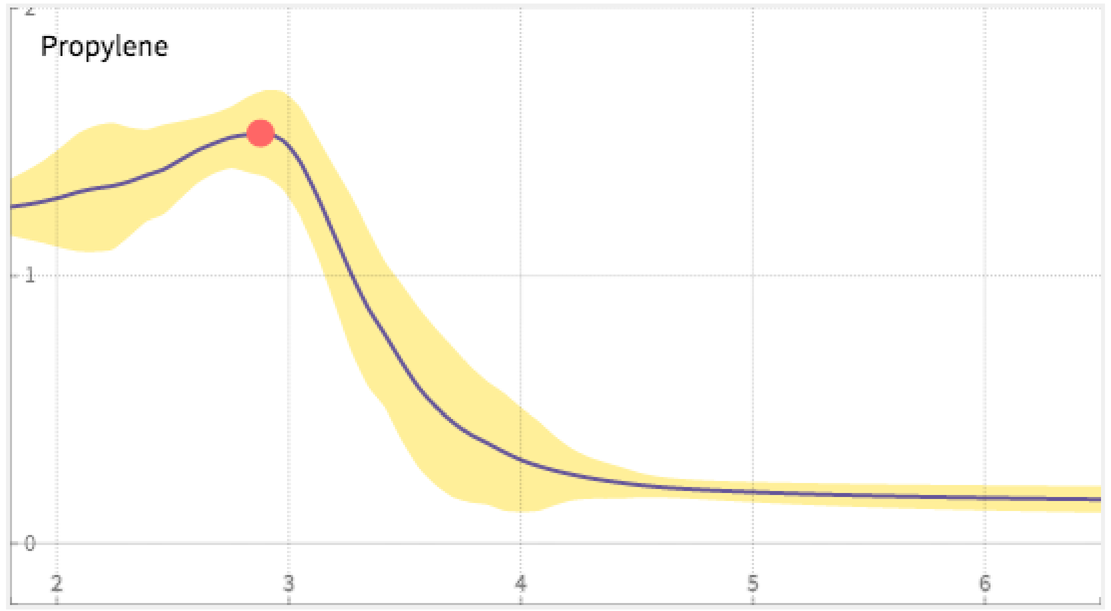
Minimize Reset Maximize

What-If's

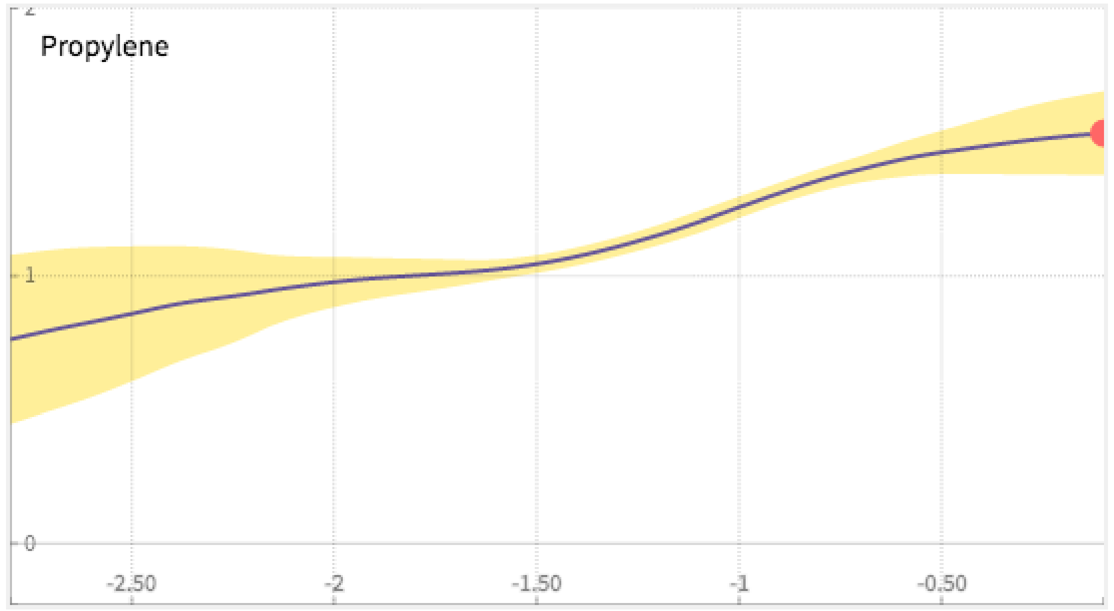
1 1

PLAY DIFFERENT WHAT-IF SCENARIOS BELOW TO SEE WHAT HAPPENS TO THE KPI. HOVER OVER THE GRAPHS FOR MORE INFO

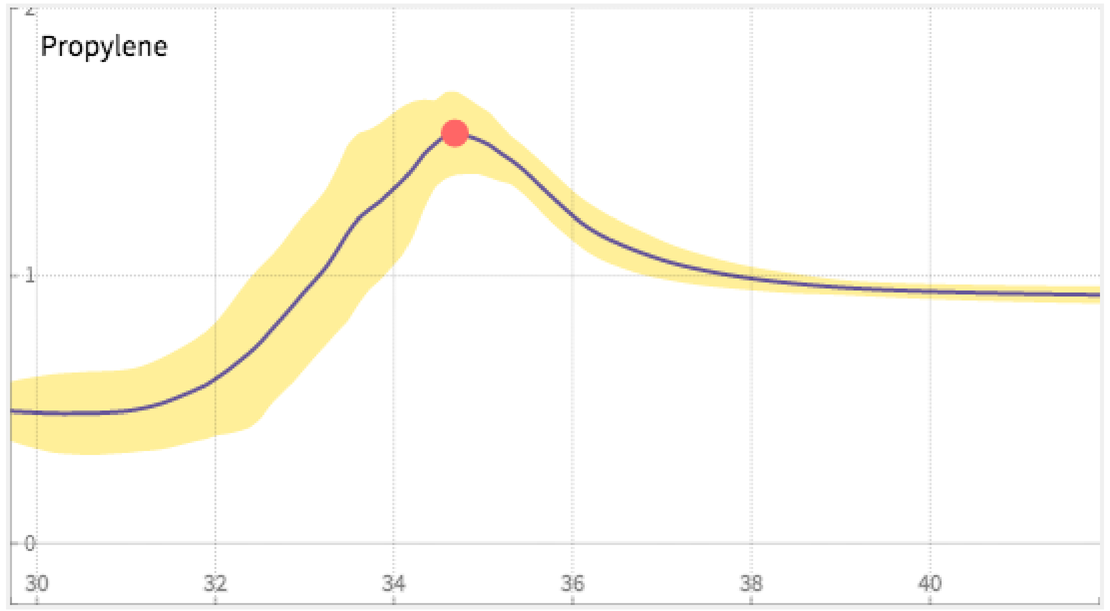
Propylene: 1.53



Delta_temperatureTray2_temperatureTray1: 2.88



Delta_temperatureTray4_feedTemperature: -0.10



Delta_temperatureTray7_temperatureTray1: 34.68

deltaPressure	OverheadPressure	OverheadTemperature	refluxFlow	feedTemperature	feedFlow1	feedFlow2	temperatureTray1	temperatureTray2	temperatureTray3	temperatureTray4
4.04	382.0	-152.0	135.0	44.6	269.0	1730.0	39.1	42.0	42.0	43.1
4.05	382.0	-152.0	135.0	44.4	264.0	253.0	38.4	41.4	41.5	42.9
4.05	382.0	-152.0	135.0	43.2	264.0	188.0	36.8	39.6	39.8	41.2
4.06	382.0	-152.0	135.0	44.0	264.0	440.0	39.2	42.1	42.0	43.1
4.06	382.0	-152.0	135.0	42.5	263.0	271.0	36.6	39.7	39.8	40.9
4.06	382.0	-152.0	135.0	44.0	269.0	554.0	38.4	41.3	41.3	42.5
4.06	382.0	-152.0	135.0	44.3	267.0	1040.0	38.4	41.4	41.5	42.8
4.06	382.0	-152.0	135.0	44.3	267.0	438.0	38.7	41.6	41.7	42.8

Showing 1 to 9 of 2,199 entries

Minimize Reset Maximize

Conclusions of the DataStory Predicting Propylene Output (extended dataset)

HERE IS WHAT WE LEARNED ABOUT YOUR DATA

We analyzed **Predicting Propylene Output (extended dataset)** to assess what drives your key performance metric **Propylene** using **68** columns you provided. We explored the data health of your data and rated it at **60** in general. Your data had **2199** rows, and the KPI has **158** unique values.

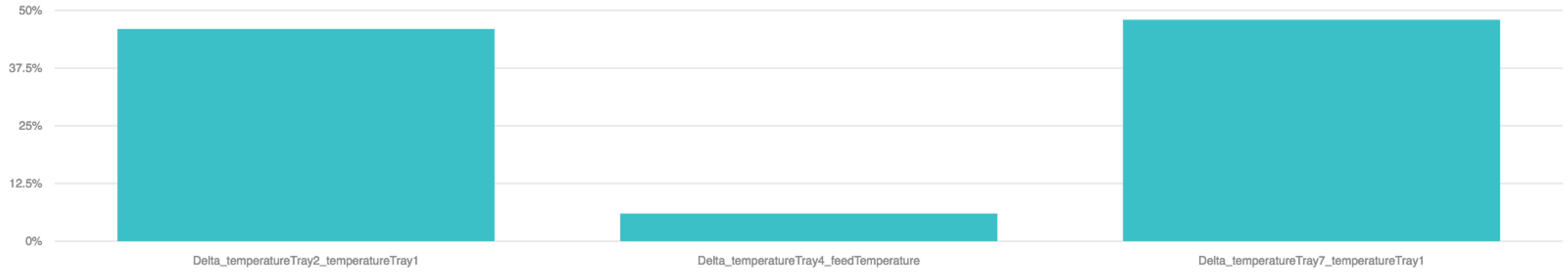
Your data set had **none** missing cells . With respect to predicting the **Propylene** the health of your data is **58**. At this stage DataStories focus on finding reliable relationships between numeric metrics and your KPI. So, we had to look at **68** metrics remaining after omitting **0** We first looked at how your metrics impact the KPI individually. For this we performed a standard correlation analysis and a more involved analysis of the mutual information content between **Propylene** and all other inputs individually. Because the data you provided only had **68** columns on top of the KPI we also computed all individual pairwise relationships (correlation and mutual information) among the metrics to see how things are connected to each other. Based on initial results we could conclude that out of **68**, **1** could be removed from the consideration whatsoever, because they do not have even slight independent relationships to your **Propylene**.

From this preliminary analysis we could conclude that only **1388** inputs are individually related to your KPI, but many of them are correlated to each other. This means, that you could further improve focus and keep searching for a minimal set of metrics that matter. We did this for you! After deeply learning your prediction problem and having created and challenged **25579** models, we discovered that **3** are sufficient to predict your **Propylene** at **XX % correlation**. These driver metrics have various influence on the KPI and have to be used together to make robust predictions. The drivers are **Delta_temperatureTray2_temperatureTray1 (Importance: 46%)**, **Delta_temperatureTray4_feedTemperature (Importance: 6%)**, **Delta_temperatureTray7_temperatureTray1 (Importance: 48%)**, alltogether their importances sum up to 100%. You can play with how they impact **Propylene** in the What-if scenario tools (here).

If by exploring the drivers you realized that some of them are very difficult to measure or control, or might be coupled with your performance, try to re-run the DataStory while eliminating them from the list of candidate metrics during the DataStory setup. Now when we have the models we can identify outliers, or optimize the models to find optimal settings to achieve desired **Propylene** levels. This is a premium feature, please, contact us to discuss this.

We are working very hard to add model evaluation functionality and model export functionality and for now you can upload the data with empty KPI values, and we will fill them in with predictions.

Let us know how you liked it!
DataStories




datastories ™
Best Prize
THE MOST
HOLISTIC
APPROACH
Manufacturing
Sustainability

datastories

P&G

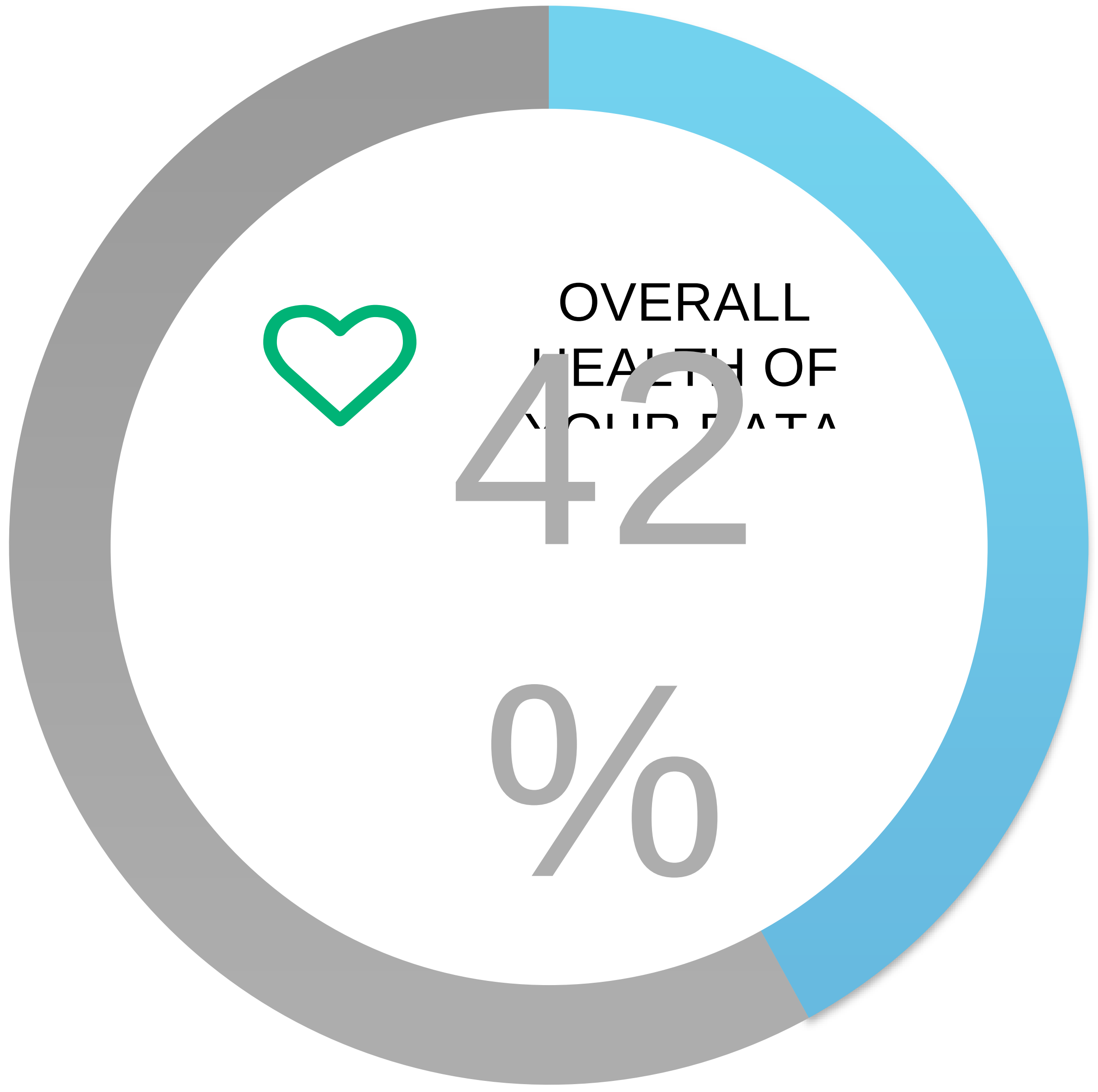
Challenges

Predictive Models

Interactive Dashboard

Data-Driven Innovation guiding Sustainability

Process Data: 154,260 energy observations



TOTAL NUMBER OF TAGS	TAGS WITH AT LEAST 50% DATA	TOTAL NUMBER OF ENERGY OUTPUTS
1,239	1,183	12

CONSTANT TAGS	DISCRETE TAGS	CONTINUOUS TAGS
292	609	574

Our Approach to the Hackathon Challenge

1. Collect logged measurements from all tags
2. Process, organize, and aggregate the data
3. Run predictive modeling, find energy consumption drivers and predict energy consumption
4. Deploy predictions in a dashboard with interactive what-if scenarios

Our Hackathon Outcomes

1. Built Non-linear Predictive Models for Power Consumtion with local error bounds

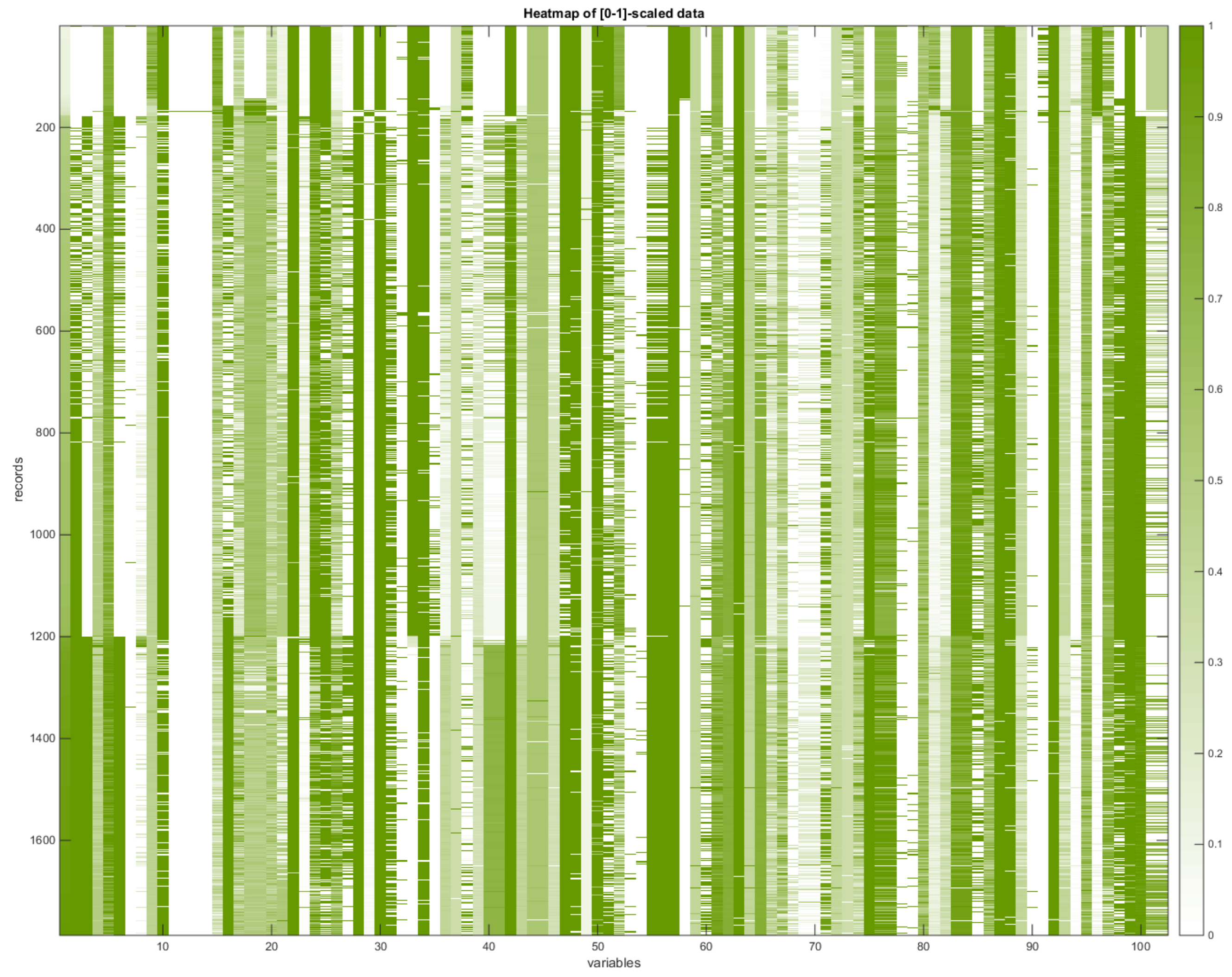
2. Which lead to a Dashboard with focus on the variables that currently matter most

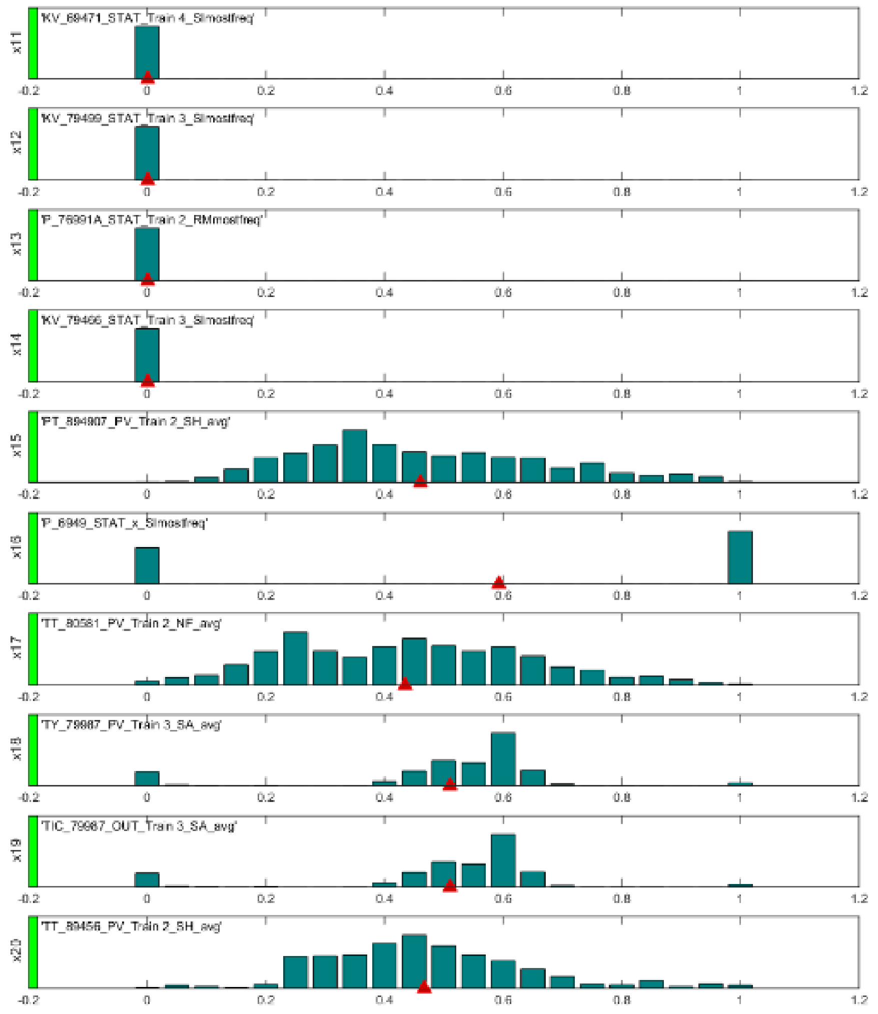
3. That also integrates an innovative approach to run What-If Scenarios starting from the current process state

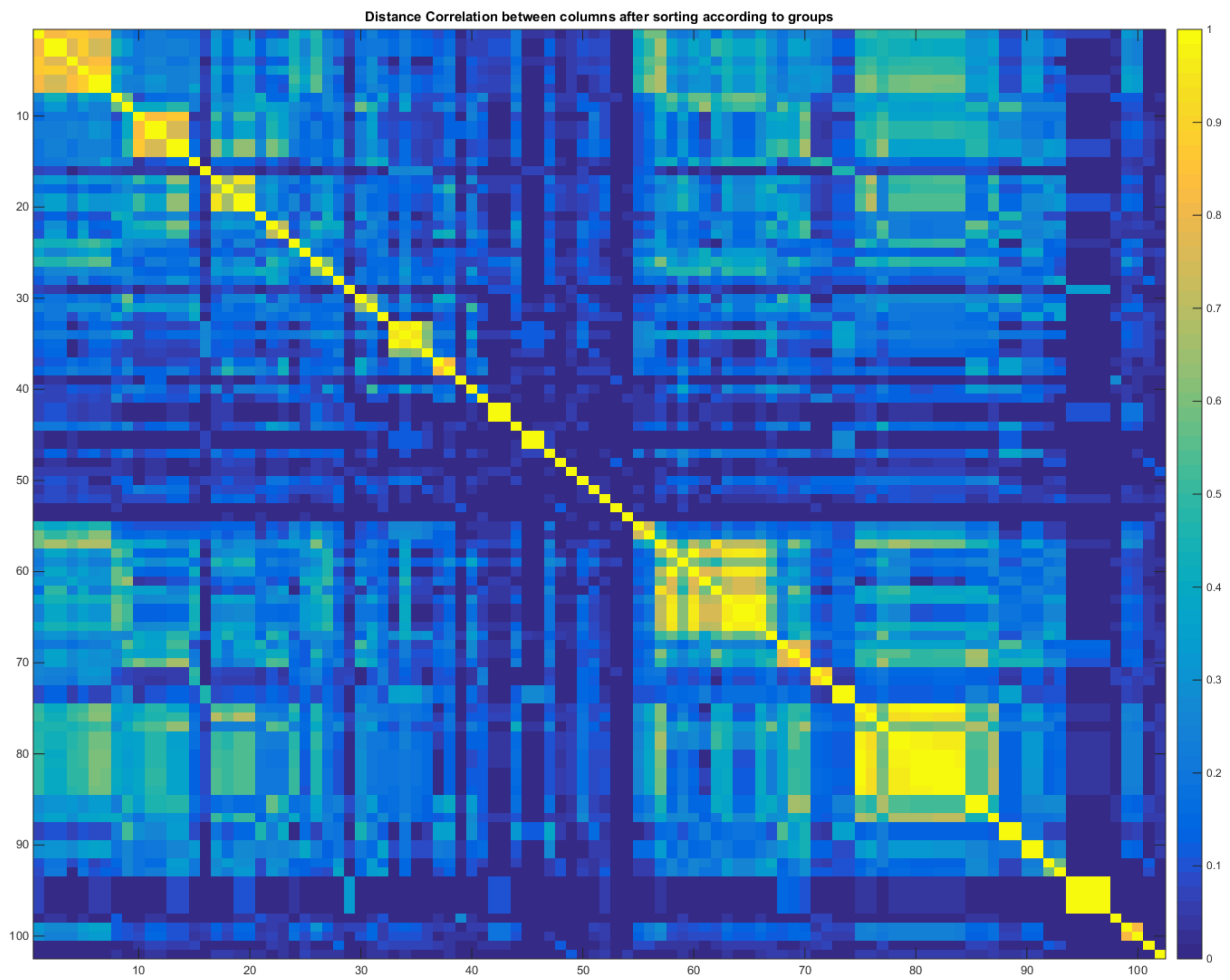
4. And allows you to maximize the Sustainability of your process at all times.

Data Pre-Processing & Analysis

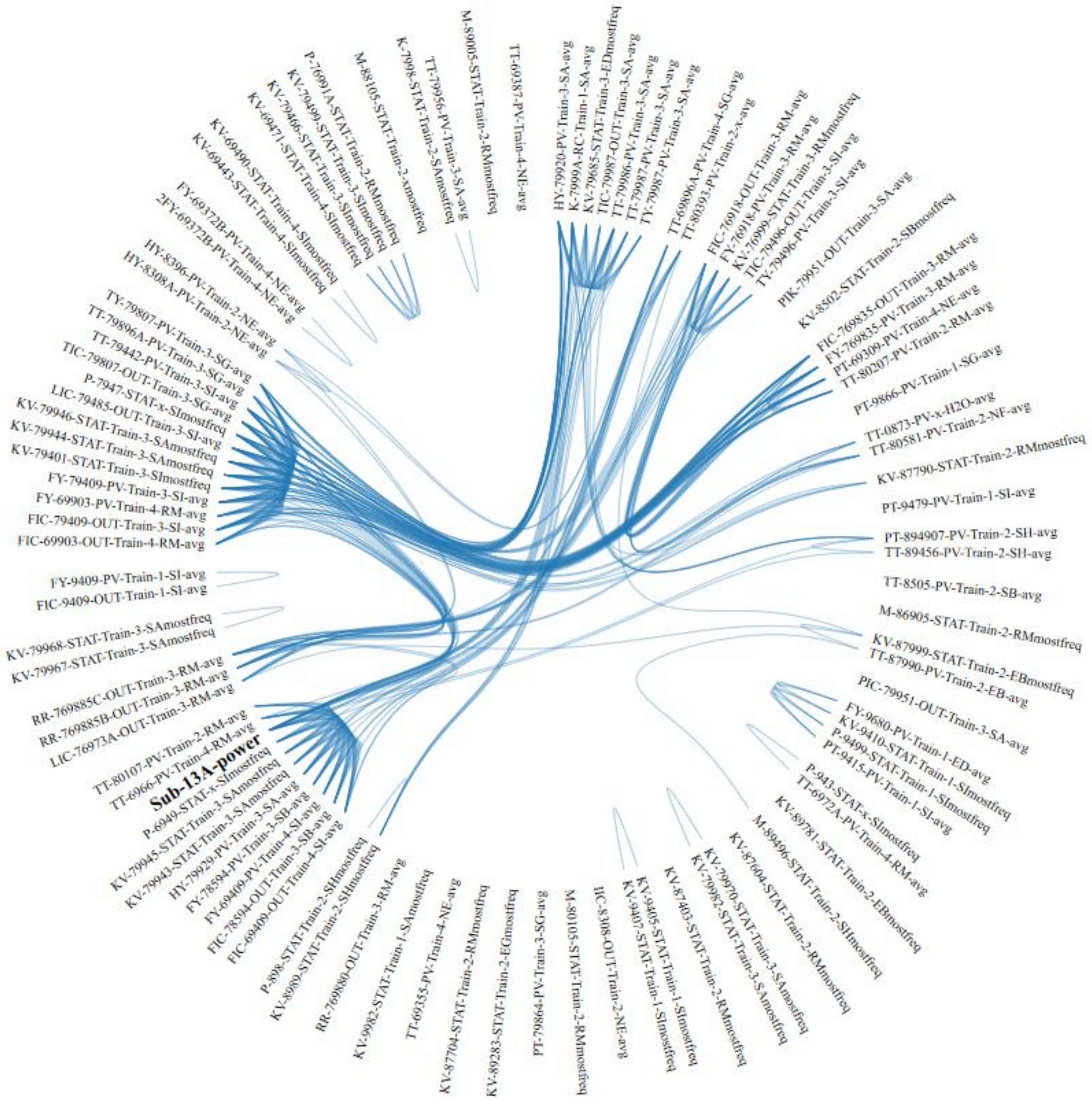
- ✓ Integrated and aligned all available tags with energy data
- ✓ Generated monthly and quarterly datasets with 5 min averages
- ✓ Looked at Data Health, Data distributions, Linear Correlations, Mutual Information Content and Variable Connections & Grouping







Interactive Explorer of Tag Relationships



Our Modeling Process

Organize all your process data into one big table

↓

Define your KPIs

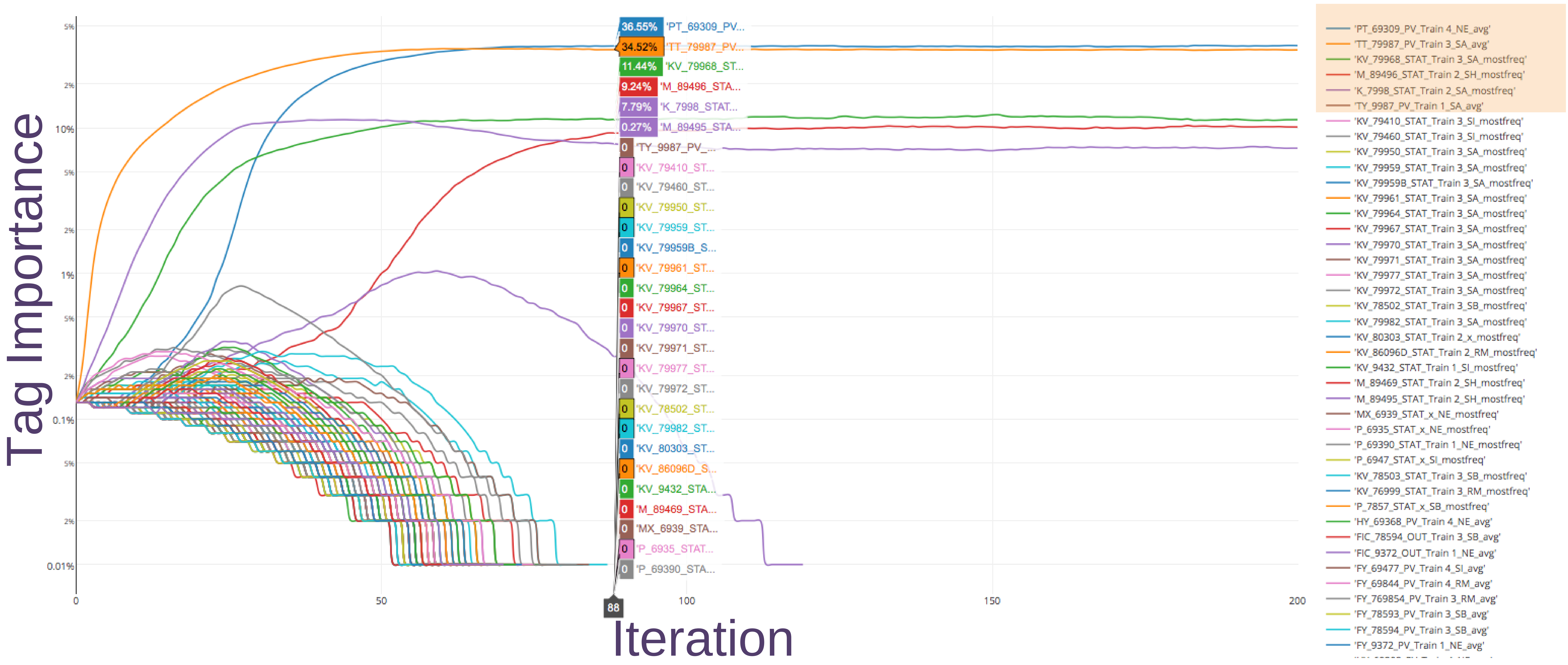
↓

Apply advanced machine learning

↓

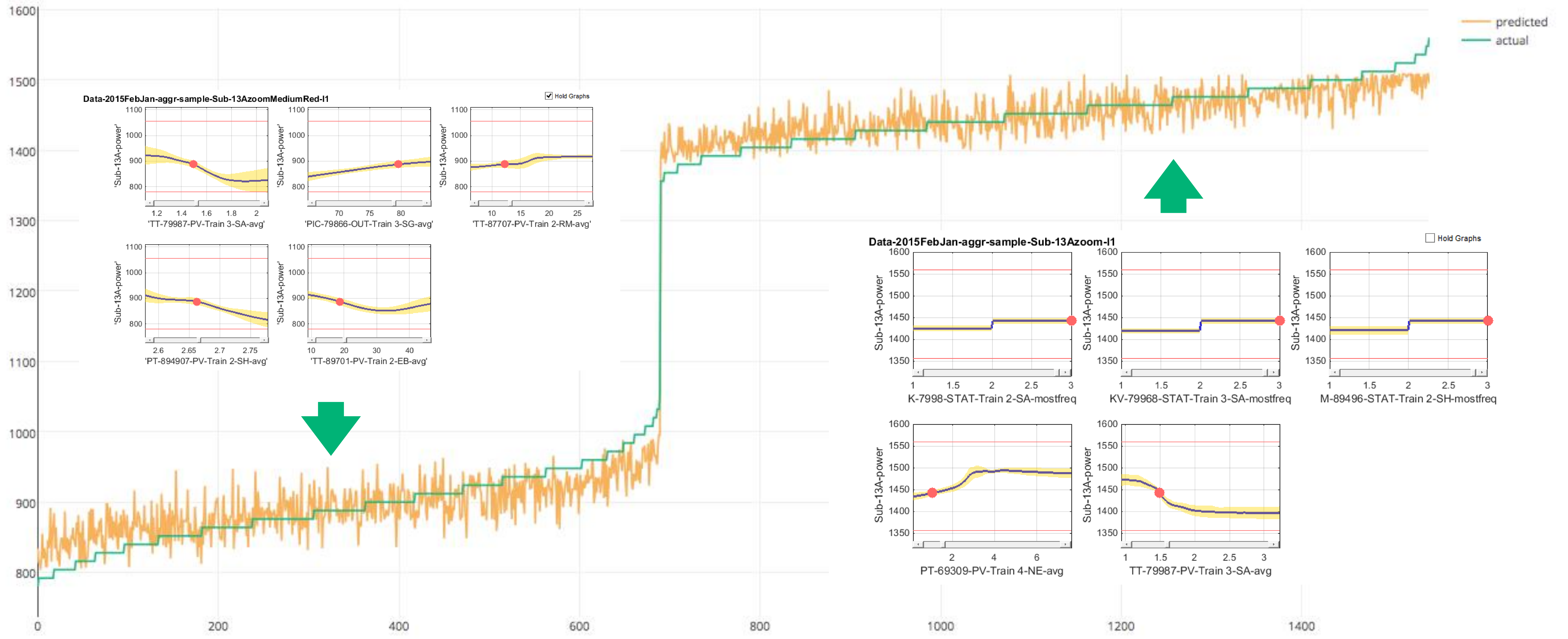
$$$ Identify driving metrics & Deploy Predictive Models $$$

Build Compact Non-linear Models per regime using extensive process of variable competition and elimination



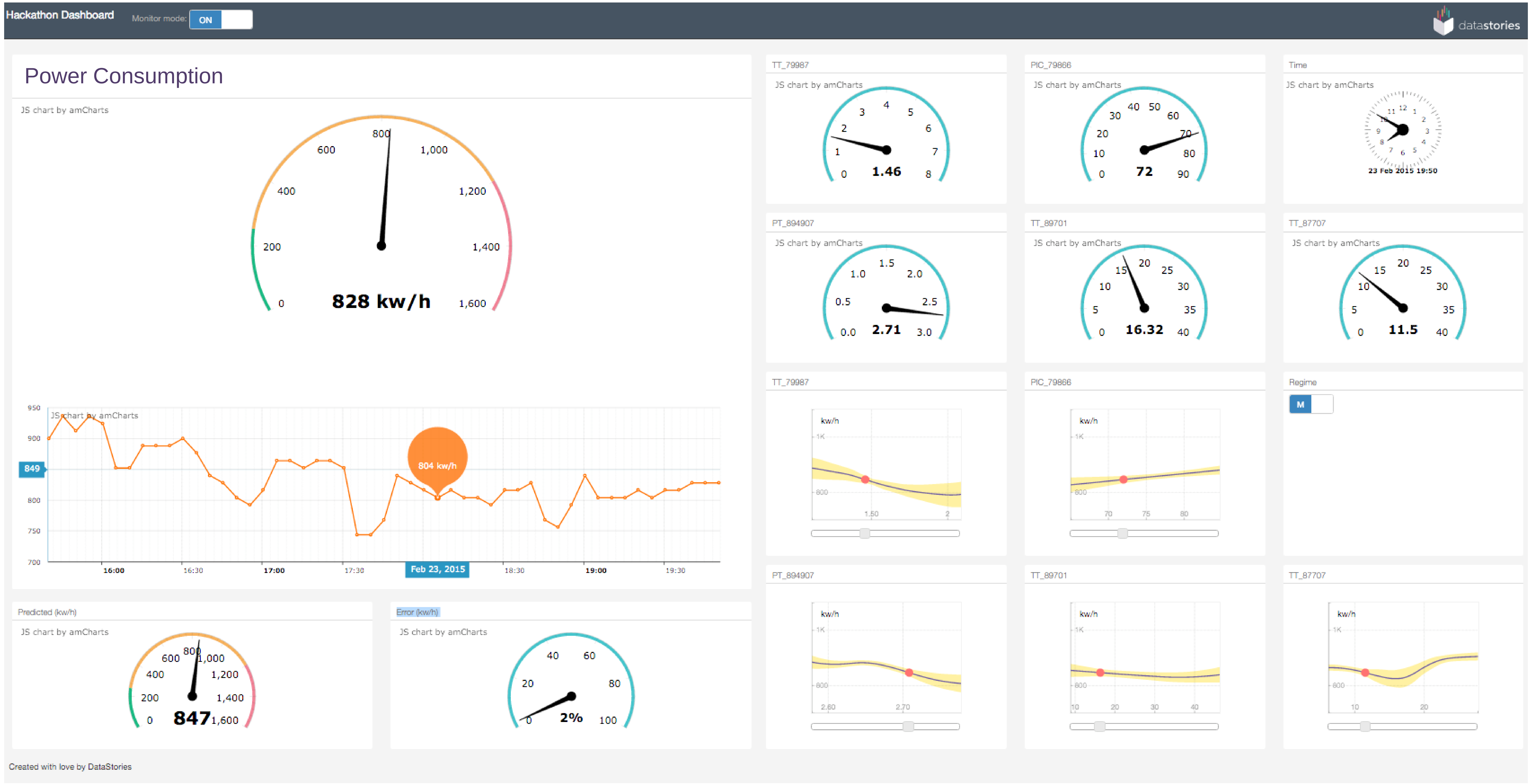
Result is compact 5-variable model-ensembles with error limits for each regime

Energy Consumption



Measurements sorted by Power Consumption

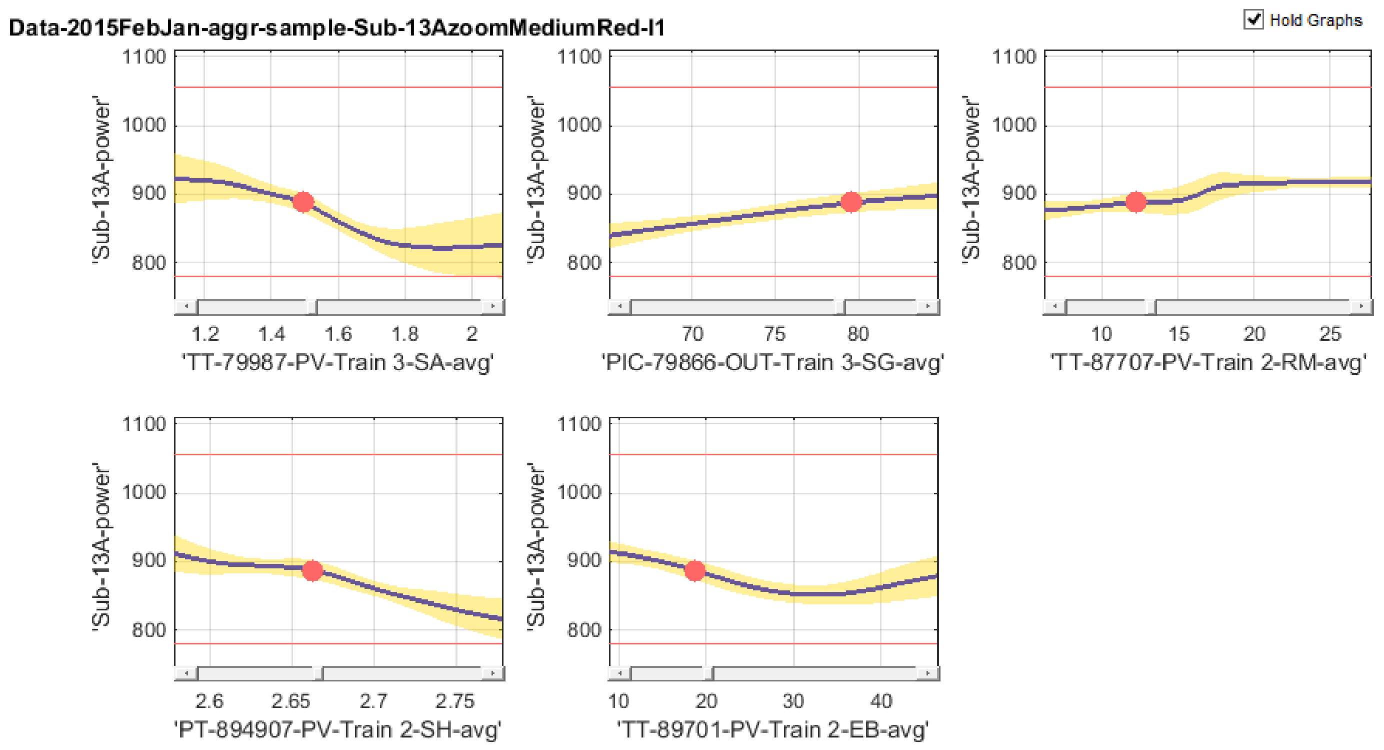
Dashboard with a focus

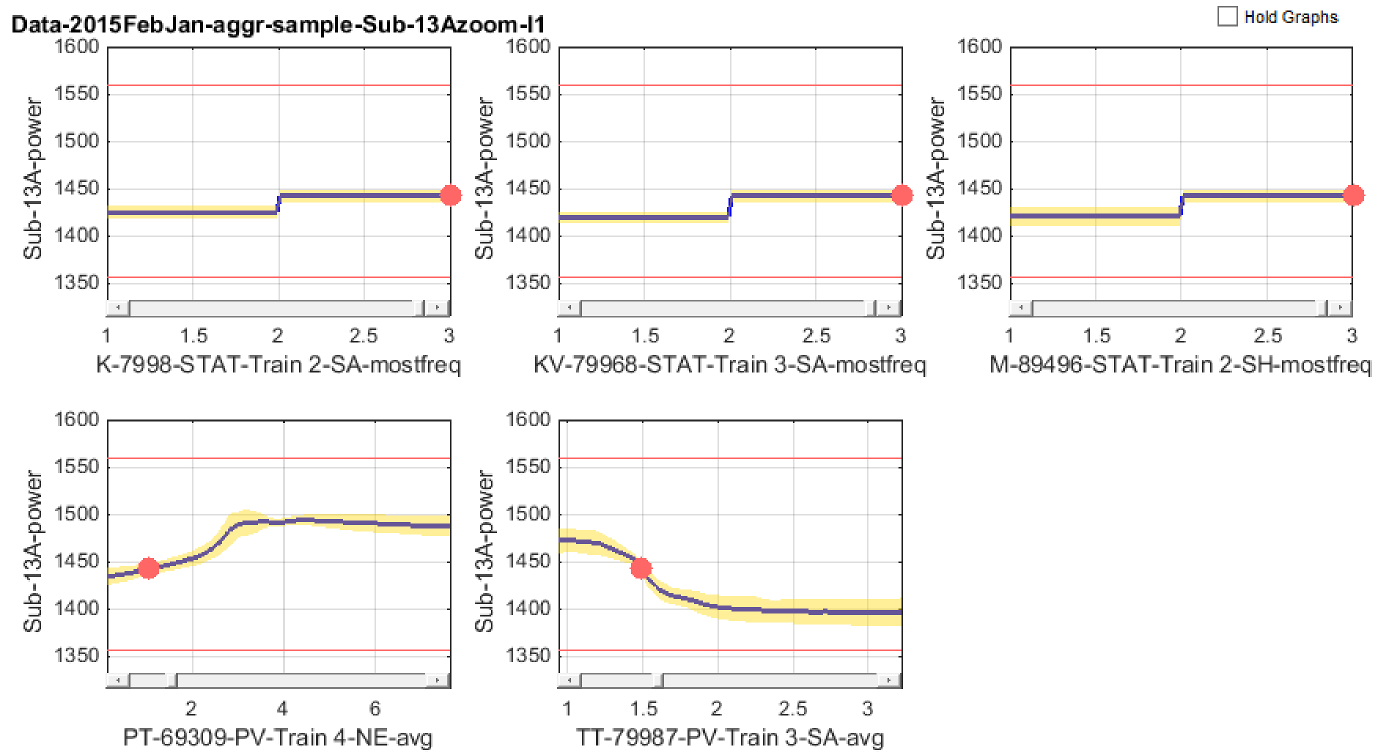



Chrome File Edit View History Bookmarks People Window Help
Tue 10:38
Katya Vladislavleva
Dashboard#1
localhost:8000/Hackathon/dashboard_monday/dashboard.html
Hackathon Dashboard
Monitor mode:
ON
datastories
Power Consumption
888 kw/h
864 kw/h
806
Feb 23, 2015
08:30
09:00
09:30
10:00
10:30
11:00
11:30
12:30
TT_79987
1.52
PIC_79866
72.21
Time
23 Feb 2015 12:35
PT_894907
2.69
TT_89701
16.23
TT_87707
12.47
TT_79987
kw/h
PIC_79866
kw/h
Regime
M
PT_894907
kw/h
TT_89701
kw/h
TT_87707
kw/h
Predicted (kw/h)
847
Error (kw/h)
5%
Created with love by DataStories

Models are Robust by Design

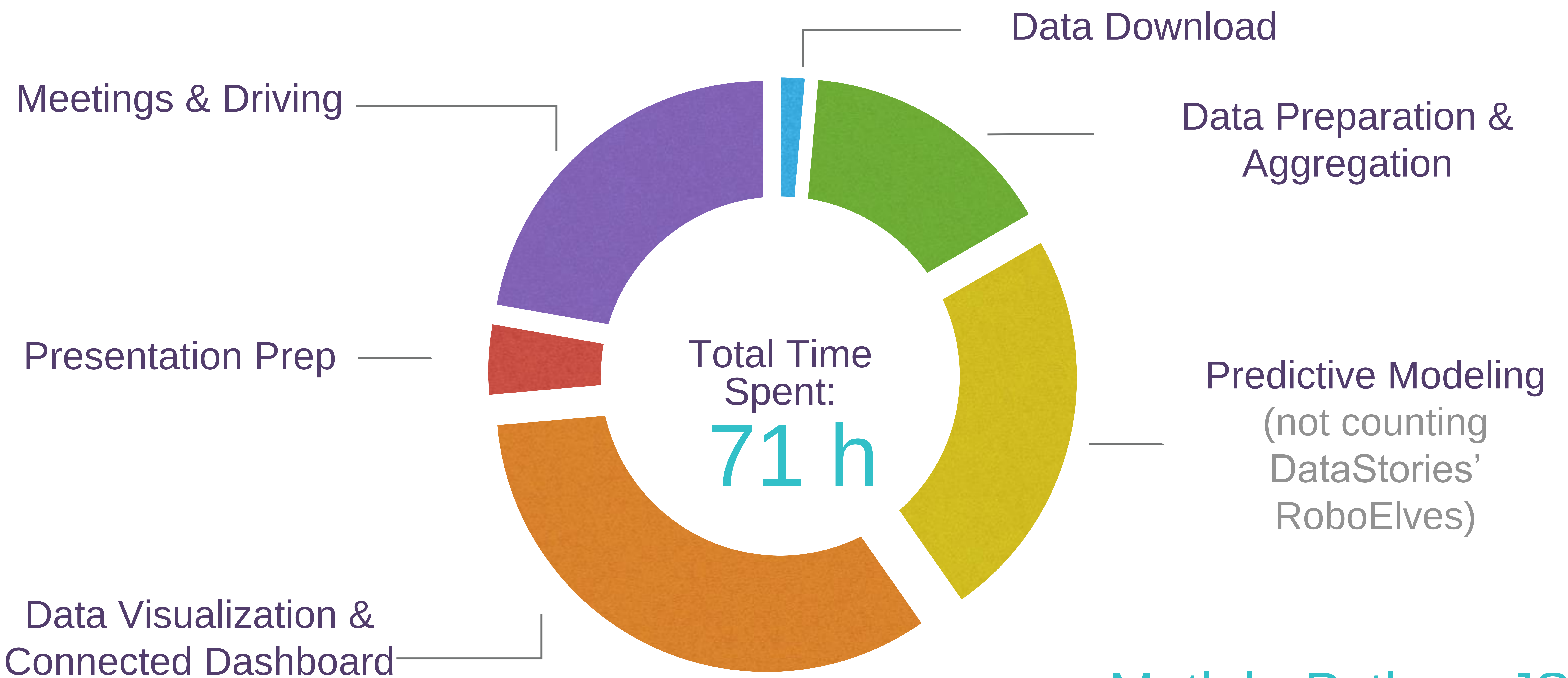
Energy Consumption





- ✓ 600,000 models created to produce a final ensemble for one region
- ✓ 2 regions of electricity consumption lead to 1,200,000 models
- ✓ 50% of the modeling effort is spent on cross-validation and making sure the models are predictive and not over-fitting
- ✓ From 1,183 potential inputs only five (5) metrics per region are necessary and sufficient
- ✓ Global R2 0.97; per regime R2 is 0.68-0.7
- ✓ Final ensembles consist of 100 models each and also provide confidence limits

Time It took Us



Matlab, Python, JS, HTML5

Benefits of using Matlab

- ✓ Super fast implementation
- ✓ Reliable deployment + flexibility (Matlab package, Stand-alone, Cloud)
- ✓ Code protection
- ✓ Matlab users can integrate it easily in their Matlab-based routines
- ✓ Plug&Play Hadoop integration

Business Outcomes

✓ Project of high business value

✓ Perfect product validation

Special Thanks to Brussels Data Science Community for organizing the hackathon

datastories

Subscribe for the VIP beta at beta.datastories.com