Introduction to Data Science for PI Data for PI Professionals

Ahmad Fattahi – Manager, Data Science Enablement, OSIsoft

Kleanthis Mazarakis – Data Scientist, OSIsoft

Objective

- Gain a basic understanding of data science process
- Understand how PI System fits in the picture

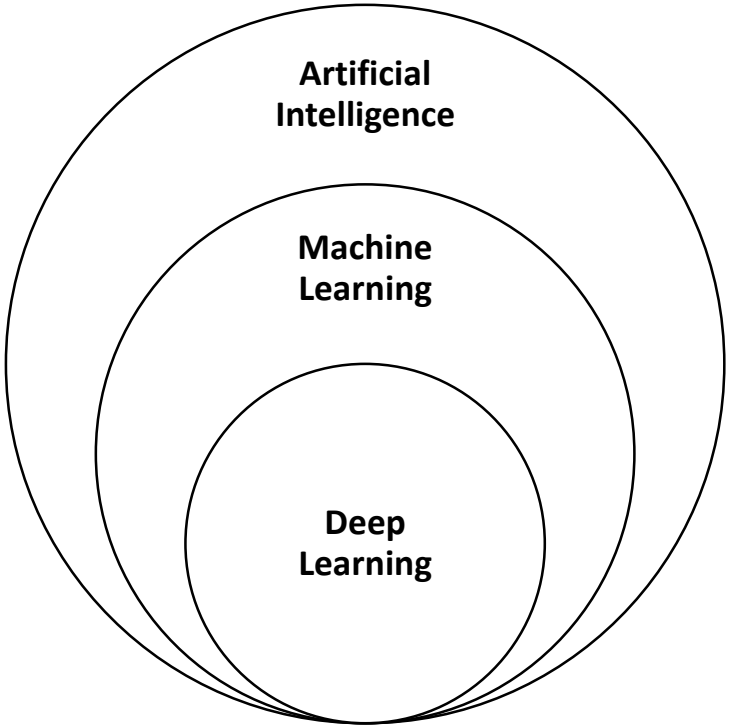
Agenda

- Nomenclature
- A primer on data science practice
- CRISP-DM
- Data science for PI data: a case study

Nomenclature

Data Science is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms.

-wikipedia



How do you tell myth from fact?

It’s all about VALUE!

Data Science should be a core organizational capability as opposed to a technical practice.

Always start with a sharp question.

Not so sharp: how can we improve our bottom line?

Sharp: how can we save on energy to cool our building and what factors contribute to that?

Start Small

How Do You Pick the Right Analysis?

Descriptive/Exploratory	Inferential/Predictive	Causal/Mechanistic
• What is the number of days in which the set point was never reached? • What is the average power usage across all those days?	• Which features can predict the power usage on a certain day?	• What causes power consumption to fluctuate?

Is the goal of the project to…

… predict?

… control?

What Functions Do You Need For Success?

Data Engineer

- Architecture, infrastructure, data governance

Data Scientist

- Cleaning, analysis and communication

Subject Matter Expert

- Intuition, what matters and what doesn't

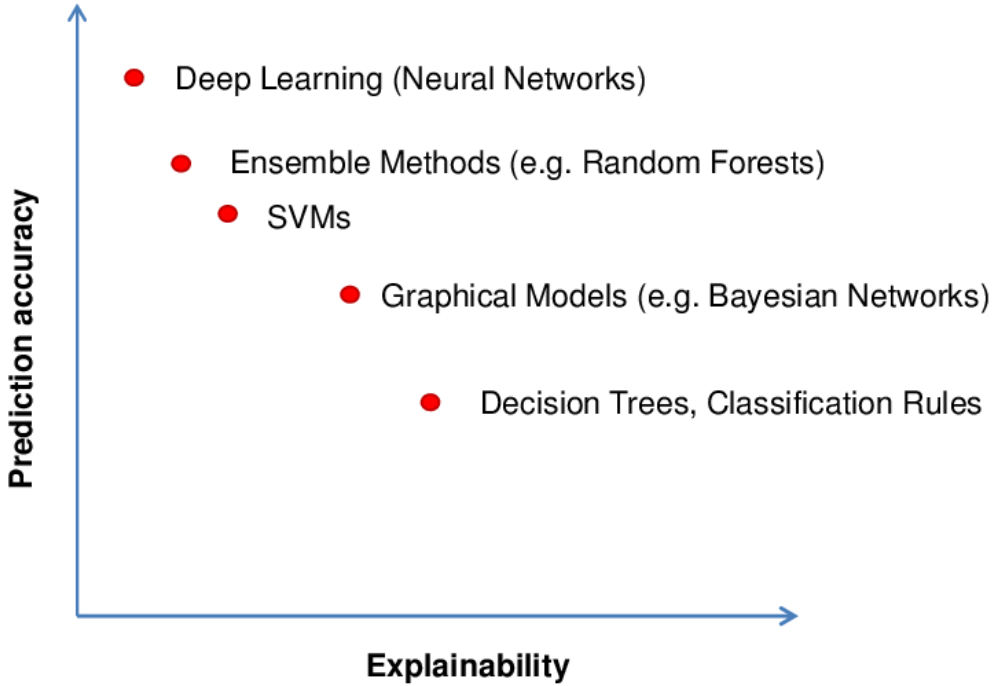
Data and Subject Matter Management

- Ownership, support

Communication is key.

Explainability

The Explainability Tradeoff

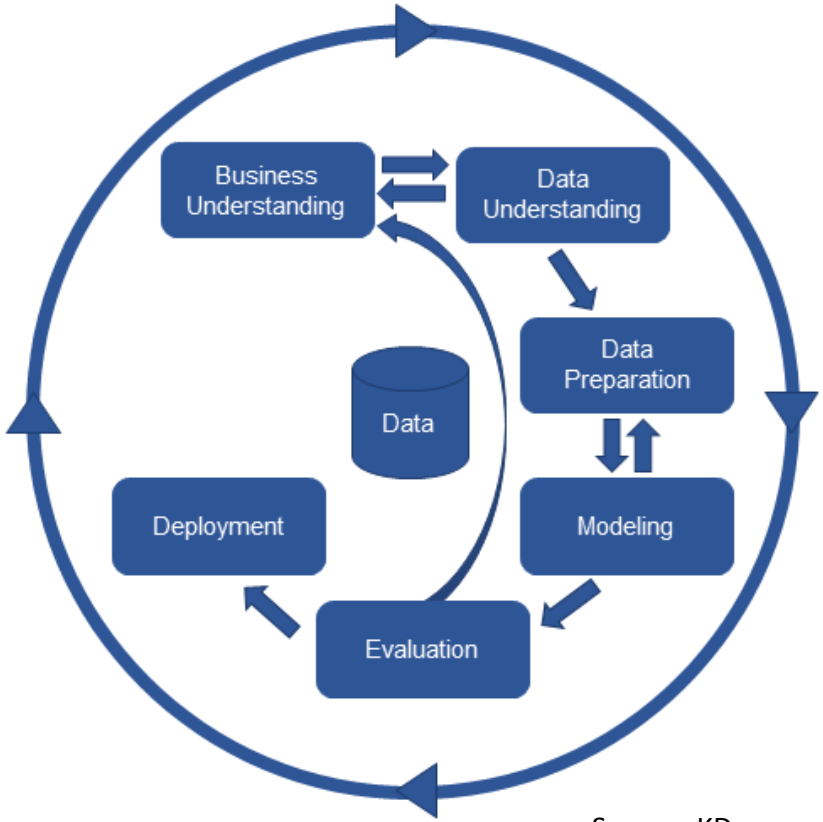


Source: ResearchGate GmbH

Story: Optimize Building Energy Consumption

CRISP-DM

- CRoss Industry Standard Process for Data Mining
- Among most popular methodologies
- Emphasizes cycles and iterations



Source: KDnuggests

Reproducible Work Is the Differentiator

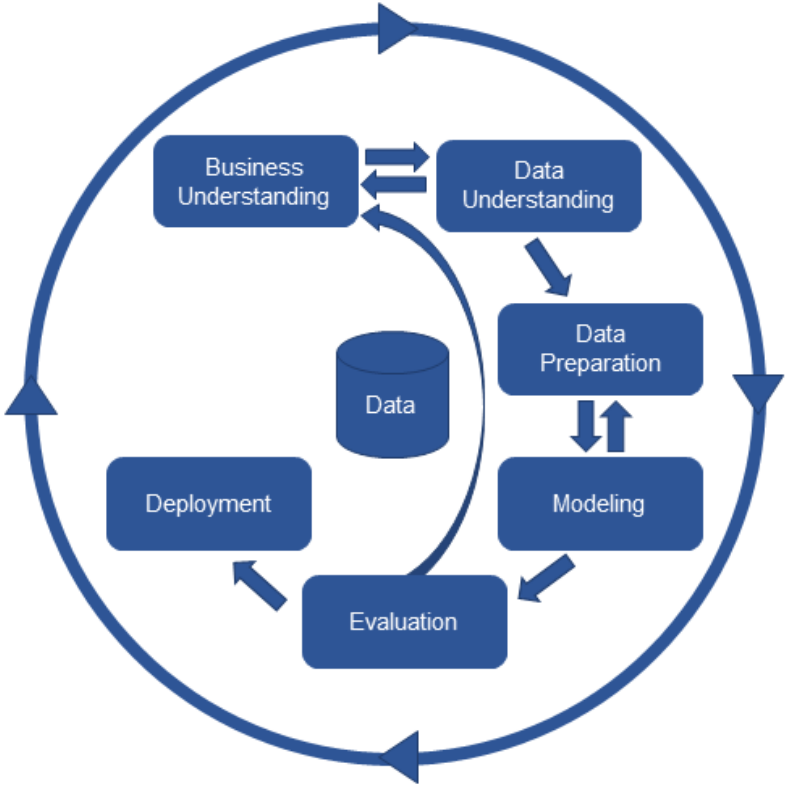
Assume your work is going to be repeated and tweaked frequently

Over time:

- Models veer off
- Physical systems change
- Priorities evolve
- New business owners come
- You get reassigned!

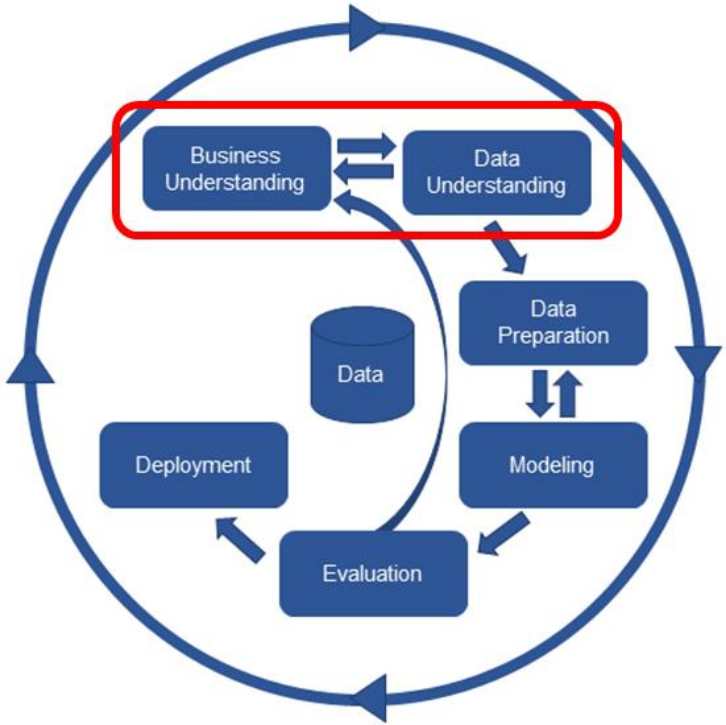
Leverage tools such as Jupyter Notebooks or other commercial platforms

The Cycle Repeats



Case study: Interacting with PI System data

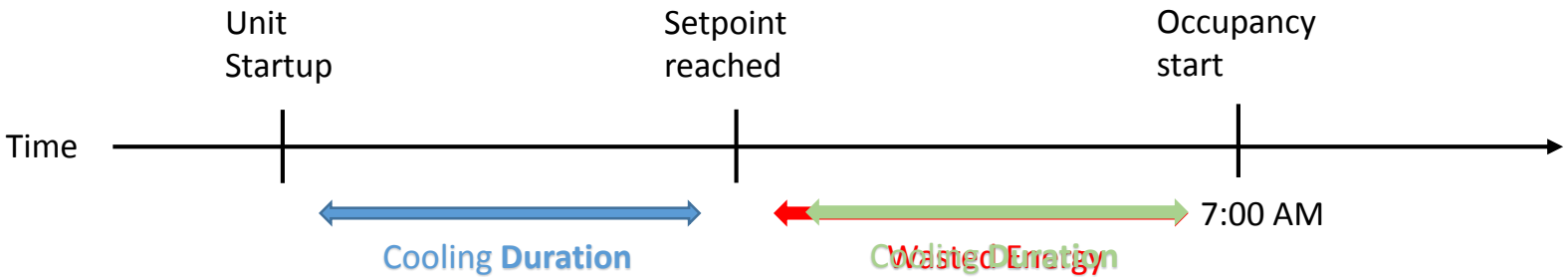
Business and Data Understanding



- What is the Business Objective?
- What data is available?
- What data is missing?

Case Study: Facilities Optimization

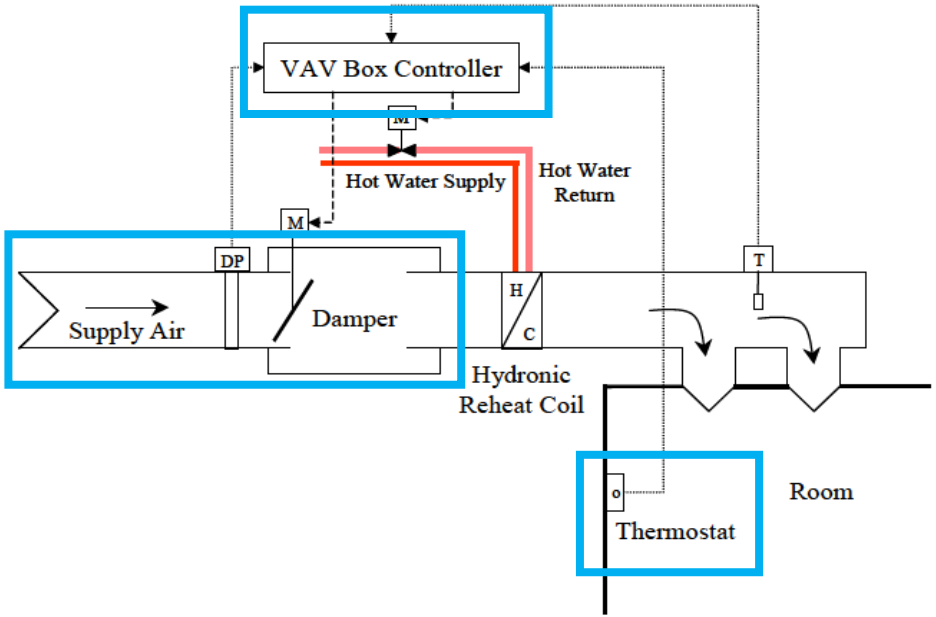
Improve Energy Efficiency of the building, by optimizing the startup of the **Variable Air Volume Cooling (VAVCO)** units



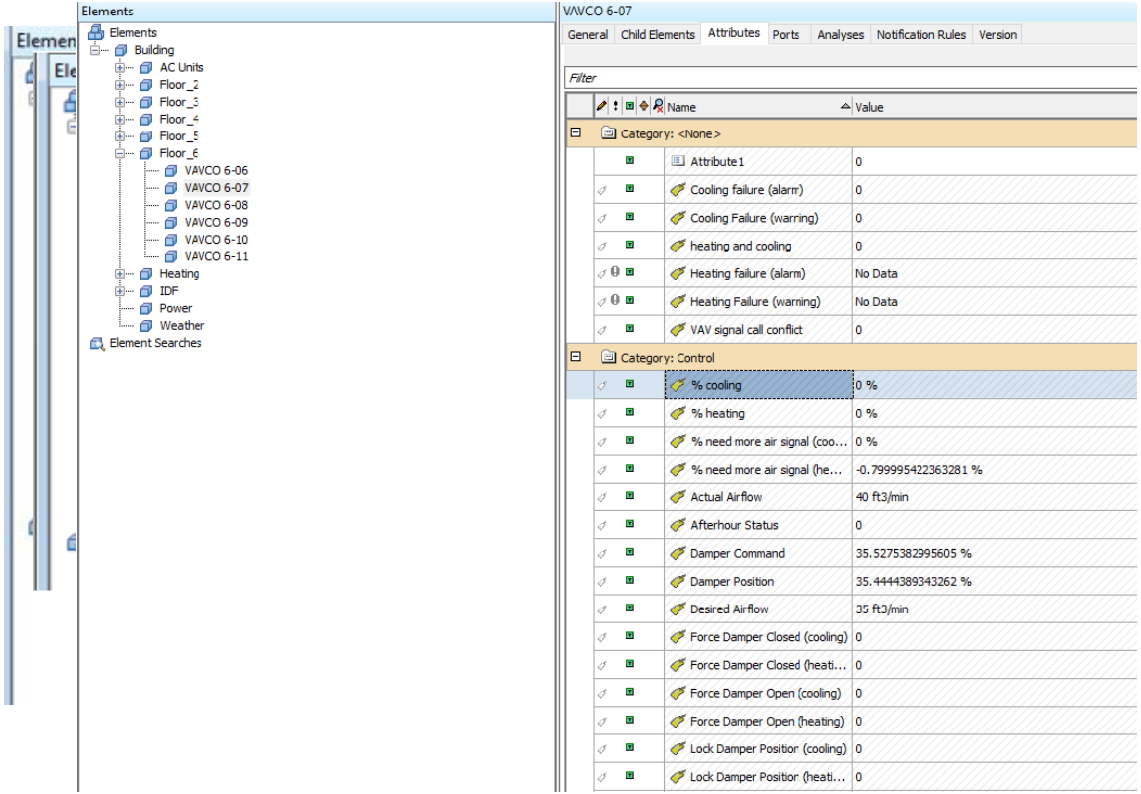
Value

- "If you saved 1 hour a day for 261 Working Days that is a significant amount of energy" – Subject Matter Expert

How does a VAV unit operate?

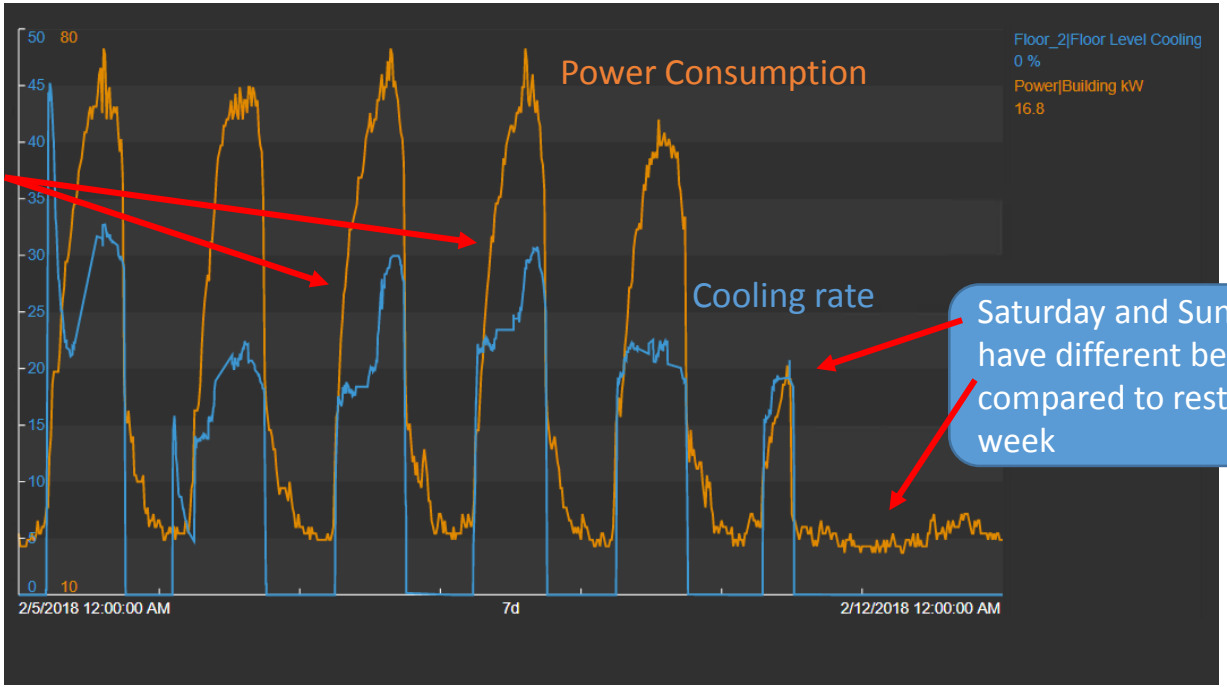


Available Data



Data understanding – Power Consumption

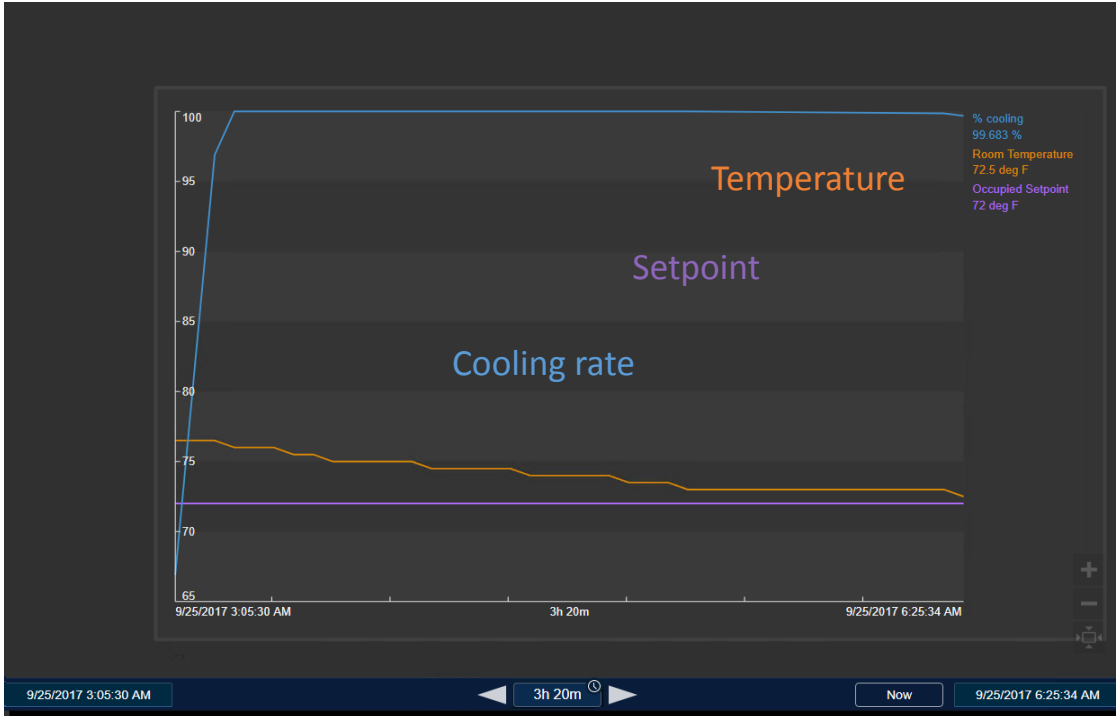
Daily power usage aligned with **Average % Cooling**



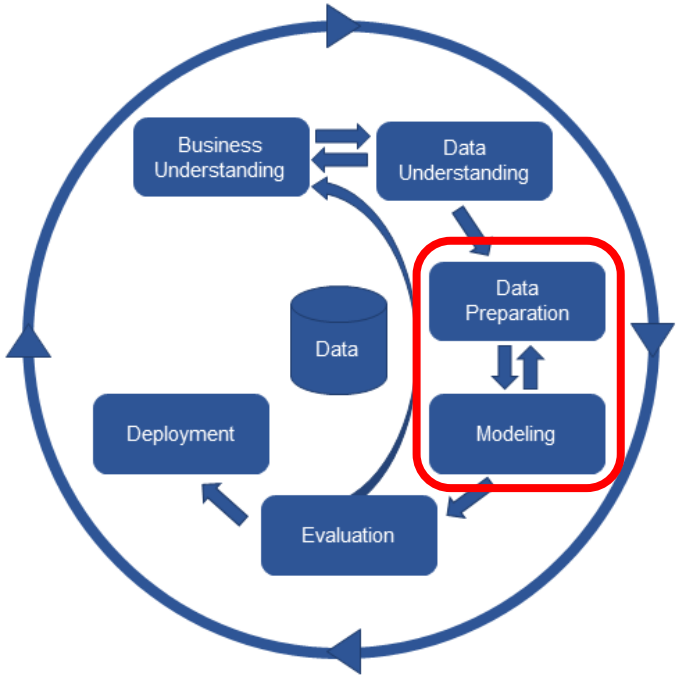
Saturday and Sunday have different behavior, compared to rest of the week

Data understanding – Daily Cooling operation

- Morning startup event
- Feature extraction with Event Frames
- Predict Duration of Event Frame



Data Preparation and Modelling



- How do I access PI Data from Advanced Analytics platforms?
- Which features are relevant for model prediction and can I identify any relationships between them?
- Which model should I use?

Client libraries for Data Science (Python & R)

- Client libraries for R and Python available in GitHub
  - https://github.com/osimloeff/PI-Web-API-Client-R
  - https://github.com/osimloeff/PI-Web-API-Client-Python
- Jupyter Notebooks - Data Exploration with Python
- Asset Analytics - MATLAB integration

jupyter Building Power Consumption example Last Checkpoint: 03/15/2018 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

Create the PI Web API connection

```
In [ ]: client = PIWebApiClient("https://localhost/piwebapi", False, "student01", "P1school!", True)
```

Load the Power data and do some transformation and feature creation

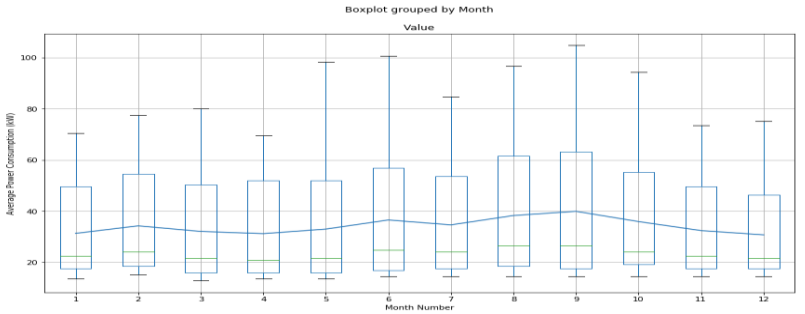
```
In [ ]: PowerData = client.data.get_recorded_values("af:\\\\PISRV01\\Lab Building Data\\Building\\Power|Building kW", None, None, '*',
                                                    None, None, 50000, "items.value;items.timestamp", "1-mar-2017", None)
        PowerData.dtypes
```

Perform data type transformation and feature creation

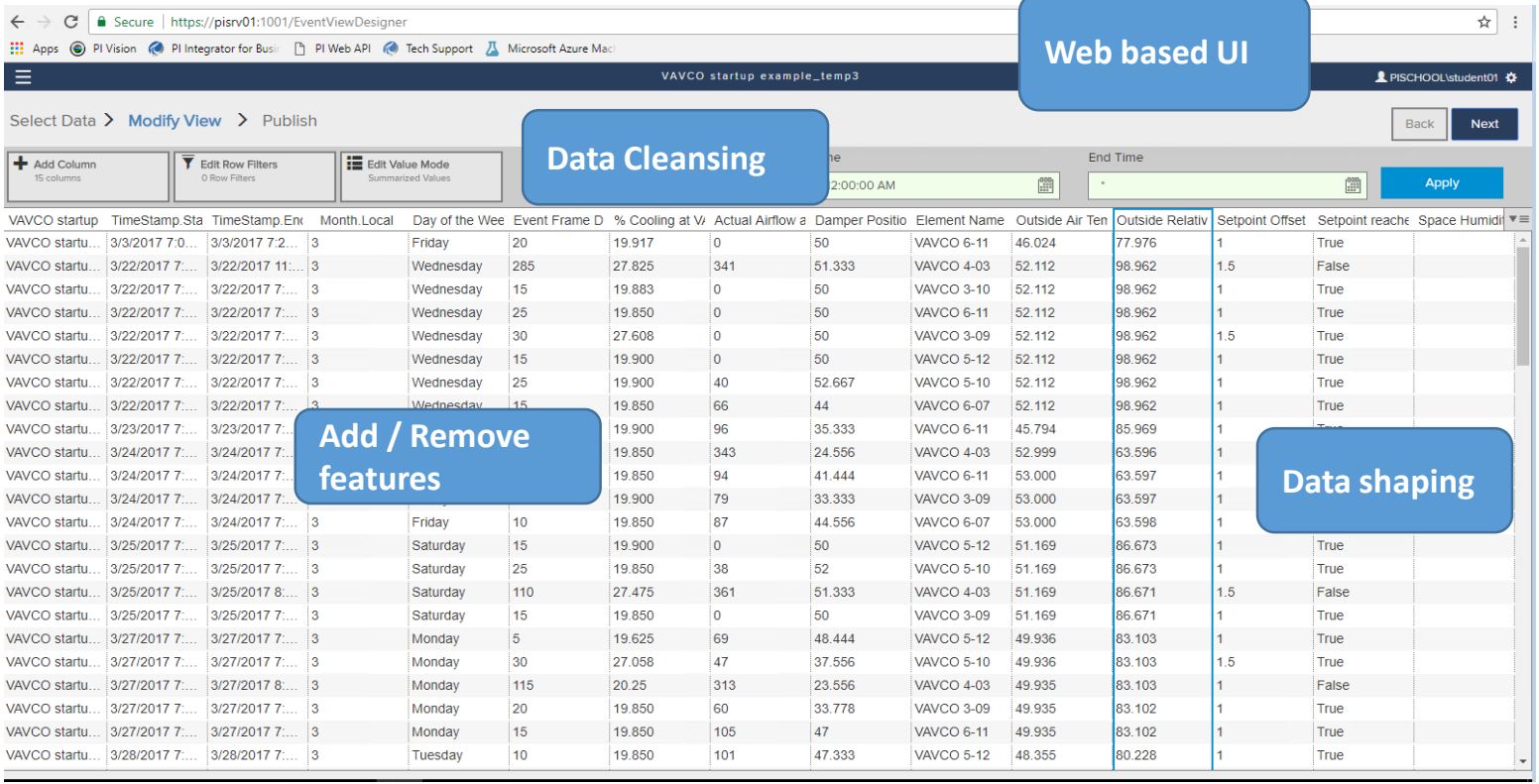
```
In [ ]: PowerData.Timestamp = pd.to_datetime(PowerData.Timestamp)
        PowerData['Month'] = PowerData.Timestamp.dt.month
        PowerData.head()
```

Boxplot of Average Power consumption per Month

```
In [ ]: PowerData.boxplot(by='Month', figsize = (14,8))
        plt.plot(PowerData.groupby('Month').mean())
        plt.ylabel('Average Power Consumption (kW)')
        plt.xlabel('Month Number')
        plt.show()
```



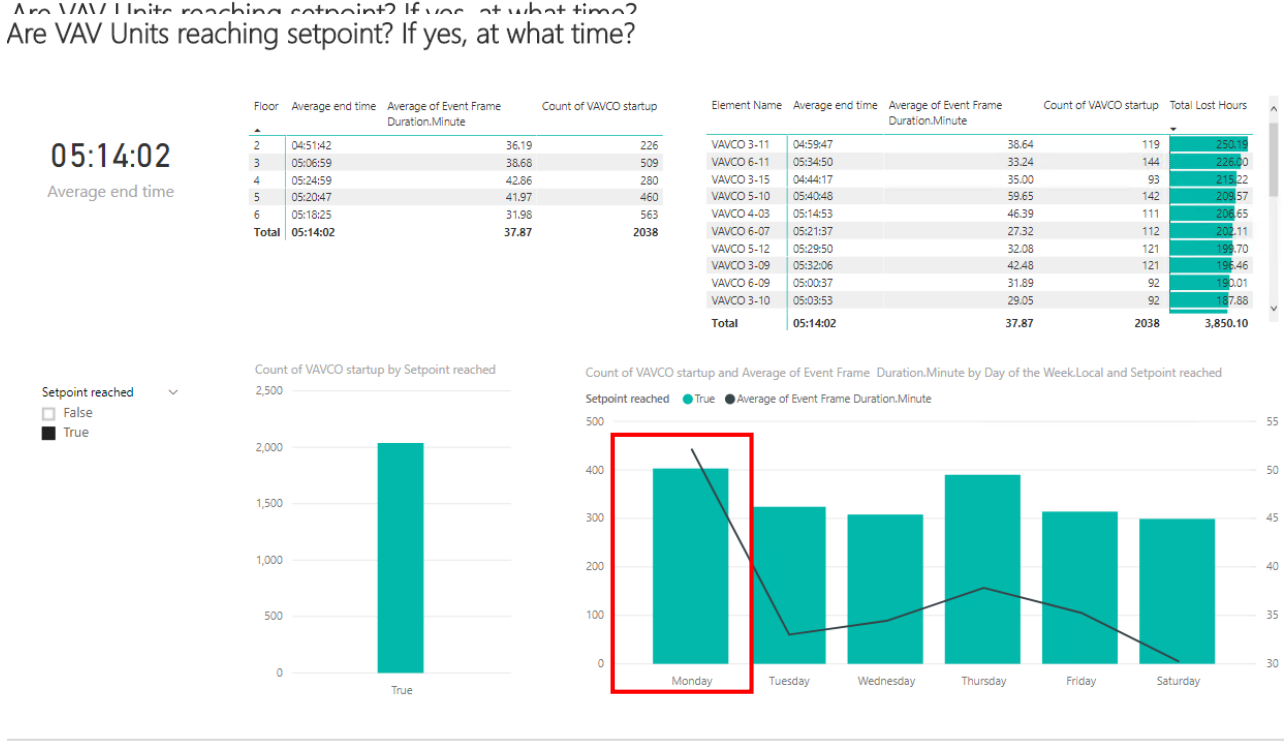
Event Frames published using PI Integrator for Business Analytics



VAVCO startup	TimeStamp.Sta	TimeStamp.En	Month.Local	Day of the Wee	Event Frame D	% Cooling at V	Actual Airflow a	Damper Positio	Element Name	Outside Air Ten	Outside Relativ	Setpoint Offset	Setpoint reache	Space Humidi
VAVCO startu...	3/3/2017 7:0...	3/3/2017 7:2...	3	Friday	20	19.917	0	50	VAVCO 6-11	46.024	77.976	1	True	
VAVCO startu...	3/22/2017 7:...	3/22/2017 11:...	3	Wednesday	285	27.825	341	51.333	VAVCO 4-03	52.112	98.962	1.5	False	
VAVCO startu...	3/22/2017 7:...	3/22/2017 7:...	3	Wednesday	15	19.883	0	50	VAVCO 3-10	52.112	98.962	1	True	
VAVCO startu...	3/22/2017 7:...	3/22/2017 7:...	3	Wednesday	25	19.850	0	50	VAVCO 6-11	52.112	98.962	1	True	
VAVCO startu...	3/22/2017 7:...	3/22/2017 7:...	3	Wednesday	30	27.608	0	50	VAVCO 3-09	52.112	98.962	1.5	True	
VAVCO startu...	3/22/2017 7:...	3/22/2017 7:...	3	Wednesday	15	19.900	0	50	VAVCO 5-12	52.112	98.962	1	True	
VAVCO startu...	3/22/2017 7:...	3/22/2017 7:...	3	Wednesday	25	19.900	40	52.667	VAVCO 5-10	52.112	98.962	1	True	
VAVCO startu...	3/22/2017 7:...	3/22/2017 7:...	3	Wednesday	15	19.850	66	44	VAVCO 6-07	52.112	98.962	1	True	
VAVCO startu...	3/23/2017 7:...	3/23/2017 7:...				19.900	96	35.333	VAVCO 6-11	45.794	85.969	1		
VAVCO startu...	3/24/2017 7:...	3/24/2017 7:...				19.850	343	24.556	VAVCO 4-03	52.999	63.596	1		
VAVCO startu...	3/24/2017 7:...	3/24/2017 7:...				19.850	94	41.444	VAVCO 6-11	53.000	63.597	1		
VAVCO startu...	3/24/2017 7:...	3/24/2017 7:...				19.900	79	33.333	VAVCO 3-09	53.000	63.597	1		
VAVCO startu...	3/24/2017 7:...	3/24/2017 7:...	3	Friday	10	19.850	87	44.556	VAVCO 6-07	53.000	63.598	1		
VAVCO startu...	3/25/2017 7:...	3/25/2017 7:...	3	Saturday	15	19.900	0	50	VAVCO 5-12	51.169	86.673	1	True	
VAVCO startu...	3/25/2017 7:...	3/25/2017 7:...	3	Saturday	25	19.850	38	52	VAVCO 5-10	51.169	86.673	1	True	
VAVCO startu...	3/25/2017 7:...	3/25/2017 8:...	3	Saturday	110	27.475	361	51.333	VAVCO 4-03	51.169	86.671	1.5	False	
VAVCO startu...	3/25/2017 7:...	3/25/2017 7:...	3	Saturday	15	19.850	0	50	VAVCO 3-09	51.169	86.671	1	True	
VAVCO startu...	3/27/2017 7:...	3/27/2017 7:...	3	Monday	5	19.625	69	48.444	VAVCO 5-12	49.936	83.103	1	True	
VAVCO startu...	3/27/2017 7:...	3/27/2017 7:...	3	Monday	30	27.058	47	37.556	VAVCO 5-10	49.936	83.103	1.5	True	
VAVCO startu...	3/27/2017 7:...	3/27/2017 8:...	3	Monday	115	20.25	313	23.556	VAVCO 4-03	49.935	83.103	1	False	
VAVCO startu...	3/27/2017 7:...	3/27/2017 7:...	3	Monday	20	19.850	60	33.778	VAVCO 3-09	49.935	83.102	1	True	
VAVCO startu...	3/27/2017 7:...	3/27/2017 7:...	3	Monday	15	19.850	105	47	VAVCO 6-11	49.935	83.102	1	True	
VAVCO startu...	3/28/2017 7:...	3/28/2017 7:...	3	Tuesday	10	19.850	101	47.333	VAVCO 5-12	48.355	80.228	1	True	

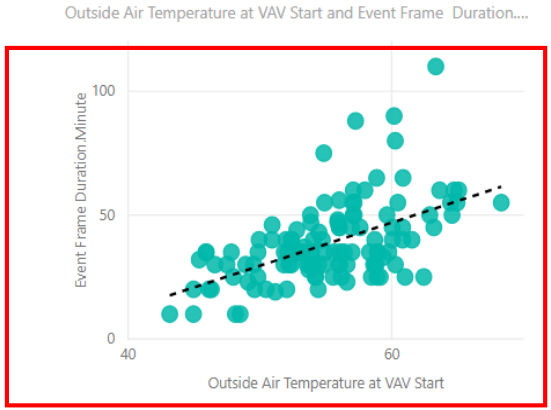
Data Exploration – Microsoft Power BI

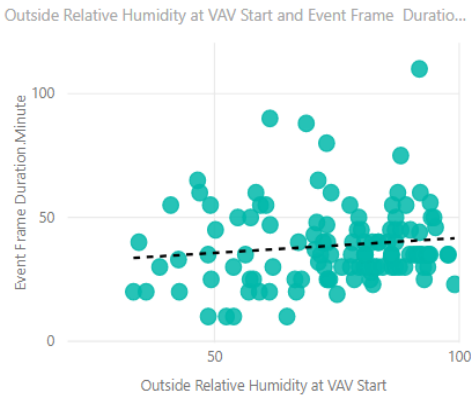
- **Target** Average end time is **7 AM** ➔ **1.5 hours** of wasted Energy
- **Total Lost Hours** points to worst performing units
- **Monday** has significantly higher Average Duration

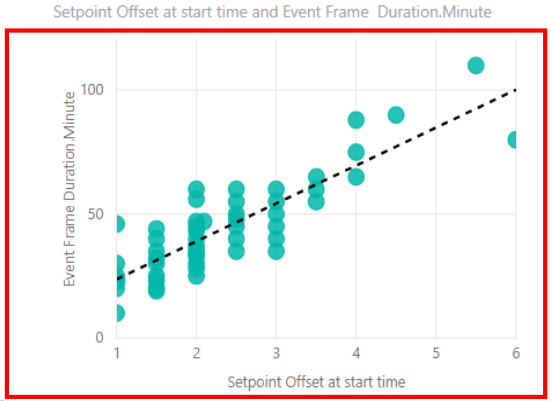


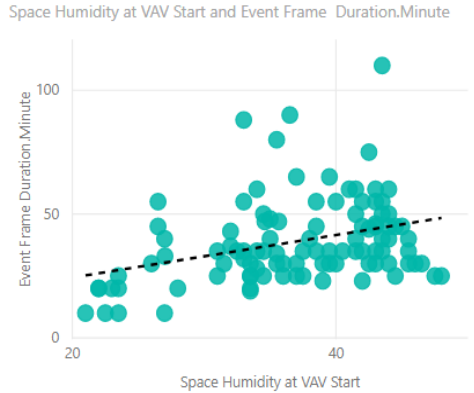
Bivariate Analysis

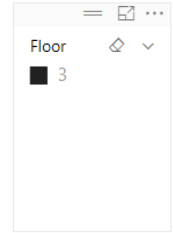
- No clear relationships when looking at whole dataset
- Linear relationship between **Duration – Setpoint Offset** on floor level
- Linear relationship becomes more clear for specific units







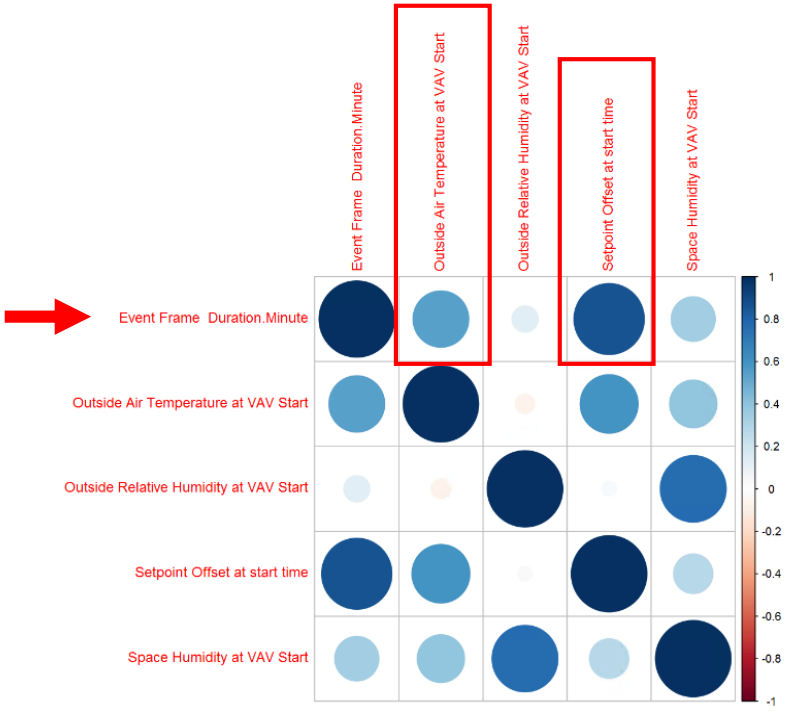




Correlation plot embedded in Power BI

Confirms Bivariate Analysis Results

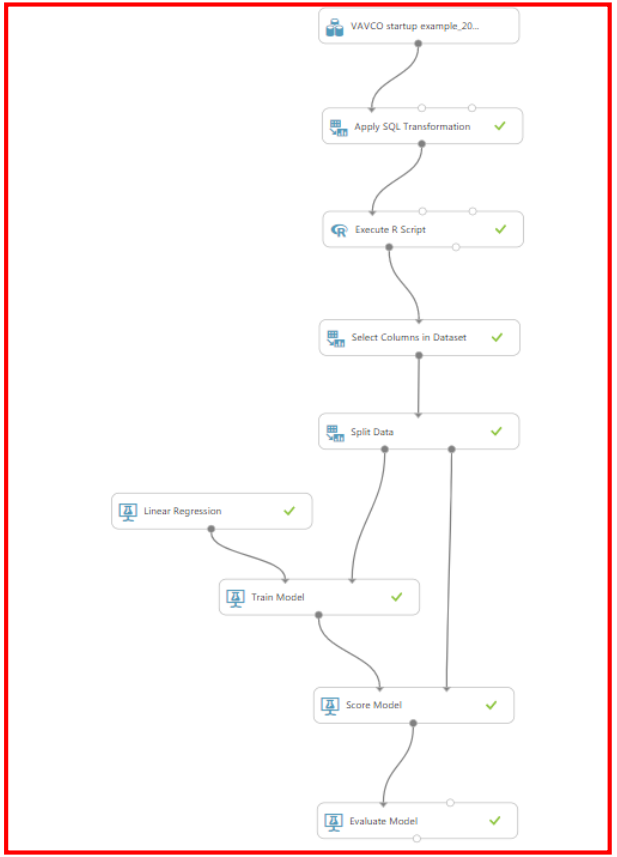
Strong correlation of **Duration** with **Setpoint Offset** and **Outside Temperature**



Floor
3

Element Name
VAVCO 2-03
VAVCO 2-09
VAVCO 2-11
VAVCO 2-13
VAVCO 3-09
VAVCO 3-10
VAVCO 3-11
VAVCO 3-15
VAVCO 3-16
VAVCO 4-03
VAVCO 4-15
VAVCO 4-16
VAVCO 4-17

Azure ML model

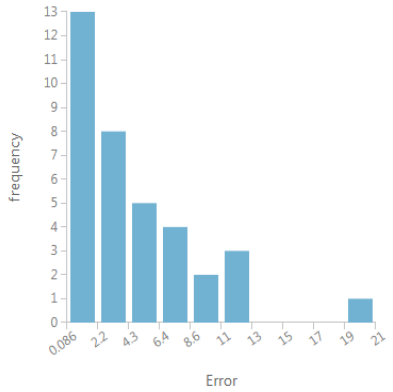


Results

Metrics

Mean Absolute Error	4.753966
Root Mean Squared Error	6.450828
Relative Absolute Error	0.418556
Relative Squared Error	0.14538
Coefficient of Determination	0.85462

Error Histogram

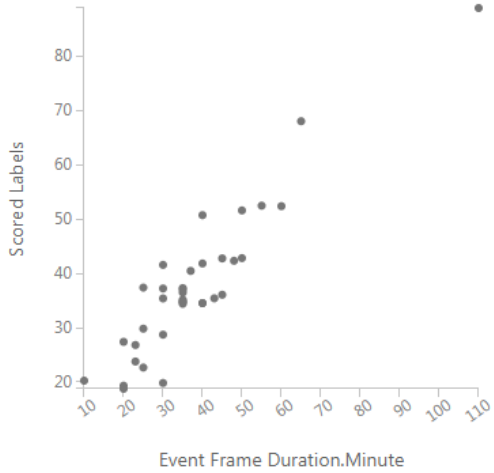


Visualizations

Event Frame Duration.Minute

ScatterPlot

compare to Scored Labels



Deployment

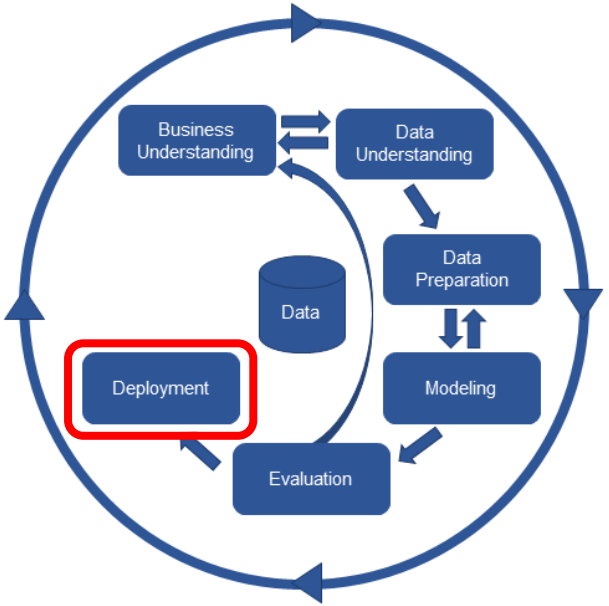
Productizing the model

- Have the model control the wakeup time of the building

Simpler models can be deployed in PI; some control models are built into the control network

- Regression or small Decision Trees can be efficiently deployed in AF
- Larger or more complex models need to go outside PI

Work in progress



Main Takeaways

Data Science delivers **Value**

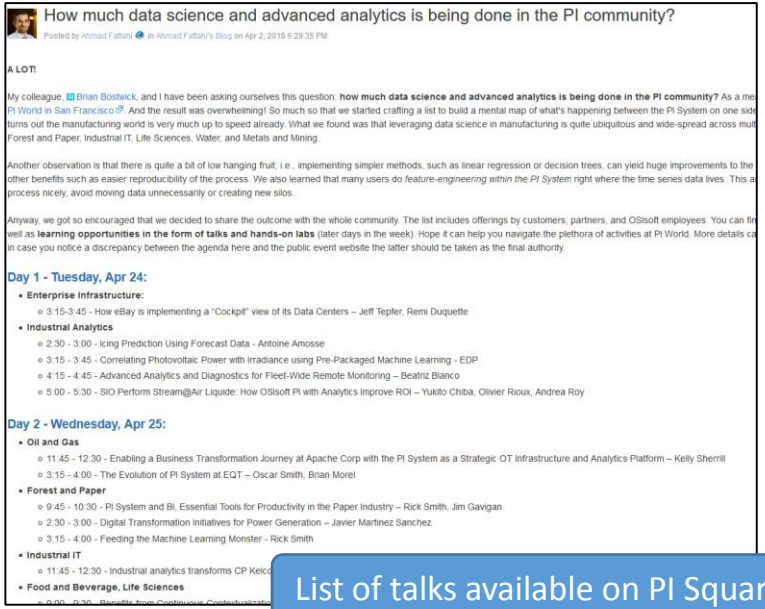
Learn the process

Iterative process – Don’t get discouraged

Leverage PI System And Subject Matter Expertise

Keep on learning!

- PI World presentations
- Labs and online courses
- Talk to other users, partners, and us



How much data science and advanced analytics is being done in the PI community?

Posted by Ahmad Fattahi in Ahmad Fattahi's Blog on Apr 2, 2018 6:29:35 PM

A LOT!

My colleague, Brian Bostwick, and I have been asking ourselves this question: **how much data science and advanced analytics is being done in the PI community?** As a me
PI World in San Francisco. And the result was overwhelming! So much so that we started crafting a list to build a mental map of what's happening between the PI System on one side
turns out the manufacturing world is very much up to speed already. What we found was that leveraging data science in manufacturing is quite ubiquitous and wide-spread across mult
Forest and Paper, Industrial IT, Life Sciences, Water, and Metals and Mining.

Another observation is that there is quite a bit of low hanging fruit, i.e., implementing simpler methods, such as linear regression or decision trees, can yield huge improvements to the
other benefits such as easier reproducibility of the process. We also learned that many users do *feature-engineering within the PI System* right where the time series data lives. This a
process nicely, avoid moving data unnecessarily or creating new silos.

Anyway, we got so encouraged that we decided to share the outcome with the whole community. The list includes offerings by customers, partners, and OSIsoft employees. You can fir
well as **learning opportunities in the form of talks and hands-on labs** (later days in the week). Hope it can help you navigate the plethora of activities at PI World. More details ca
in case you notice a discrepancy between the agenda here and the public event website the latter should be taken as the final authority.

Day 1 - Tuesday, Apr 24:

- **Enterprise Infrastructure:**
  - 3:15-3:45 - How eBay is implementing a "Cockpit" view of its Data Centers – Jeff Tepfer, Remi Duquette
- **Industrial Analytics**
  - 2:30 - 3:00 - Icing Prediction Using Forecast Data - Antoine Amosse
  - 3:15 - 3:45 - Correlating Photovoltaic Power with Irradiance using Pre-Packaged Machine Learning - EDP
  - 4:15 - 4:45 - Advanced Analytics and Diagnostics for Fleet-Wide Remote Monitoring – Beatriz Blanco
  - 5:00 - 5:30 - SIO Perform Stream@Air Liquide: How OSIsoft PI with Analytics Improve ROI – Yukito Chiba, Olivier Rioux, Andrea Roy

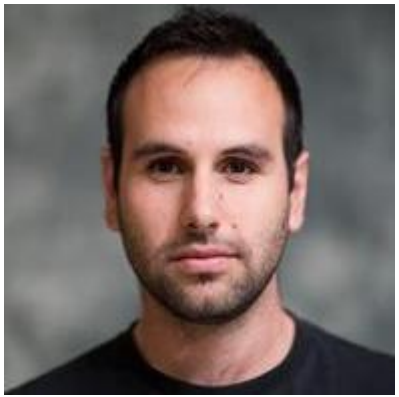
Day 2 - Wednesday, Apr 25:

- **Oil and Gas**
  - 11:45 - 12:30 - Enabling a Business Transformation Journey at Apache Corp with the PI System as a Strategic OT Infrastructure and Analytics Platform – Kelly Sherrill
  - 3:15 - 4:00 - The Evolution of PI System at EQT – Oscar Smith, Brian Morel
- **Forest and Paper**
  - 9:45 - 10:30 - PI System and BI, Essential Tools for Productivity in the Paper Industry – Rick Smith, Jim Gavigan
  - 2:30 - 3:00 - Digital Transformation Initiatives for Power Generation – Javier Martinez Sanchez
  - 3:15 - 4:00 - Feeding the Machine Learning Monster - Rick Smith
- **Industrial IT**
  - 11:45 - 12:30 - Industrial analytics transforms CP Kelco
- **Food and Beverage, Life Sciences**

List of talks available on PI Square

bit.ly/DSPIWorld18

Ahmad Fattahi
afattahi@osisoft.com
Manager, Data Science Enablement
OSIsoft



Kleanthis Mazarakis
kmazarakis@osisoft.com
Data Scientist
OSIsoft

Questions

Please wait for the **microphone** before asking your questions

State your **name & company**

Please remember to…

Complete the Online Survey for this session

Merci

谢谢

Danke

Спасибо

Gracias

Thank You

감사합니다

ありがとう

Grazie

Obrigado