info|pro

Advanced Analytics Solutions

TH Rosenheim

DATA SCIENCE IN PRACTICE

05 Dec 2022

InfoPro

info|pro
Advanced Analytics Solutions

- Founded 2001 in Freilassing
- Broad experience in a range of IT services: Project Management, Risk Management, Change Management, Software Development, Documentation
- More than 10 years of consulting and software development for digitization efforts, with a focus on machine learning and AI, including platform product
- Currently exclusively active for regional industry and retail, primarily with Advanced Analytics.

What is Advanced Analytics?

info|pro
Advanced Analytics Solutions

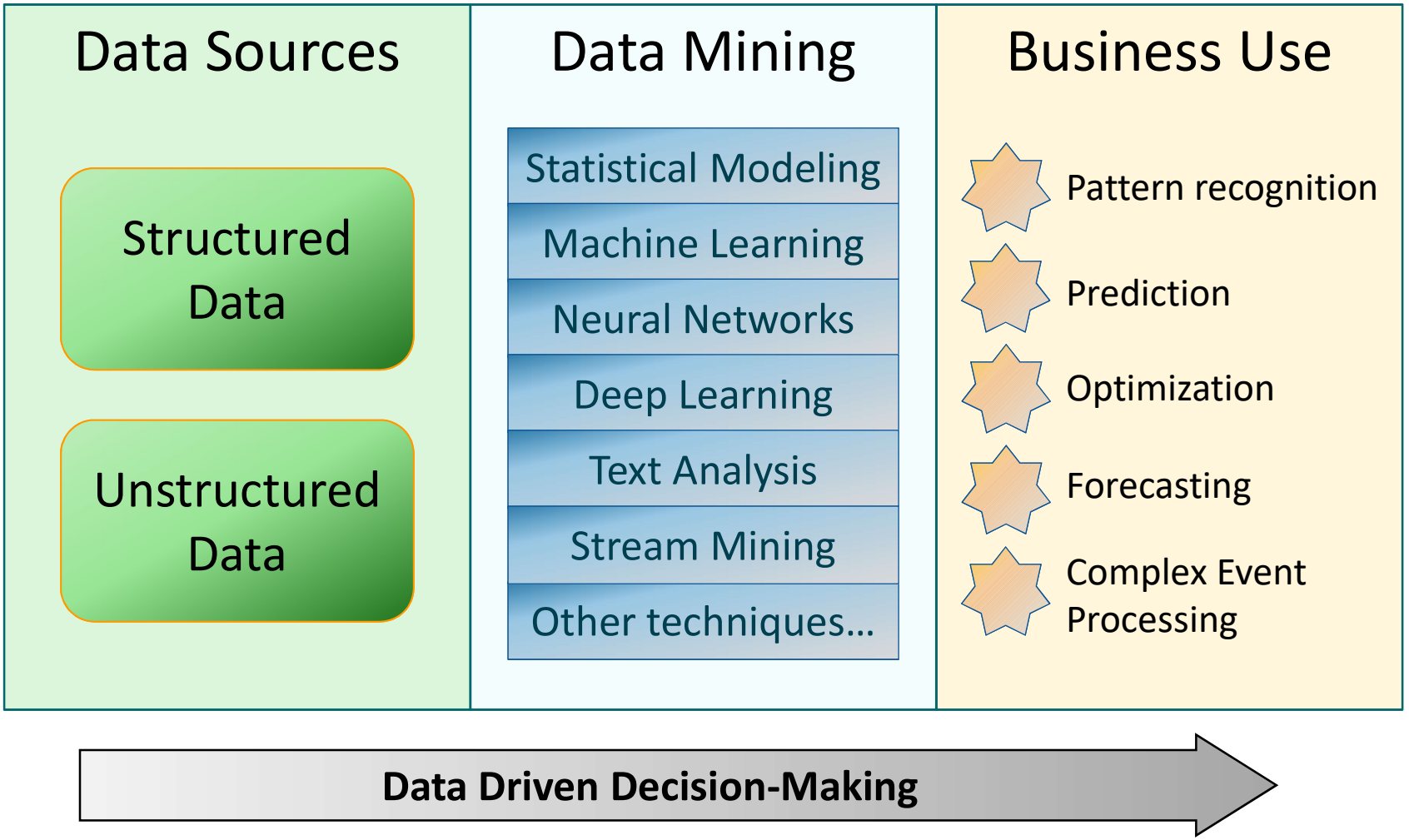
"Advanced analytics provides algorithms for complex analysis of either structured or unstructured data. It includes sophisticated statistical models, machine learning, neural networks, text analytics, and other advanced data mining techniques.

Among its many use cases, it can be deployed to find patterns in data, prediction, optimization, forecasting, and for complex event processing/analysis.

Examples include predicting churn, identifying fraud, market basket analysis, or understanding website behavior. Advanced analytics does not include database query and reporting and OLAP cubes."

- *Dr. Fern Halper / TDWI Research Director Advanced Analytics, 2010 (https://fbhalper.wordpress.com/2010/12/20/what-is-advanced-analytics/)*

Advanced Analytics



TH Innovation Lab SS 2022: mobileGLAAS

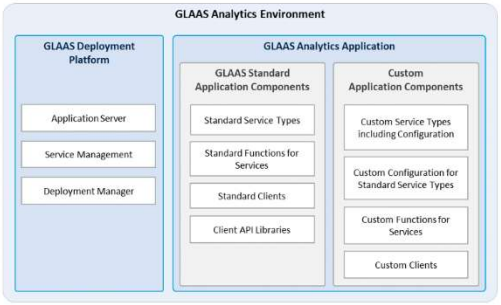
info|pro
Advanced Analytics Solutions

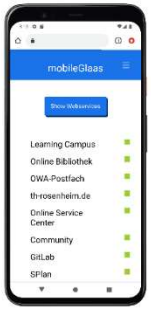
mobileGLAAS

InfoPro GLAAS - die Predicition Toolbox

Entwicklung und Deployment perfekt orchestrierter, skalierbarer Advanced Analytics-Lösungen:

- preiswerter Zugang zur vollen Machine Learning-Bandbreite
- einfache Entwicklung sich selbst anpassender Machine Learning-Anwendungen
- schnelle Umsetzung durch konfigurierbare Standard-Module (Low-Code-Ansatz):
  - → Übernahme von Basisdaten
  - → Vorbereitung der Daten für die Analyse
  - → Erstellen von Machine Learning-Modellen
  - → Durchführung von Analysen
  - → Verarbeitung und Bereitstellung von Ergebnissen und Prognosen
- mühelose Erweiterung um individuelle Komponenten
- dynamische Konfiguration, Ausführung und Monitoring von Services
- für beliebige Hardware-Umgebungen, von Edge-Computing bis zur Cloud





info|pro
Advanced Analytics Solutions

Technische Hochschule Rosenheim

Lessons Learned the Hard Way

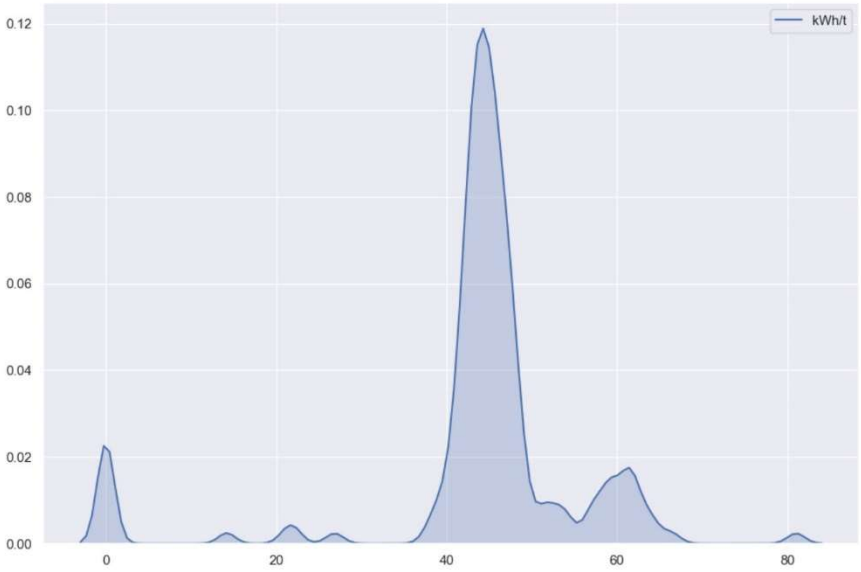
info|pro
Advanced Analytics Solutions

- You need tools to see.
- It's never about the obvious stuff.
- Let the data talk; data preparation is key.
- Your advantage is in not having a clue.
- Learn to love correlation.
- Causation is for children; the best you get is influence.
- Models are easy, interpretation is hard.
- For business, the real challenge is deployment.
- Good enough is good enough.

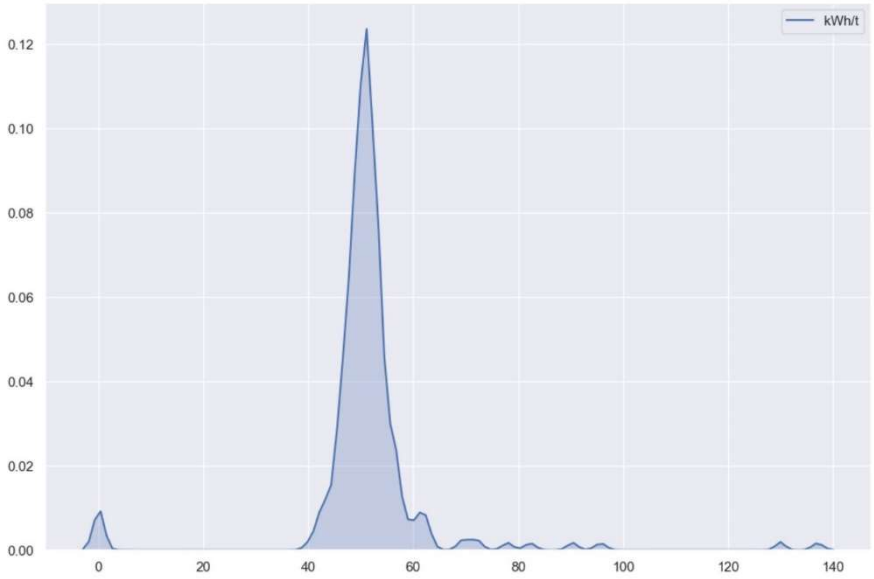
What Do Distributions Look Like?

info|pro
Advanced Analytics Solutions

2018



2022



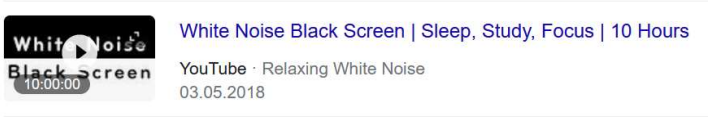
Distribution of daily total energy per ton of cement (kWh/t), before and after energy saving measures.

Was there an improvement?

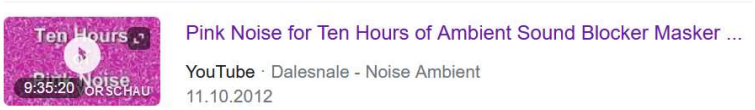
What Do Distributions Sound Like?

info|pro
Advanced Analytics Solutions

White Noise: https://www.youtube.com/watch?v=nMfPqeZjc2c



Pink Noise: https://www.youtube.com/watch?v=ZXtimhT-ff4

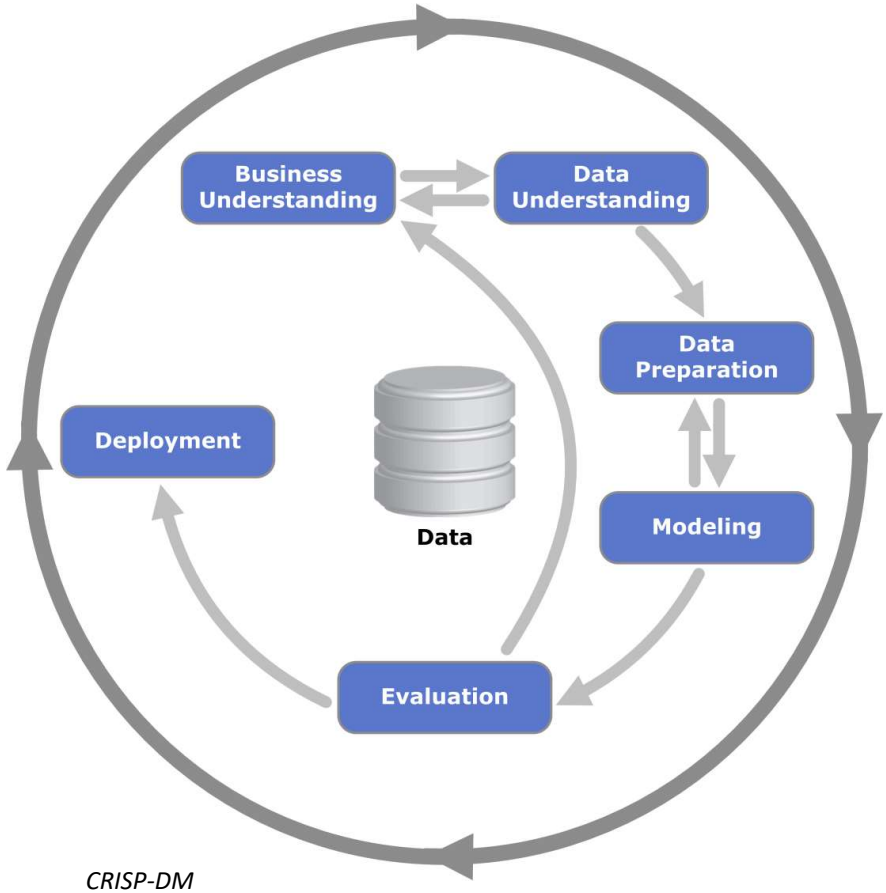


… and what do distributions feel like?

How a Data Scientist Works

info|pro
Advanced
Analytics
Solutions

A Data Mining Project Needs a Process



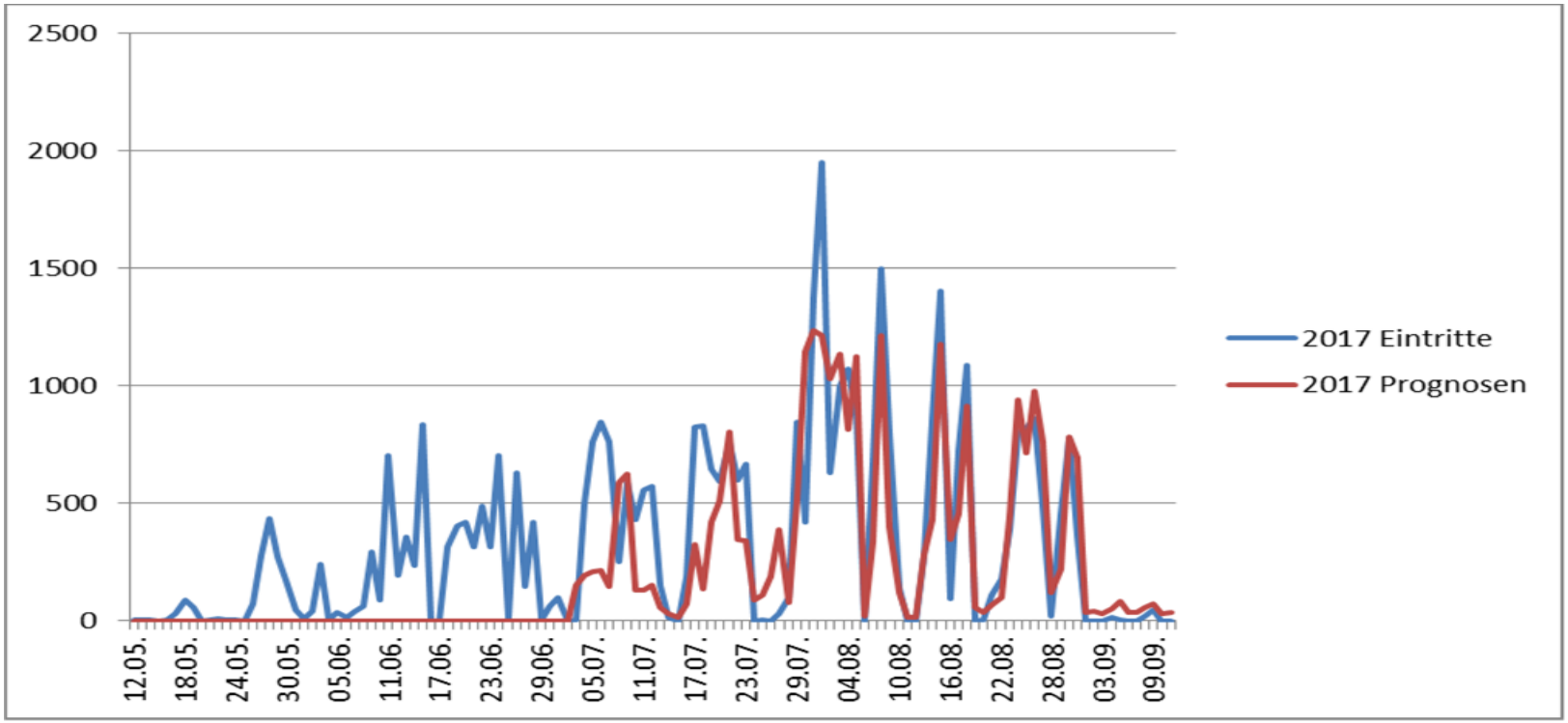
CRISP-DM

Three distinct types of discussions:

- Data Talks
- Analysis Talks
- Technical Talks

How a Data Scientist Works

info|pro
Advanced Analytics Solutions

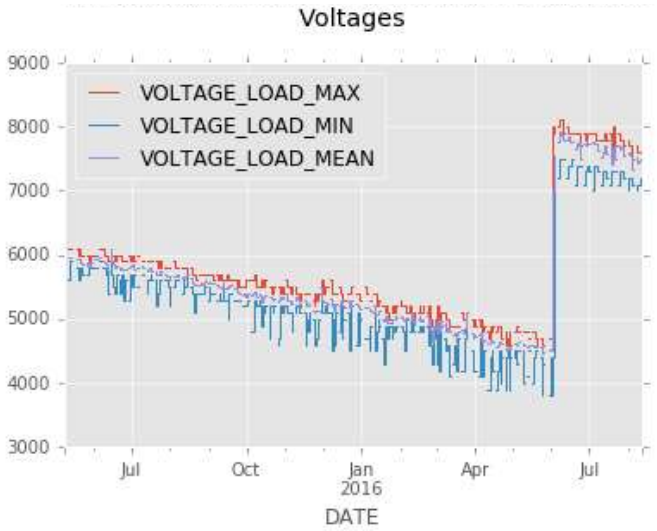


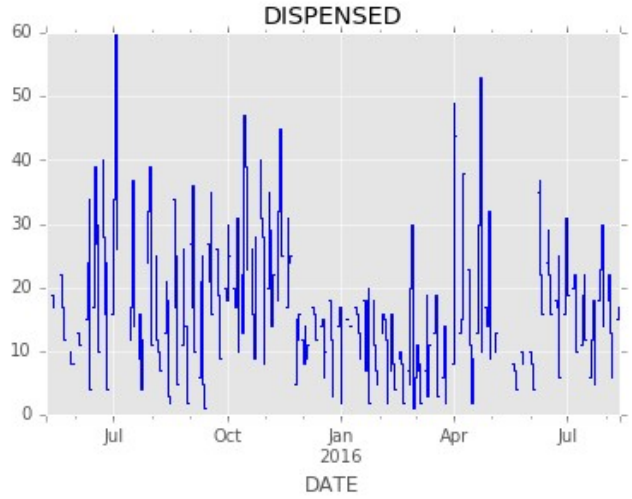
info|pro
Advanced
Analytics
Solutions

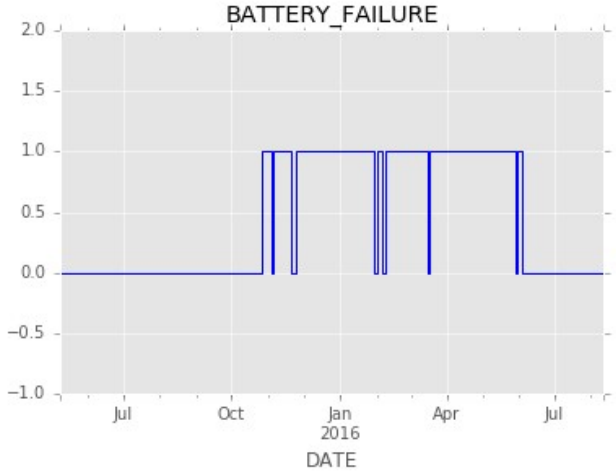
Example: Washroom Technology

Device Measurements

info|pro
Advanced Analytics Solutions







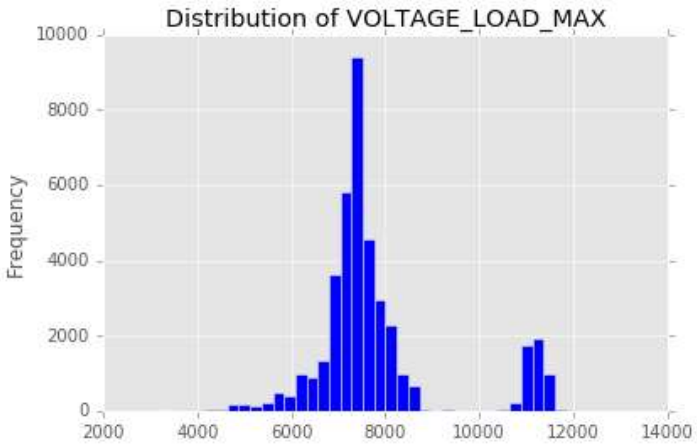
Prediction of Battery Failures

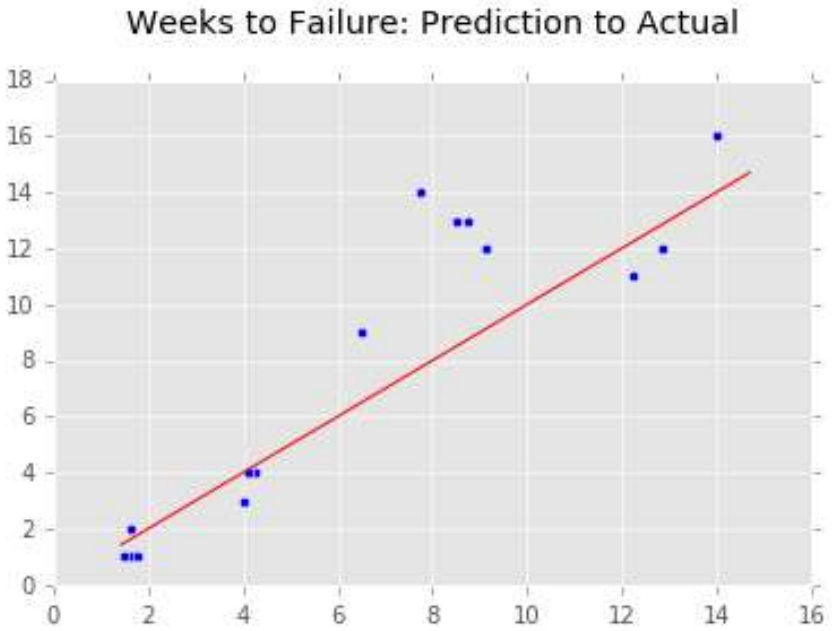
info|pro
Advanced
Analytics
Solutions

- Number of "Device-Days": 127.144
- Number of failed batteries: 61
- Challenges:
  => many "gaps" (days without values)
  => very few (real) battery failures
- The predictions are very good

Value Distribution of Failure Prognosis

info|pro
Advanced Analytics Solutions





info|pro
Advanced
Analytics
Solutions

Example: Transport Logistics

Predictions of Refrigeration Failures

info|pro
Advanced Analytics Solutions

- Developed model which predicts 95% of refrigeration failures during the next trip
- Real-time monitoring of telematics data: temperature, doors, operation times and states, refrigeration settings, etc.
- Valuable information for: vehicle disposition, vehicle planning, maintenance, asset management
- Advantages: increased availability of vehicles, lower operational costs, less legal risk, lower insurance costs

Training and Evaluation

info|pro
Advanced Analytics Solutions

- Algorithm: ADA-Boosted Random Forest Classifier
- Data Overview (one month of data):
  - Boxes: 15.397
  - Events: 112.209.909
  - On-Phases: ca. 1.302.167
- Modeling Challenges:
  - strongly unbalanced distribution NoAlert/Alert
  - large amount of missing or incorrect data
  - chosen algorithm can be very sensitive to small changes in the data

Data Preparation

info|pro
Advanced Analytics Solutions

- Calculate Training Vectors:
  - "Clean" the raw data (type conversions, filter or correction of incorrect data, etc.)
  - Derive new features
  - Determine operational phases:
    On Phases, Driving Phases, Alert Phases
  - Extract extensive features (total of 85) about the On Phases, including "Label": ALERT_IN_NEXT
- Data Quality Challenges:
  - Incorrect data which falsifies the results
    (e.g. voltage 1000 V, air temperature -800°C).
    One Box was ignored due to incorrect data.
  - Deciding how to best interpret missing data

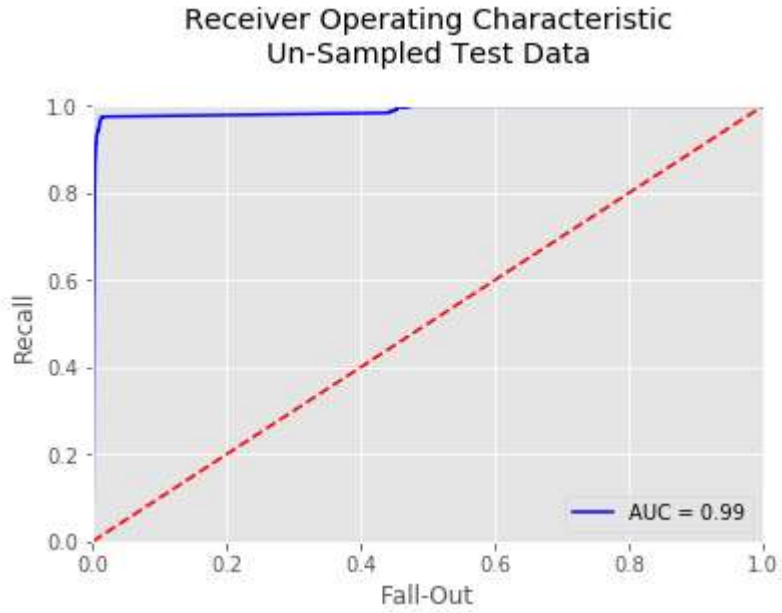
Analysis Results

Counts: 15.397 Boxes in January 2017, On-Phases: 1.302.167
Split of un-sampled data for training/test: 75:25
Sample of NoAlerts for training only: 1%

Confusion Matrix: Un-sampled test data

	Pred. False	Pred. True
Actual False	321634	3345
Actual True	25	538

```
Accuracy:                          0.990
Recall    => tp / (tp + fn):       0.956
Fall-Out  => fp / (fp + tn):       0.010
Precision => tp / (tp + fp):       0.139
NPV       => tn / (tn + fn):       0.999922
```



Hit-Rate: 95,6% (Recall)
Precision Strength: 80,11 (Precision / Probability of Actual True)

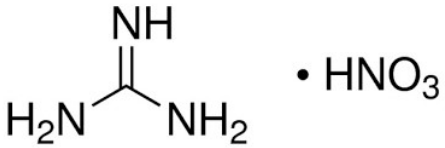
info|pro
Advanced
Analytics
Solutions

Example: Crystallizer

What does a Crystallizer look like?

Guanidine Nitrate

info|pro
Advanced
Analytics
Solutions



Motivation

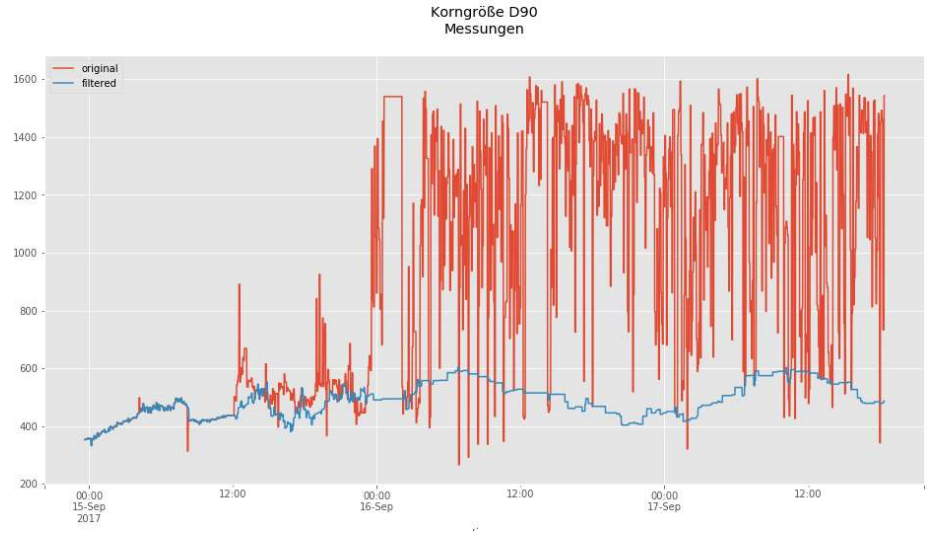
info|pro
Advanced Analytics Solutions

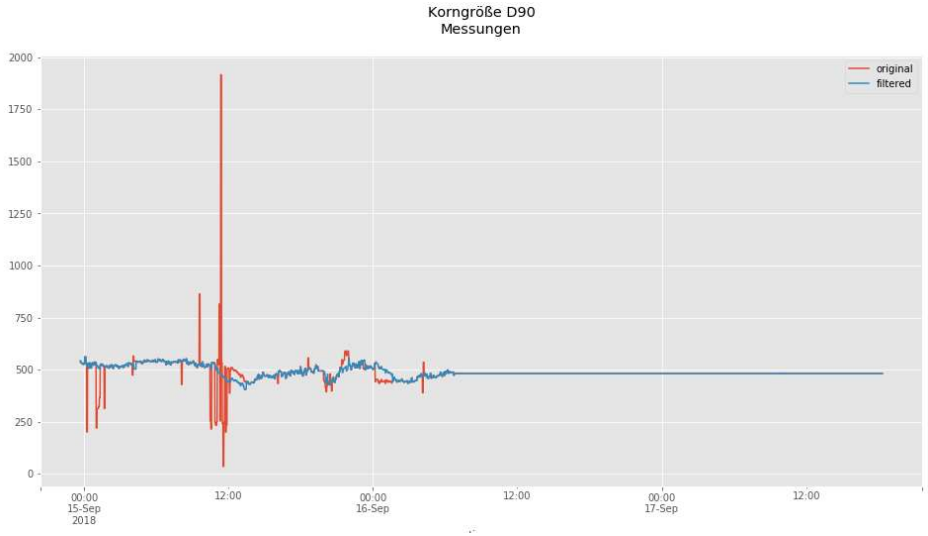
Guanidine nitrate is a salt. The grain size and its distribution are the principle quality attributes:

- How can the quality be improved?
  - Determine factors influencing grain size and distribution
  - Predict grain size and distribution
- Important Issues
  - Data sources, data quality, metric quality
  - Integration of prognoses in daily operations
  - Prognosis quality: what is "good enough"?
  - Assessment of "influence" in light of correlation

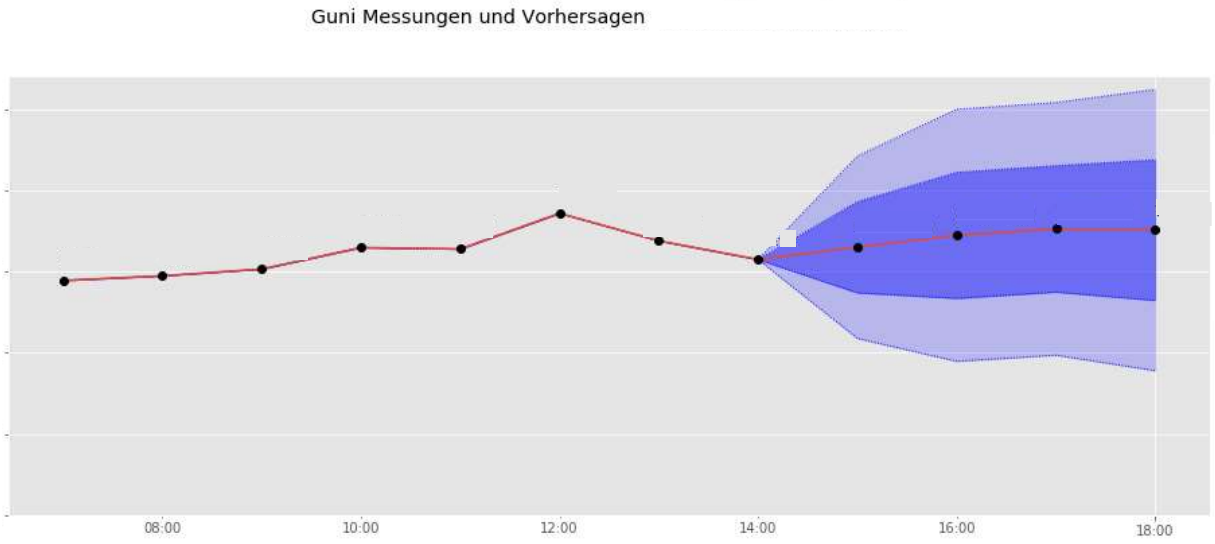
Challenges: Bad Metrics, Missing Values

info|pro
Advanced
Analytics
Solutions



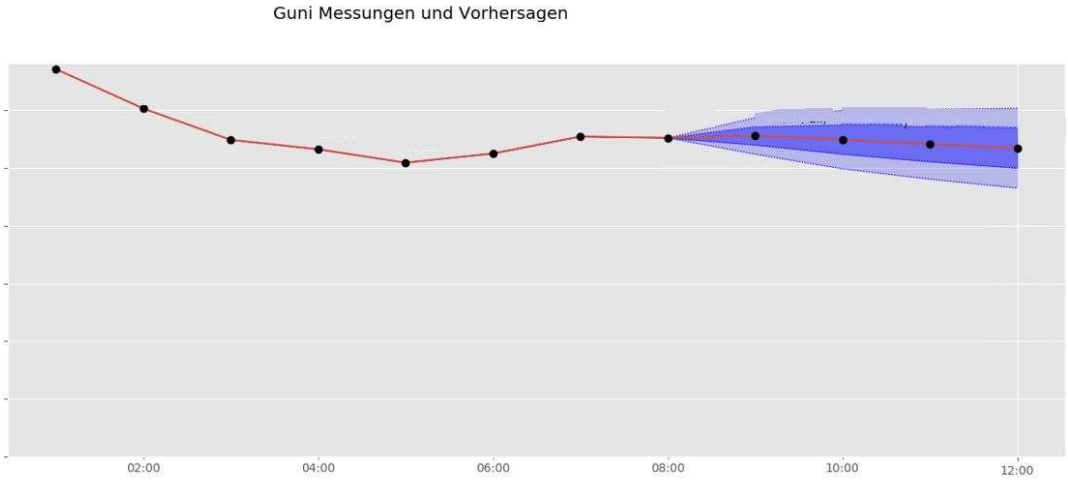


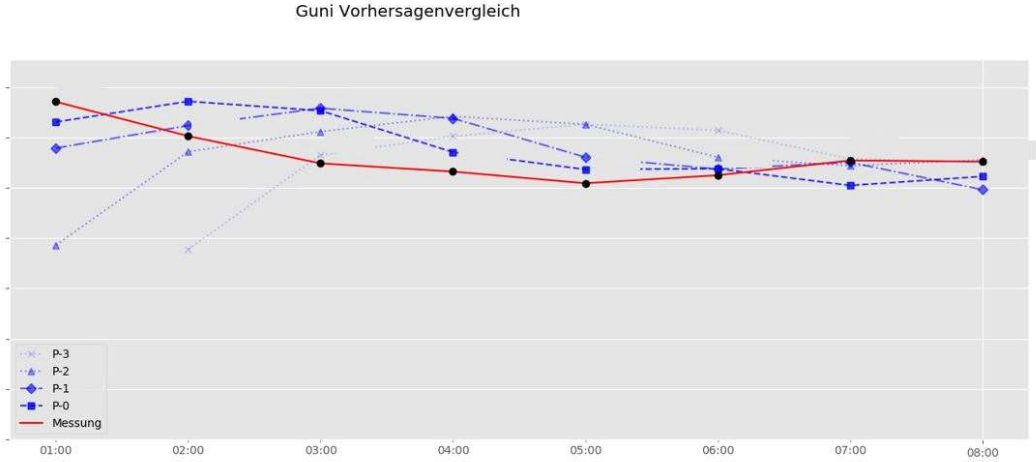
Challenges: Poor Prognosis Quality



Results

info|pro
Advanced
Analytics
Solutions





info|pro
Advanced
Analytics
Solutions

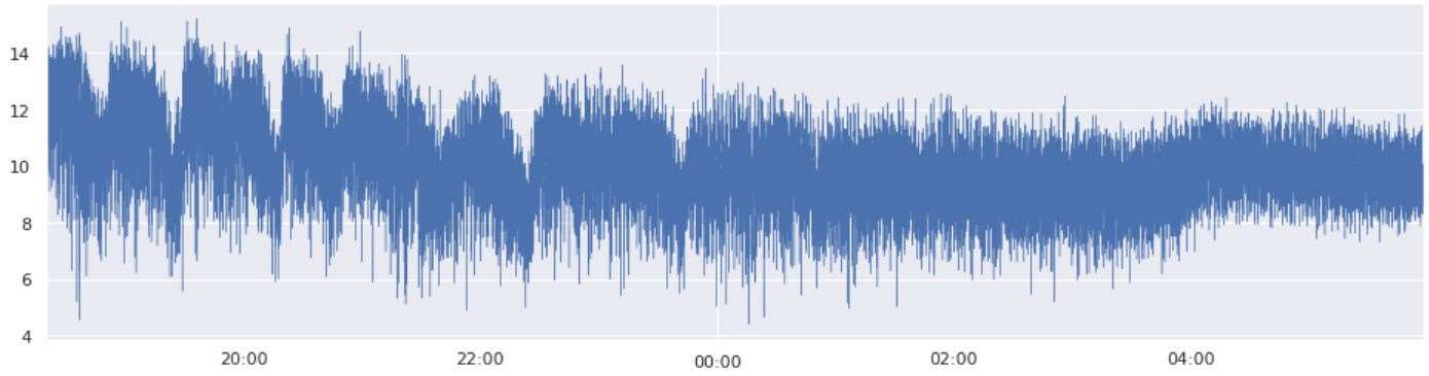
Example: Rotary Furnace

What a Rotary Furnace Looks Like…

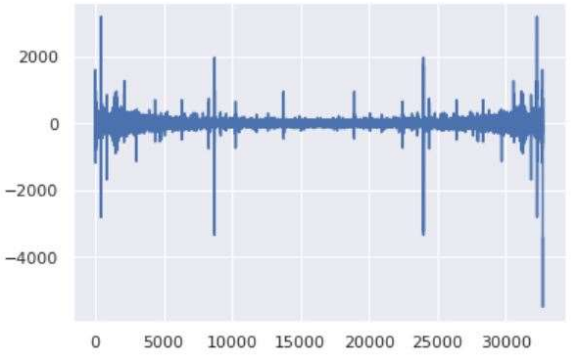
info|pro
Advanced
Analytics
Solutions

Copyright: Reimund Taubert

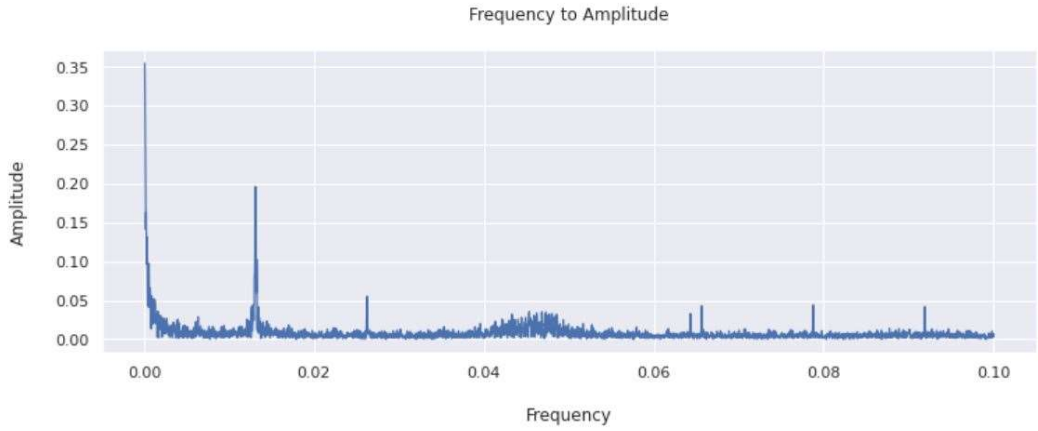
Power Analysis




```
spectrum = fft(np.array(trace_df[ofen.power]))
plt.plot(np.real(spectrum[10:]));
```



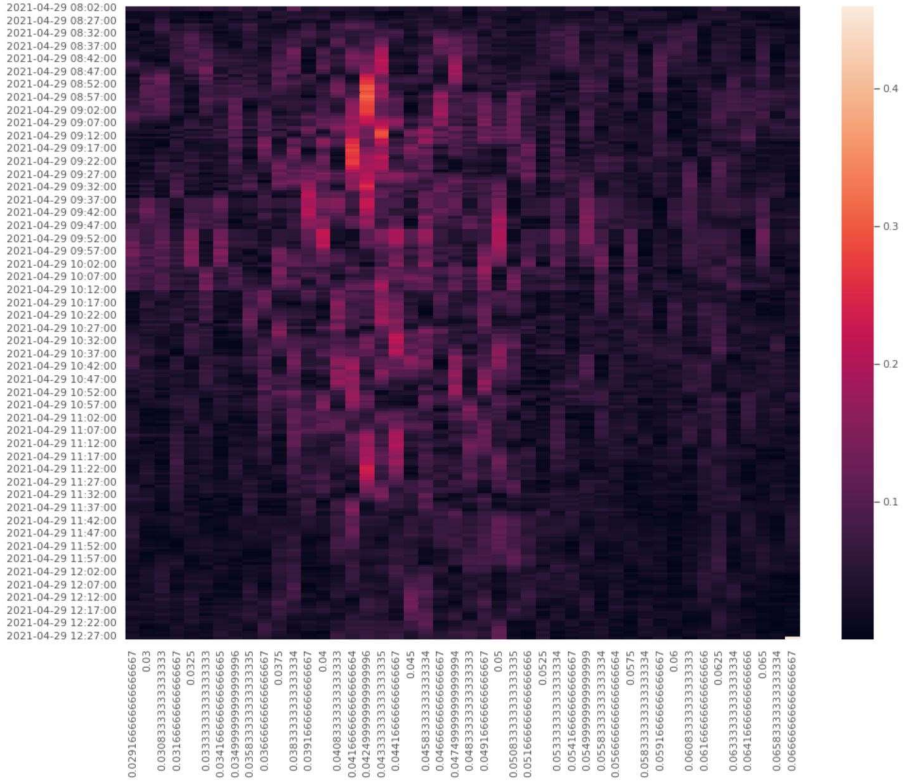

```
dft = DFT(trace_df[ofen.power])
dft.plot(kind='frequency', span=Interval(0, 0.1))
```



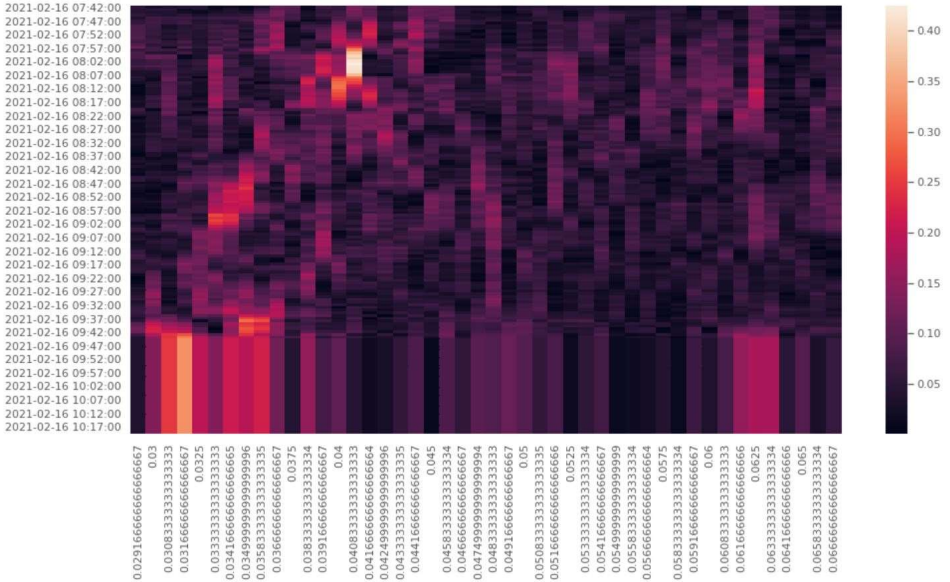
Material Spectrum

info|pro
Advanced
Analytics
Solutions

Normal



Problem



Lottozahlen vorhersagen können wir nicht, aber bei Produktionsereignissen klappt die Prognose schon ganz gut

Viele Produktionsanlagen in der chemischen Industrie werden über lange Zeiträume hinweg betrieben. Dabei stellt sich immer wieder die Frage, wie die Anlagen weiter optimiert werden können. In diesem Zusammenhang wurde vor einigen Jahren ein Team aufgestellt, das unter Verwendung der Six Sigma* Methodik entsprechende Verbesserungsprojekte durchführt.

Ziel der Projekte ist in der Regel die Steigerung der Anlagenausbeute. Aber auch andere Parameter wie zum Beispiel eine Erhöhung des Durchsatzes oder die Verbesserung der Produktqualität sind als entsprechende Schwerpunkte durchaus üblich. Gerade in einer Zeit, in der der Begriff „Nachhaltigkeit" immer stärker an Bedeutung gewinnt, können Six Sigma Projekte einen signifikanten Zukunftsbeitrag leisten. Um den Unternehmenserfolg langfristig zu sichern, rücken somit auch energetische Prozessoptimierungen oder die Verminderung von Reststoffen in den Vordergrund.

Bei einem vor zwei Jahren begonnenen Projekt konnte nach zahlreichen Versuchen und der Auswertung vieler Daten die beabsichtigte Steigerung der Ausbeute von ca. 1,0% erreicht werden. Im Rahmen dieser Datenauswertung kamen auch mathematische Methoden zum Einsatz, die unter anderem zentraler Bestandteil in der Nachrichtentechnik (Radio, Mobilfunk, etc.) sind. Hierbei wurde ein bis dato nicht bekanntes Phänomen an den Reaktoren entdeckt.

Da eine gezielte Analyse dieses Phänomens im Rahmen des Six Sigma Projekts nicht mehr möglich war, die bisher gewonnenen Erkenntnisse jedoch einen sehr vielversprechenden Eindruck machten, wurde im Rahmen des AIzFIT-Programms ein zweites Projekt mit dem Namenszusatz „Analytics" gestartet. Unter Verwendung des entdeckten Phänomens sollte ein System realisiert werden, das einen Blick in die nahe Zukunft des Produktionsprozesses ermöglicht, um die Prozesssteuerung zu verbessern und letztendlich Kosten zu sparen.

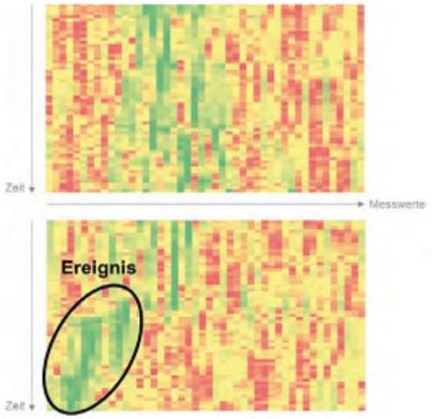
Konkret bedeutet dies, dass ein System implementiert wurde, das auf Basis von aktuellen Produktionsdaten und bekannten Datenmustern frühzeitig die Entwicklung größerer Prozessstörungen bemerkt, lange bevor ein Erkennen mit menschlichem Auge möglich ist. So kann dem nahenden Ereignis bereits vor der Entstehung schnell und effizient entgegengewirkt werden. Zudem lässt sich aus den Datenmustern eine weitere Funktion realisieren, die eine automatisierte Empfehlung an den Anlagenfahrer ausspricht, wann er gezielt in den Prozess eingreifen muss. Diese beiden Funktionen sollen wiederum die Ausgangsbasis bilden, um das „biologische Alter" einer Prozessreaktion zu ermitteln und den optimalen Zeitpunkt für eine Reinigungspause zu bestimmen.

„Anfangs", so Michael Kluge, verantwortlicher Anlagenfahrer, „herrschte bei mir schon Skepsis, ob unser Arbeitsplatz durch eine Automatisierung von Steuerungsfunktionen in Gefahr sein könnte. Auch hatte ich Zweifel, ob eine Realisierung aufgrund der verschiedenen Einflüsse auf das Verfahren überhaupt möglich sei." Sein Kollege Franz Mannseicher, stellvertretender Schichtführer, war zunächst wegen der vielen verschiedenen Daten, die ausgewertet werden müssen, ebenfalls nicht von einem Projekterfolg überzeugt. Dennoch unterstützten beide das Projekt von Beginn an nach vollen Kräften und zeigten sich den Neuerungen gegenüber stets aufgeschlossen.

Nach drei Monaten wurde ein erster Prototyp geschaffen und den Anlagenfahrern in einem Dashboard zur Verfügung gestellt. Seitdem findet eine kontinuierliche Optimierung und Erweiterung der Funktionen statt. F. Mannseicher dazu: „Man sieht, dass das System immer weiter verbessert wird. Obwohl es sich zurzeit nur um einen Prototyp handelt, können wir das System schon als Hinweisgeber und zur Bestätigung eigener Überlegungen nutzen." Auch sein Kollege M. Kluge pflichtet ihm bei: „Die Meldungen des Prototyps veranlassen uns bereits jetzt, noch genauer auf mögliche Ereignisse am Reaktor einzugehen."

Auch für die Zukunft erhoffen sich F. Mannseicher und M. Kluge noch einiges und blicken nach ersten positiven Erfahrungen zuversichtlich auf den weiteren Projektverlauf. Sie erwarten sich, dass das System bei der Steuerung der Reaktoren wertvolle Unterstützung liefern und zudem das Reinigungsverhalten verbessern wird. Als nächster Schritt ist die Übertragung des optimierten Prototyps in die direkte Produktionsumgebung geplant.

Matthias Bachmeier, Leitung Six Sigma



Grafik oben: Signalfolge eines normalen Reaktionsablaufs
Grafik unten: Signalfolge bei Auftreten einer größeren Prozessstörung (Ereignis)

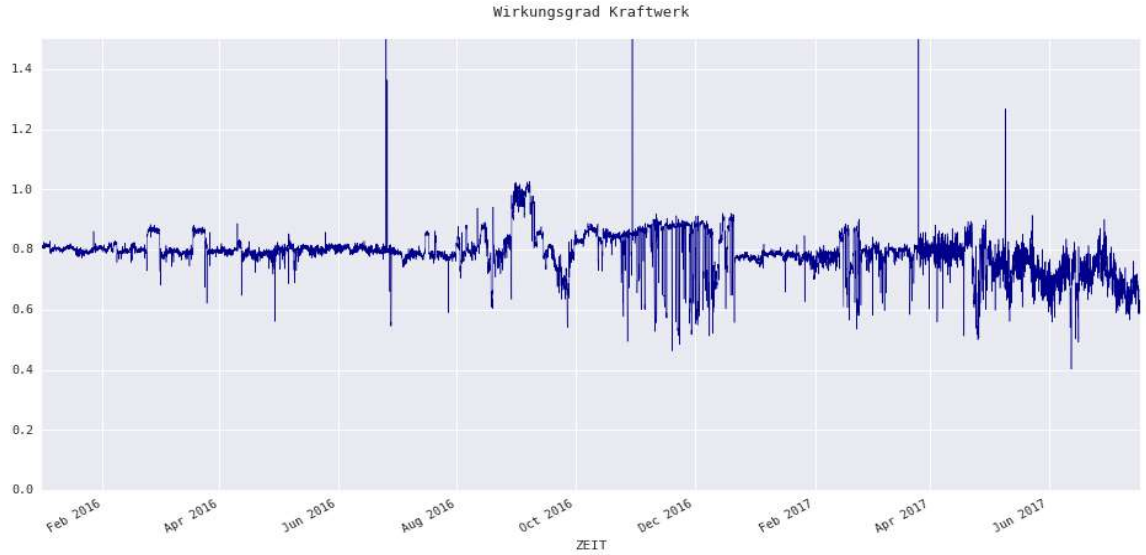
"Six Sigma Projekte sind u.a. dadurch gekennzeichnet, dass gezielt Versuche in einer Anlage durchgeführt und anhand der beobachteten Anlagenreaktion optimale Betriebsparameter oder sogar bisher nicht bekannte Wirkungszusammenhänge ermittelt werden.

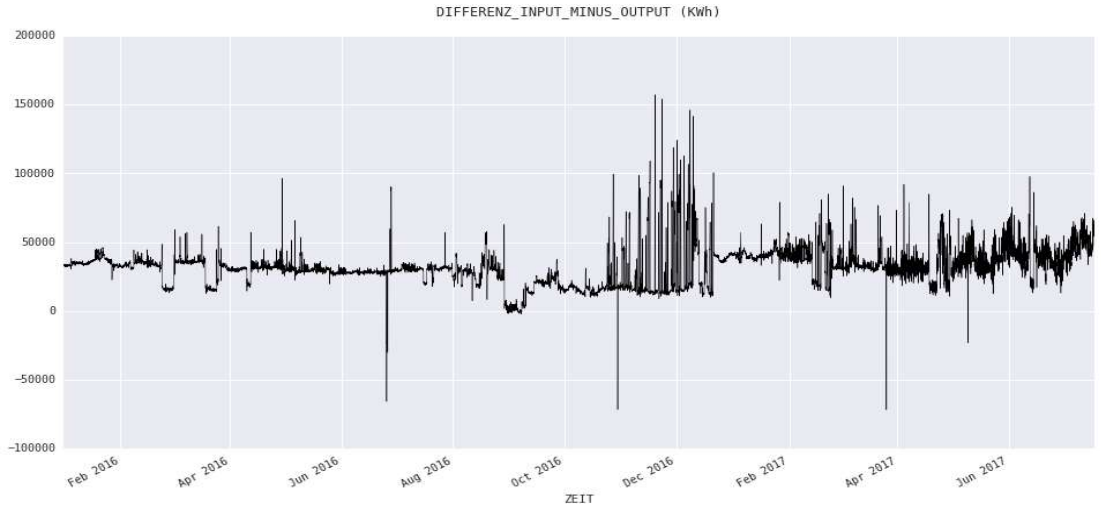
info|pro
Advanced
Analytics
Solutions

Example: Power Plant

Power Plant: Efficiency and Loss

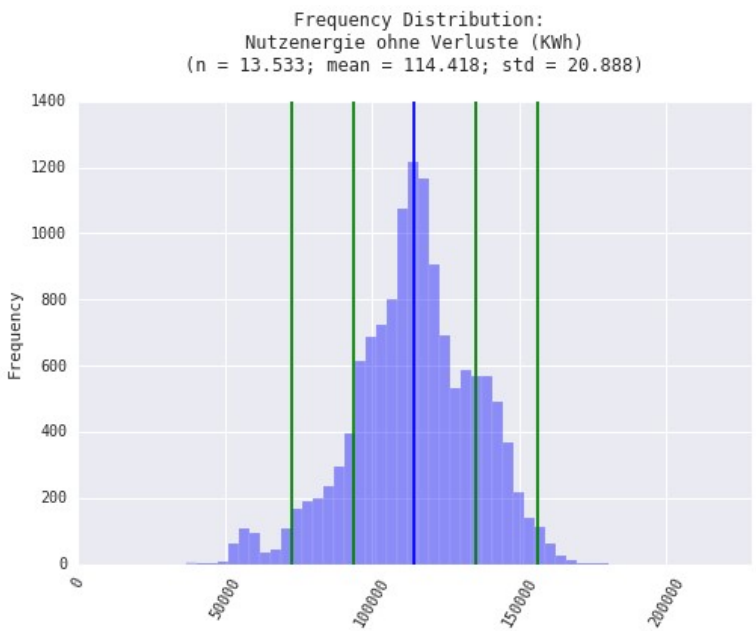
info|pro
Advanced
Analytics
Solutions

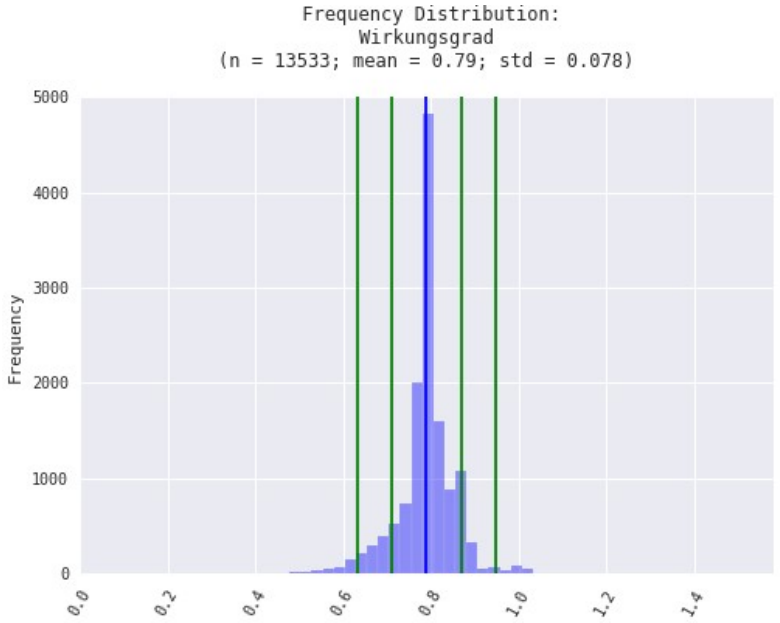




Data Source:
Energiedatenanalyse (Excel)
=> Energy Balance
01 Jan 2016 – 17 Jul 2017
Aggregation: hourly

Distribution of Output and Efficiency





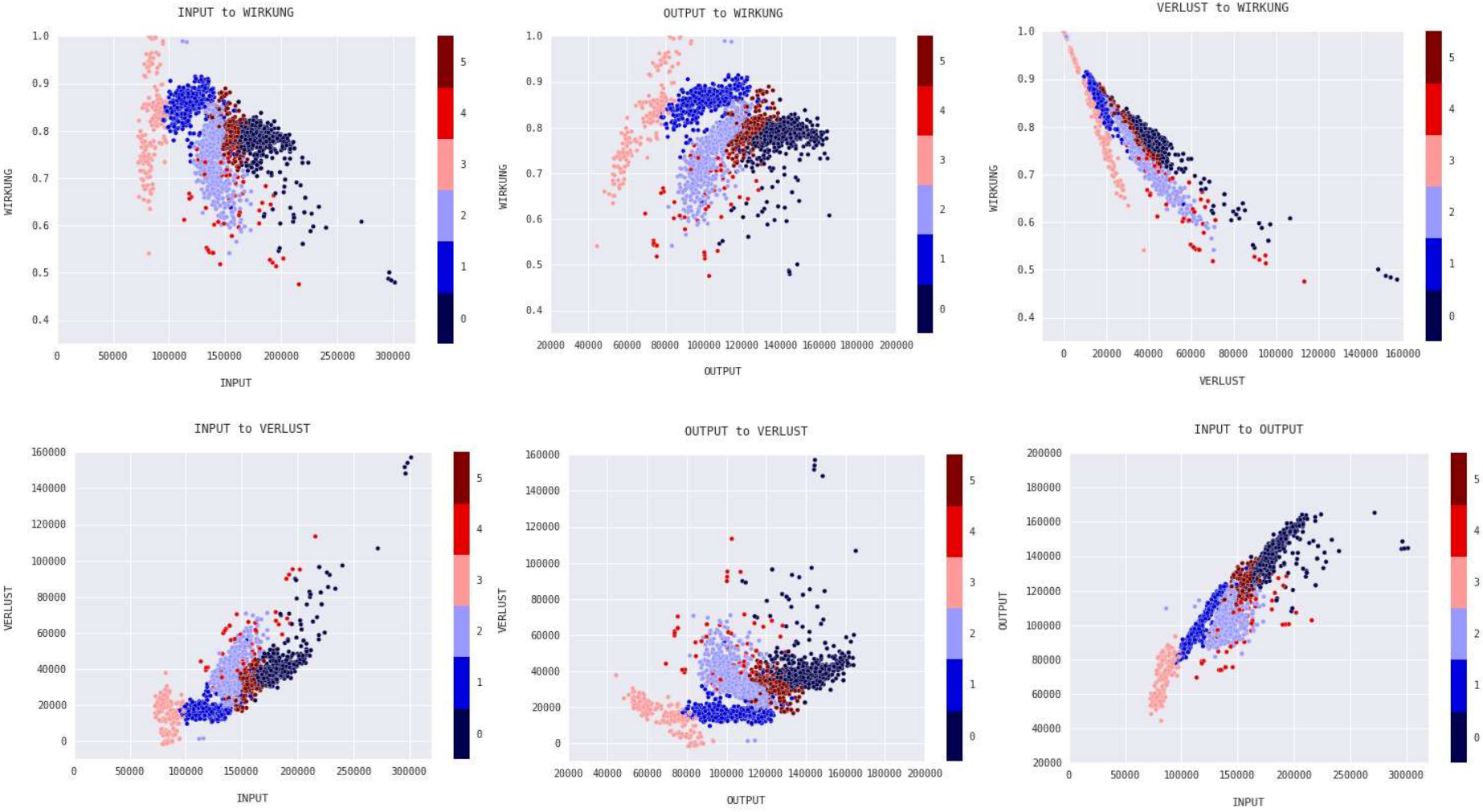
Hypothesis 1: There are two "stable" operational states:

1. 115 MW and 79% Efficiency
2. 135 MW and 87% Efficiency

Hypothesis 2: There are significant, high-end false readings

Power Plant:
Clustering in Six Operational States

info|pro
Advanced
Analytics
Solutions



Power Plant: Insights from Initial Analysis

Insights:
There is a strong correlation between Input and Loss: the Loss grows faster than the input energy.
The highest efficiency is between 100 and 150 MWh input energy.
But: only ca. 40% of the times in this energy range are in the "Sweet Spot". The rest is "Degraded", with markedly different operational state and lower efficiency.
Ca. 25% of the time the power plant is in "Overdrive", with relatively high loss.
Many of the outliers are certainly due to incorrect readings; they are not unusual operational states (e.g. efficiency near to or above 1).
When below 100 MWh input energy, the efficiency is random.

Recommendations

info|pro
Advanced Analytics Solutions

- Optimize the quality of the data sources
- Analysis of loss, especially when input increases
- Concentrate operation efforts on the "Degraded" mode, in order to increase efficiency: determine ways to move to "Sweet Spot" or "High" mode
- Analyze the reasons for the random efficiency when the input energy is under 100 MWh
- Clarify instable performance since Feb. 2017
- Further breakdown of inputs and outputs in "Energy Balance Components", including analysis of those components
- Modeling and simulation of operational states, with consideration of overall profitability, technical requirements and regulatory restrictions

info|pro
Advanced
Analytics
Solutions

Exercise: Cement Mills

Exercise / Data

info|pro
Advanced Analytics Solutions

In a CSV file:
(mill.csv)

```
mill_df
```

	Sorte	ZM	Datum	Betriebszeit	Gesamtenergie	Produktion	kWh/t	t/h
0	01	ZMA	2021-02-09 00:00:00	3.82	10639.81	372.20	28.586271	97.434555
1	01	ZMA	2021-02-17 00:00:00	17.29	24280.89	858.17	28.293800	49.633892
2	01	ZMA	2021-02-18 00:00:00	0.00	0.00	0.00	NaN	NaN
3	01	ZMA	2021-07-16 00:00:00	1.61	2212.74	68.62	32.246284	42.621118
4	01	ZMA	2021-07-29 00:00:00	1.47	1960.19	62.97	31.128950	42.836735
...	...	...	...	...	...	...	...	...
4546	04	ZMC	2022-11-28 00:00:00	20.08	11956.19	287.21	41.628739	14.303287
4547	04	ZMC	2022-11-29 00:00:00	24.00	14245.37	338.76	42.051511	14.115000
4548	04	ZMC	2022-11-30 00:00:00	22.30	13235.85	324.15	40.832485	14.535874
4549	04	ZMC	2022-12-01 00:00:00	0.43	257.95	5.56	46.393885	12.930233
4550	04	ZMC	2022-12-02 00:00:00	0.58	346.99	8.37	41.456392	14.431034

4551 rows × 8 columns

```
mill_df.Sorte.unique()
```

```
array(['01', '02', '03', '04'], dtype=object)
```

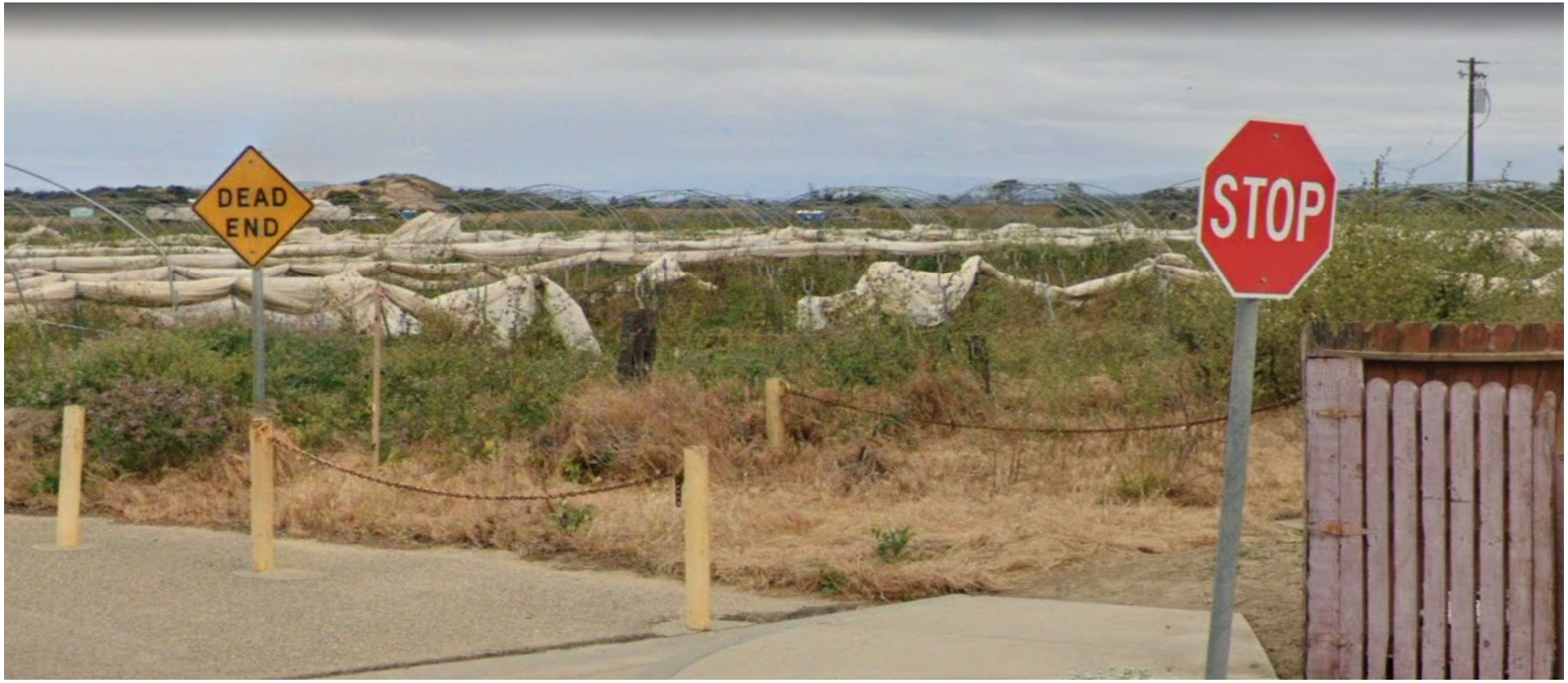
```
mill_df.ZM.unique()
```

```
array(['ZMA', 'ZMB', 'ZMC'], dtype=object)
```

Exercise Task

info|pro
Advanced Analytics Solutions

- Search for meaningful, significant clusters in the cement mill data provided. These may well be best found for individual combinations of mill and sort.
- Compare the results of multiple clustering algorithms, for example:
  - k-Means
  - Hierarchical Clustering
  - DBSCAN
  - Agglomerative Clustering
- Describe and quantify the challenges you have with the data.



Information Professionals GmbH
Sägewerkstraße 3
D-83395 Freilassing

Telefon: +49 (0) 86 54 / 77 63 420
E-Mail: kontakt@infopro-gmbh.de
Internet: www.infopro-gmbh.de

info|pro
Advanced
Analytics
Solutions