
Feinstaub Hackathon

Veranstaltet von der Stuttgarter Zeitung
am 20.01.2018

Ergebnisse der Datenanalyse von
Dr. David James, Jonathan v.d.Kamp,
Olga Moreva, Dr. Simon Müller,
Dirk Rönsch, Joachim Rosskopf



Prämisse des Hackathons 🤖

- Wir haben die Daten aus dem [Luftdaten-Archiv](#) des OK Lab Stuttgart mit [DWD Wetterdaten](#) verschnitten.
- Download der **~500x10e6 Datenpunkte** bzw. **~ 1010000 Dateien** bzw. **~ 47 GB** dauerte > 4 Stunden.
- Anschliessendes parsen & komprimieren der Daten ins [Parquet-Format](#) > 1 Stunden. Führt jedoch zu einer Reduktion der Datenmenge auf ~ 8 GB. ([Download Link](#))
- Selektion von Sensoren in Stuttgart, Zusammenfassung von Sensoren, ver-joinen mit DWD Daten, Bereinigung, dauert weitere Zeit (> 1,5 Stunden) und führt zu einem **Ausgangs-Datensatz** für die Analyse von ca. **500 MB**.
- Die Daten für Stuttgart [roh](#), [aggregiert](#) und als [Zeitreihen](#) sind unter den jeweiligen Links zu erreichen. Ein [Github Repository mit Notebooks](#) gibt es auch.

Das OK Lab besitzt einen sehr interessanten, potentiell wertvollen und erstaunlich umfangreichen Datensatz. Wir haben den Datensatz deshalb vorbereitet mitgebracht.



Fragestellung: Datenqualität 🙄

Nutzen:

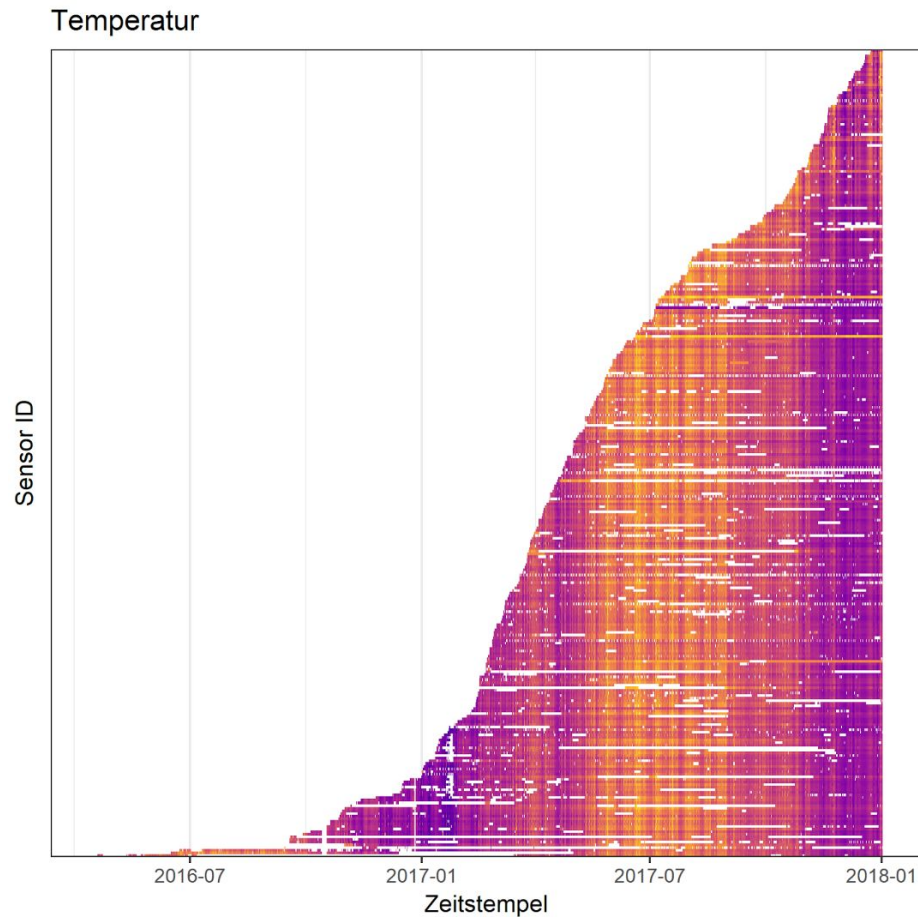
Für weitere Betrachtungen (wissenschaftlich / kommerziell) der durch Citizen Scientist gesammelten Daten, ist die Datenqualität eine notwendige Voraussetzung. Wir stellen hier ein Framework vor (siehe [3]), welches den „Wert“ der gesammelten Daten auf ein professionelles Niveau heben kann.

Mögliche ableitbare Aktionen:

- Sensor-Paten könnten aktiv per Push-Benachrichtigung bei ungenügender Datenqualität automatisch benachrichtigt werden.
- Durch automatische Vorfilterung entsteht ein qualitativ hochwertiger Datensatz, der Weiterverarbeitung stark vereinfacht.

Datenqualität durch Referenzierung

Anhand der Temperatur und den Referenzwerten der DWD Messstationen.



Datenqualität durch Referenzierung

Anhand der Temperatur und den Referenzwerten der DWD Messstationen.

Bestimmung eines Scores:

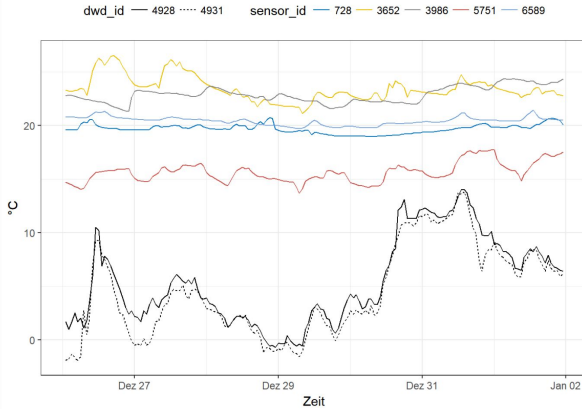
1. Grenze den zu betrachtenden Zeitraum ein (Bsp. 1 Woche)
2. Berechne den **mittleren quadratischen Abstand** zwischen Sensor- und Referenzwerten:

$$score_j = \sqrt{\frac{1}{n} \sum (y_i - y_{DWD})^2}$$

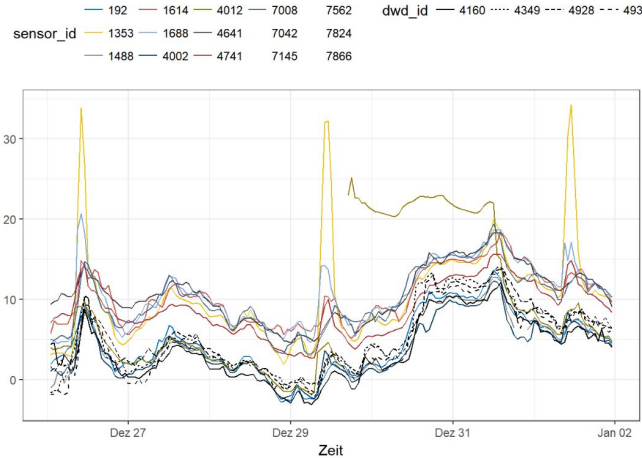
3. Standardisiere den Score und berechne basierend auf der Normalverteilungsannahme eine Warnstufe
 - **Grün**
 - **Gelb**: Score > 1SD
 - **Orange**: Score > 2SD
 - **Rot**: Score > 3SD

Datenqualität durch Referenzierung

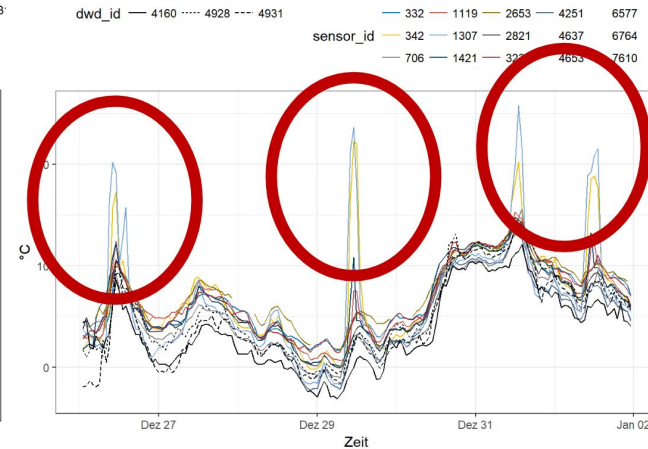
Rote Gruppe




Gelbe & Orange Gruppe



Grüne Gruppe



- Gelb/orangene Gruppe ist eine Verschiebung zu den DWD-Sensoren sichtbar
- Bei den Grünen sind Sensoren mit starken Spitzen enthalten



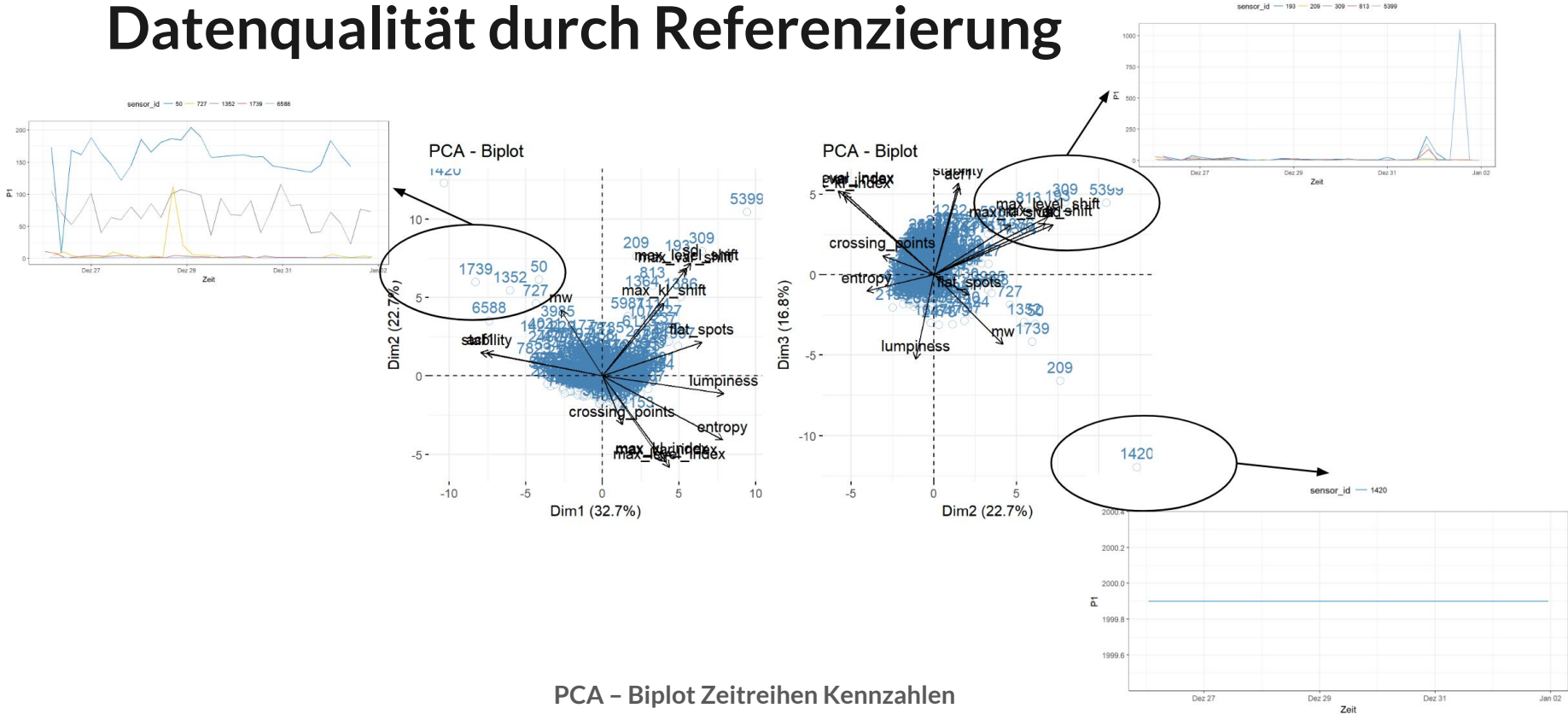
Datenqualität durch Referenzierung

Anhand der Temperatur und den Referenzwerten der DWD Messstationen.

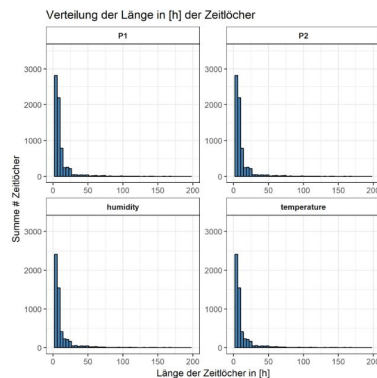
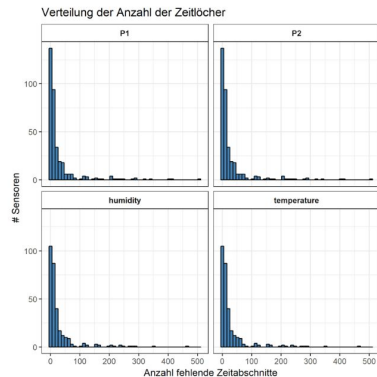
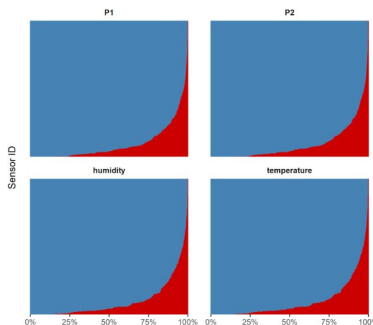
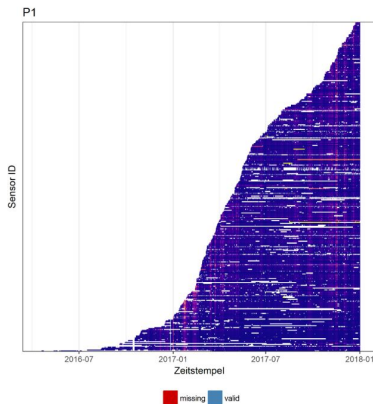
Bestimmung eines Scores:

4. Berechne unterschiedliche **Zeitreihen-Kennzahlen**, wie
 - Stabilität,
 - Lumpiness,
 - Max Var Shift, Max Mean Shift, Max KL Shift,
 - Autokorrelation,
 - Entropy,
 - Crossing Points,
 - Flat Spots, ...
5. Berechne eine **PCA** auf dieser Kennzahlen Matrix und verwende den normierten **Abstand** der Sensoren zu den **Hauptkomponentenachsen** als Score.

Datenqualität durch Referenzierung



Fehlende Werte



Analyse fehlender Zeitabschnitte:

- Die überwiegende Anzahl der Sensoren hat wenige Lücken.
- Eine kleine Anzahl hat große Lücken.
- Jedoch sind diese “Aussetzer” nur von kurzer dauer.

Mögliche ableitbare Aktionen:

- Entwicklung eines Scores.
- Aggregation ähnlicher Sensoren, mit Ziel fehlende Werte aufzufüllen.

Fragestellung: Datenqualität

Weitere ableitbare Aktionen:

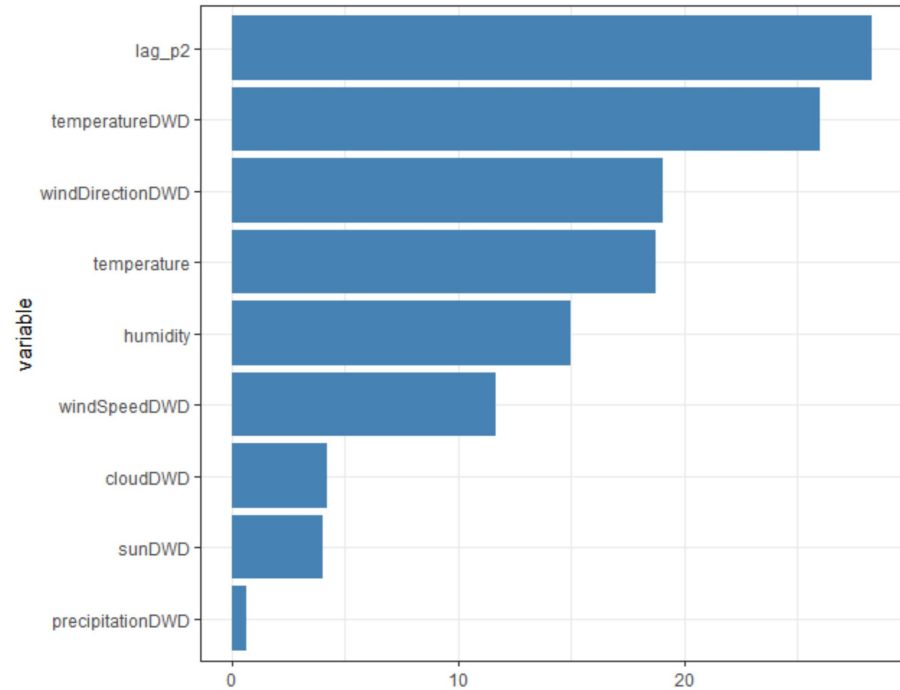
- Die Sensorqualität ist auch zeitlich gekoppelt. Berechne Warnstufen basierend auf einem gleitenden Fenster, d.h. ein zeitliches Qualitätsmaß je Sensor.
- Referenzstationen fuer Partikel-Zahl/Konzentration wuerde die Qualitaetsbetrachtung dieses Sensortyps vereinfachen.

Weitere technische Ideen:

- Passe für jeden Sensor eine Regressionskurve an, berechne die erste / zweite Ableitung und berechne dann den L2-Abstand zu den Referenzdaten.
- Verwende zusätzlich Zeitreihen-Kennzahlen für die Ähnlichkeitsbestimmung der Zeitreihen (siehe Hyndman, et. al.) und berechne daraus den PCA-Score. Dieser Ansatz wäre mehr Data Mining und weniger Statistik lastig.
- Verwende zusätzlich die k-nächsten Nachbarn (basierend auf den Geo-Koordinaten) und berechne bzgl. dieser den mittleren Abstand (Schränke den maximalen Abstand ein). Gewichte nun den Abstand zu den DWD-Daten und zu den k-NN-Daten. Die Gewichte könnte man durch den Abstand der Sensoren ermitteln, sowie dem DWD-Sensor ein höheres Gewicht aufgrund der hohen Verlässlichkeit geben. Hierdurch könnte berücksichtigt werden, daß bestimmte Gebiete wärmer sind.

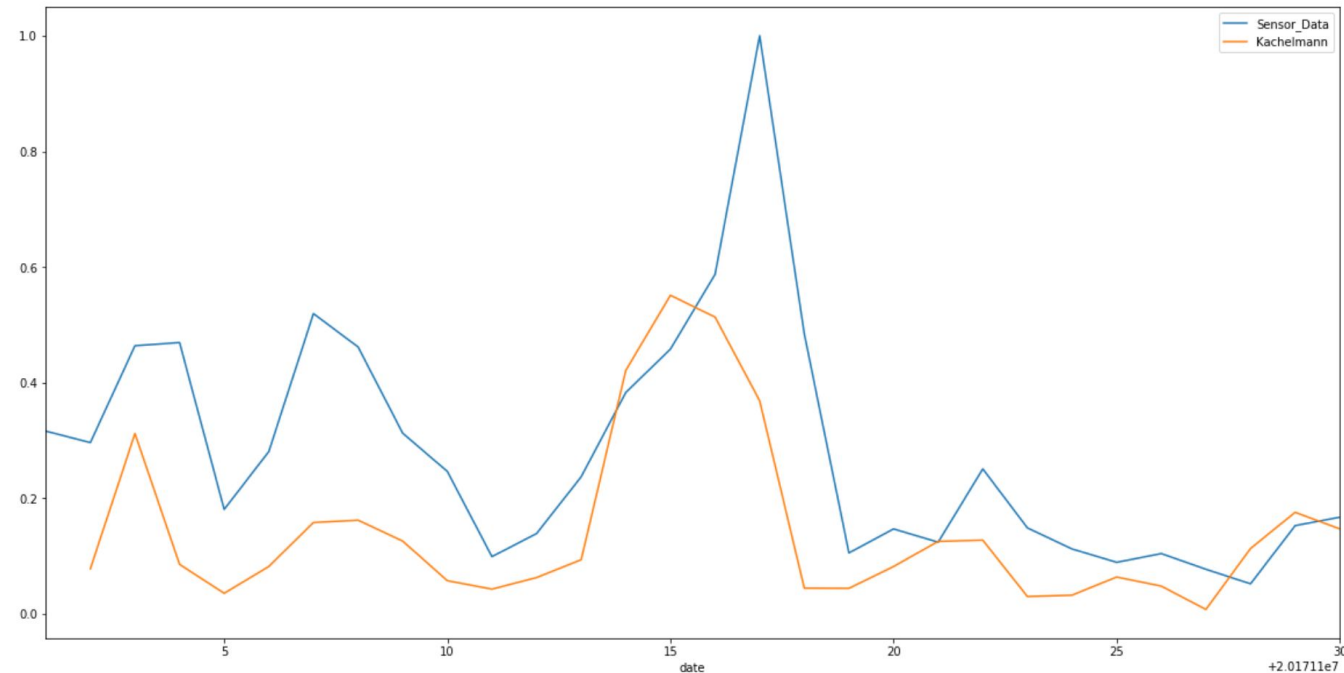
Treiberanalyse

Welche Variablen haben Einfluss
auf den Partikel-Konzentrations-
Messwert



- Obige Ausgabe ist Ergebnis eines Trainingslaufes eines Decision Tree Algo.
- Der Skalierung Abszisse ist keine physikalische Bedeutung zuzuschreiben.

Vergleich der OKLab API Werte mit Kachelmann-Index



Mögliche Aussagen:

- Kachelmann detektiert Zeitpunkte der Trendumkehr korrekt.
- Die Trends in den Vorhersagen sind korrekt.
- Abweichung in Größenordnung nicht geklärt.

Referenzen



Literatur:

- [1] Robert Hyndman. Automatic algorithms for time series forecasting. [PDF]
- [2] Wang, R.Y. and D.M. Strong, Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst., 1996. 12(4): p. 5-33. [PDF]
- [3] K. Crowston and N. R. Prestopnik. Motivation and Data Quality in a Citizen Science Game: A Design Science Evaluation, 2013 46th Hawaii International Conference on System Sciences, Wailea, Maui, HI, 2013, pp. 450-459. [PDF]

Links zu Software und Notebooks:

- Github Repository mit Notebooks: https://github.com/anofox/StZ_Feinstaub_Hackathon
- Link zu Daten:
 - Ausgangsdaten: Archiv von <http://luftdaten.info/> und CDC Daten des DWD von <ftp://ftp-cdc.dwd.de/pub/CDC/>
 - Daten fuer die Analyse: https://storage.googleapis.com/datenlager/stgt_sensors_with_date_geo_dwd.parquet.tar.gz
 - Einige derivate dieser Daten:
 - https://storage.googleapis.com/datenlager/stgt_sensors_with_date_geo_dwd_aggregated.parquet.tar.gz
 - https://storage.googleapis.com/datenlager/stgt_sensors_with_date_geo_dwd.parquet.tar.gz

Hat Spass gemacht!

Feinstaub Hackathon der Stuttgarter Zeitung
am 20.01.2018

Ergebnisse der Datenanalyse von
Dr. David James, Jonathan v.d.Kamp,
Olga Moreva, Dr. Simon Müller,
Dirk Rönsch, Joachim Rosskopf

Ich bin der **AnoFox**, eine
Softwarekomponente der
Analysten zur **Anomalie-
Erkennung** und
Zeitreihen-Vorhersage.
Mehr infos unter:
<https://anofox.com>

