

Deutsche Bahn in die Karten geschaut

IKS-Thementag

Autor: Philipp Steinweg

16.08.2022



Agenda

- Data Science & die OSEMN-Pipeline
- Die Daten & unser Ziel
- OSEMN-Pipeline am Beispiel
 - ◆ Explorative Datenanalyse
 - ◆ ML-Modelle: Training & Ergebnisse

Agenda

- **Data Science & die OSEMN-Pipeline**
- Die Daten & unser Ziel
- OSEMN-Pipeline am Beispiel
 - ◆ Explorative Datenanalyse
 - ◆ ML-Modelle: Training & Ergebnisse

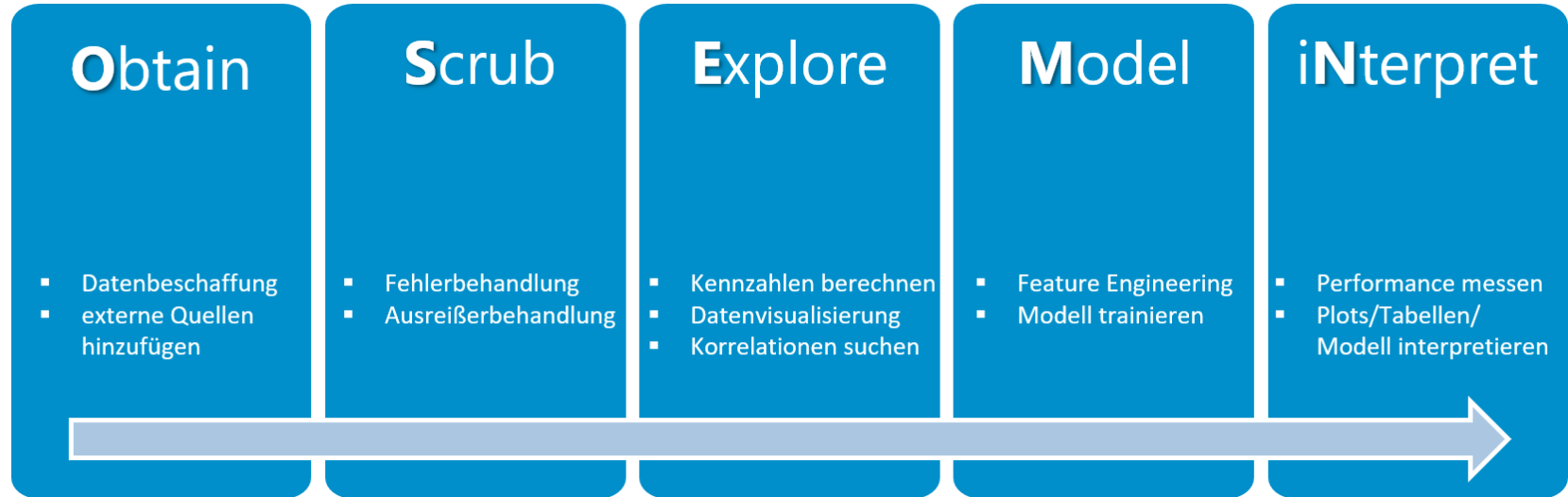
Was ist Data Science?



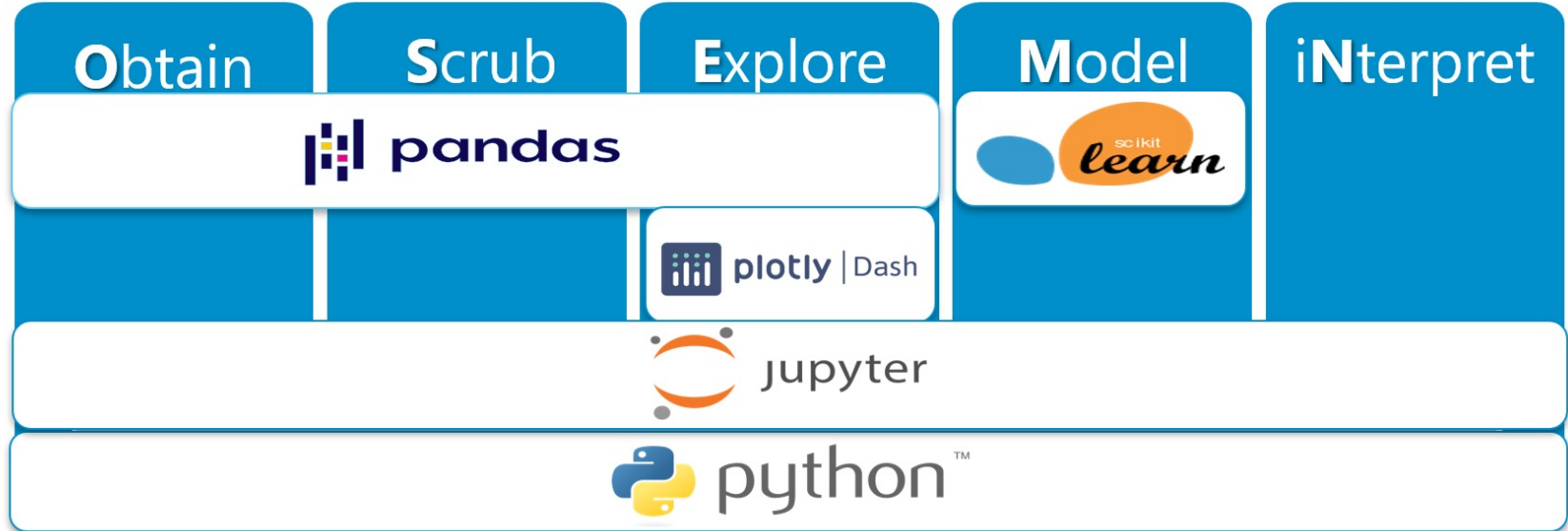
Data Science

- ◆ ist interdisziplinär,
- ◆ extrahiert Wissen,
- ◆ zeigt Muster auf
- ◆ und stellt Zusammenhänge dar
- ◆ aus strukturierten und unstrukturierten Daten.

Von Daten zu Wissen: OSEMNI!



Verwendete Technologien



Agenda

- Data Science & die OSEMN-Pipeline
- **Die Daten & unser Ziel**
- OSEMN-Pipeline am Beispiel
 - ◆ Explorative Datenanalyse
 - ◆ ML-Modelle: Training & Ergebnisse

Die Daten

➤ API liefert Abfahrtstafel zu jeweils einer Bahnstation zum Abfragezeitpunkt

22:23 Abfahrt Düsseldorf Hbf Nutzungsbedingungen Impressum 			
Zeit	Nach	Über	Gleis
21:03 RE 5	Duisburg Hbf	Düsseldorf  ca. 80 Minuten später ***** Heute Gleis 4 ***** Halte nicht in Oberhausen Hbf, Oberhausen Sterkrade, Dinslaken, Voerde(Niederrhein), Wesel.	4
22:03 RE 5	Emmerich	Sterkrade, Oberhausen-Holtten, Dinslaken, Voerde(Niederrhein), Friedrichsfeld(Niederrhein), Wesel, Mehrhoog ***** Düsseldorf ca. 35 Minuten später ***** Heute Gleis 10	10
22:21 RE 4	Aachen Hbf	Wf, Erkelenz, Hückelhoven-Baal, Lindern, Geilenkirchen, Übach-Palenberg, Herzogenrath, Aachen West, Aachen Schanz *****	4
22:22 RE 1	Hamm(Westf)Hbf	Wattenscheid, Bochum Hbf, Dortmund Hbf, Dortmund-Scharnhorst, Dortmund-Kurl, Kamen-Methler, Kamen, Bönen-Nordbögg ca. 5 Minuten später ***** Heute Gleis 5	5
22:24 IC 2307	Koblenz Hbf	Köln Hbf, Bonn Hbf, Remagen, Andernach ca. 20 Minuten später	16
22:25 S 1	Solingen Hbf	Düsseldorf Volksgarten, Düsseldorf-Oberbilk, Düsseldorf-Eller Mitte, Düsseldorf-Eller, Hilden, Hilden Süd, Solingen Vogelpark ca. 5 Minuten später	11
22:26 S 28	Kaarster See	Er Str., Düsseldorf-Hamm, Neuss Rheinparkcenter, Neuss Am Kaiser, Neuss Hbf, Kaarst IKEA, Kaarst Mitte/Holzbüttgen, Kaar	12
22:26 VIA RE19	Arnhem Centraal	Sterkrade, Oberhausen-Holtten, Dinslaken, Voerde(Niederrhein), Friedrichsfeld(Niederrhein), Wesel, Wesel Feldmark, Mehrhoog	7
22:30 S 6	Kettwig	Düsseldorf-Rath Mitte, Düsseldorf-Rath, Ratingen Ost, Höselt, Kettwig Stausee ***** Düsseldorf Wehrhahn, Düsseldorf-Zoo, C	14
22:32 S 6	Langenfeld(Rhld)	Hild, Düsseldorf-Reisholz, Düsseldorf-Benrath, Düsseldorf-Garath, Düsseldorf-Hellerhof, Langenfeld(Rhld)-Berghausen ***** I ca. 10 Minuten später	11
22:33 S 8	Mönchengladbach Hbf	Er Str., Düsseldorf-Hamm, Neuss Rheinparkcenter, Neuss Am Kaiser, Neuss Hbf, Büttgen, Kleinenbroich, Korschebroich, Mö	12

Deutsche Bahn in die Karten geschaut

Die Daten: Beispielantwort API

✱ Eine Zeile der Ankunftstafel
Dortmund Hbf. 04.06 um
10:29:46

```
{'delayArrival': None,
'delayDeparture': 0,
'destination': 'Dorsten',
'isCancelled': 0,
'messages': {'delay': [], 'qos': []},
'missingRealtime': False,
'platform': '23',
'route': [{'name': 'Dortmund Hbf'},
          [...],
          {'name': 'Dorsten'}],
'scheduledArrival': None,
'scheduledDeparture': '12:32',
'scheduledPlatform': '23',
'train': 'RB 43',
'trainClasses': ['N'],
'trainNumber': '14767',
'via': ['Dortmund-Huckarde Nord', 'Dortmund-Rahm',
        'Wanne-Eickel']}
```

Die Daten: Hinweise

- ✱ Wir fragen Start / Zielstation der Linien RE1-RE9 ab (insg. 14 Stationen)
- ✱ Abfrage einer(!) Station alle 12 Sekunden
 - ◆ 12 Sek. * 14 Stationen = 2 Min. 48. Sek bis zur erneuten Abfrage
 - ◆ Grund: Limitierung durch API
- ✱ Erfassungszeitraum: 30.04.2022 - 15.08.2022

Die Daten: Welche Bahnlinien/Bahnhöfe wurden erfasst?

Linie	Start ↔ Ziel
RE1	Hamm Hbf ↔ Aachen Hbf
RE2	Osnabrück ↔ Düsseldorf Hbf
RE3	Hamm Hbf ↔ Düsseldorf Hbf
RE4	Dortmund Hbf ↔ Aachen Hbf
RE5	Emmerich Hbf ↔ Koblenz Hbf
RE6	Minden Hbf ↔ Köln/Bonn Flughafen
RE7	Rheine Hbf ↔ Krefeld Hbf
RE8	Mönchengladbach Hbf ↔ Koblenz Hbf
RE9	Aachen Hbf ↔ Siegen Hbf

Deutsche Bahn in die Karten geschaut

Data Science & OSEM-N-Pipeline | Die Daten & unser Ziel | OSEM-N-Pipeline am Beispiel | Ende

Unser Ziel

✦ Training eines ML-Modells, dass Verspätungen in Minuten vorhersagt

Folgende Einschränkungen:

- ◆ Beschränkung auf die Linien RE1-RE9
- ◆ Vorhersage nur für abgeschlossene Fahrten

Abgeschlossene Fahrten konstruieren

- ✱ Daten zeigen „Momentaufnahmen“ einer Ankunftstafel
 - Wie konstruieren wir daraus „abgeschlossene“ Fahrten?

- ✱ Hierfür müssen wir mindestens Start- und Zielbahnhof einer Linie beobachten
 - ◆ Warum?

Beispiel: RE2, Zugnummer 10203 vom 10.06

✧ Wie oft taucht der Zug auf unseren Ankunftstafeln auf?

	train	trainClasses	trainNumber	date	timestamp	delayArrival	delayDeparture	destination	delayArrival	delayDeparture	destination
35	RE 2	[N]	10203	2022-06-10	00:31:39	NaN	0.0	Osnabrück Hbf	NaN	0.0	Osnabrück Hbf
85	RE 2	[N]	10203	2022-06-10	00:34:20	NaN	0.0	Osnabrück Hbf	NaN	0.0	Osnabrück Hbf
...	...	...	...	...	...	...	...	...	...	...	...
15090	RE 2	[N]	10203	2022-06-10	05:12:39	4.0	NaN	Osnabrück Hbf	4.0	NaN	Osnabrück Hbf
15273	RE 2	[N]	10203	2022-06-10	05:15:21	5.0	NaN	Osnabrück Hbf	5.0	NaN	Osnabrück Hbf

127 rows × 27 columns

Letztes „Signal“ jeder Station

- ❖ Für die Rekonstruktion der Fahrt reicht uns die Anzeige „kurz vor Abfahrt“
→ Wir filtern das zeitlich letzte Signal jeder Station

	train	trainClasses	trainNumber	date	timestamp	delayArrival	delayDeparture	destination
1	RE 2	[N]	10203	2022-06-10	03:04:08	NaN	0.0	Osnabrück Hbf
0	RE 2	[N]	10203	2022-06-10	05:15:21	5.0	NaN	Osnabrück Hbf

Fahrt rekonstruieren

	trainNumber	date	train	scheduledDeparture	scheduledArrival	start_delay_departure	final_delay_arrival
0	10203	2022-06-10	RE 2	05:06	07:11	0.0	5.0

Agenda

- Data Science & die OSEMN-Pipeline
- Die Daten & unser Ziel
- **OSEMN-Pipeline am Beispiel**
 - ◆ Explorative Datenanalyse
 - ◆ ML-Modelle: Training & Ergebnisse

Agenda

- Data Science & die OSEMN-Pipeline
- Die Daten & unser Ziel
- OSEMN-Pipeline am Beispiel
 - ◆ Explorative Datenanalyse
 - ◆ ML-Modelle: Training & Ergebnisse

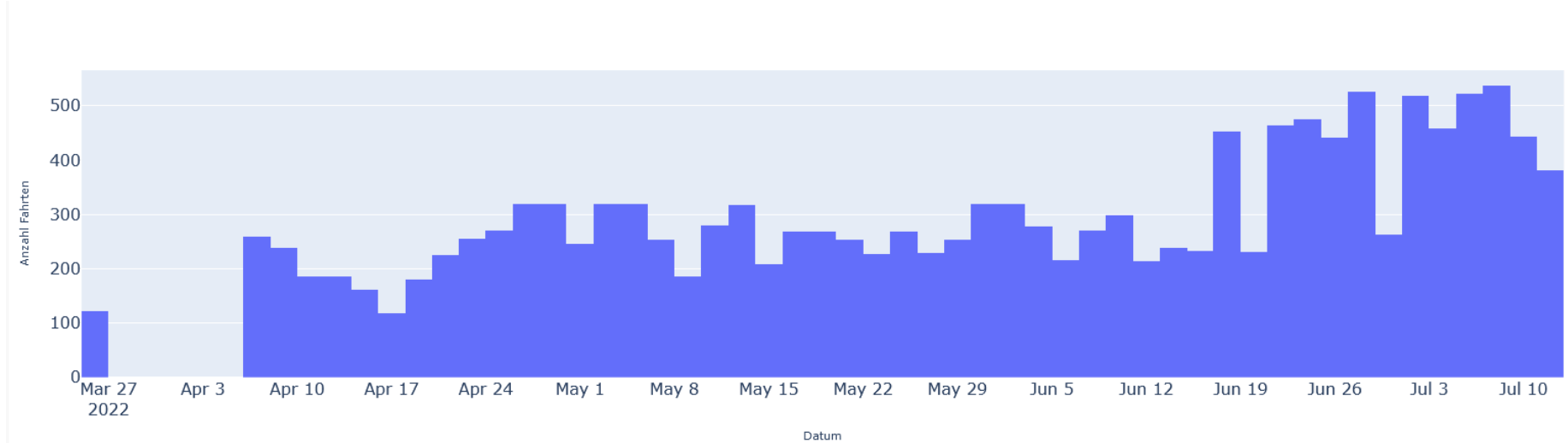
Erster Blick auf unsere Daten

	trainNumber	date	train	scheduledDeparture	scheduledArrival	start_delay_departure	final_delay_arrival	messages_delay	messages_qos	platform	route
0	10203	2022-04-07	RE 2	05:06	07:11	1.0	1.0	[]	[]	4	[{'isAdditional': None, 'isCancelled': None, '...
1	10203	2022-04-23	RE 2	05:06	07:10	0.0	0.0	[]	[]	4	[{'isAdditional': None, 'isCancelled': None, '...
...	...	...	...	...	...	...	...	...	...	...	...
14823	89912	2022-07-09	RE 3	04:03	05:12	0.0	0.0	[]	[]	26	[{'isAdditional': None, 'isCancelled': None, '...
14824	89912	2022-07-10	RE 3	04:03	05:12	0.0	0.0	[]	[]	9	[{'isAdditional': None, 'isCancelled': None, '...

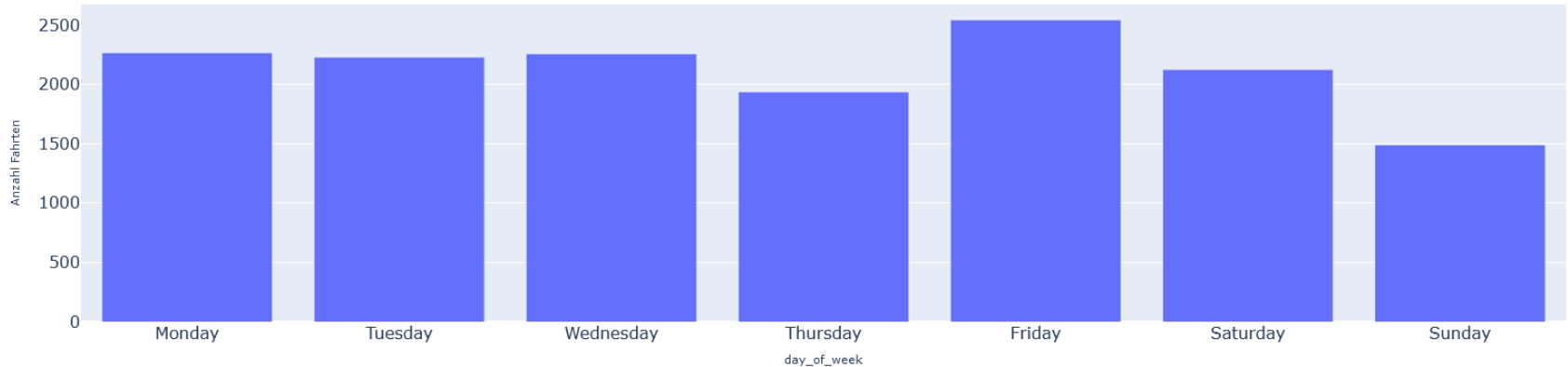
14825 rows × 14 columns

	count	mean	std	min	25%	50%	75%	max
start_delay_departure	14825.0	1.524384	6.371098	-2.0	0.0	0.0	0.0	116.0
final_delay_arrival	14825.0	6.715616	12.369276	-122.0	0.0	1.0	8.0	143.0

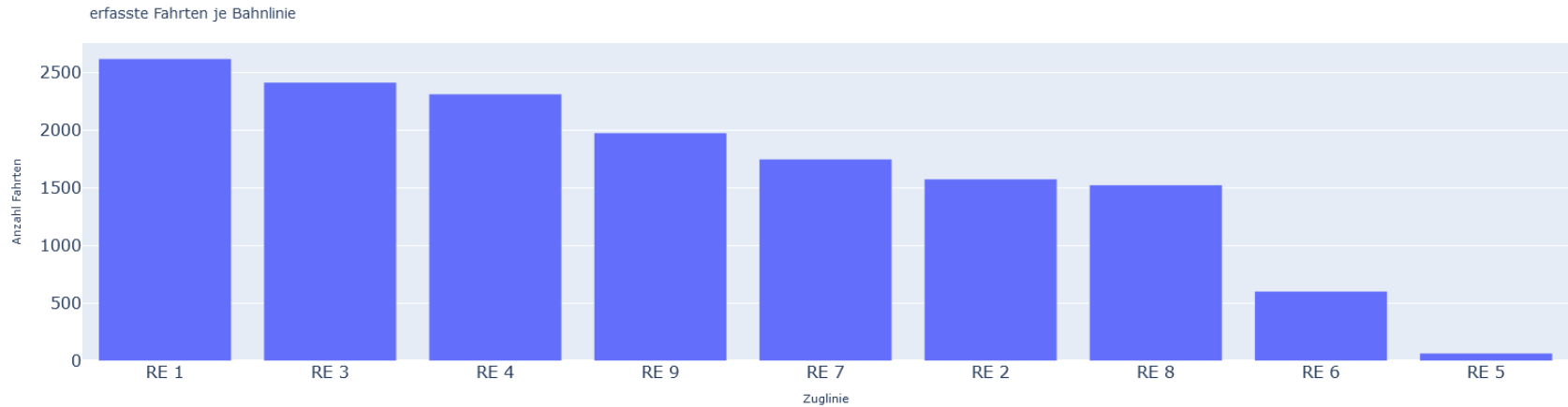
Erfasste Fahrten pro Datum



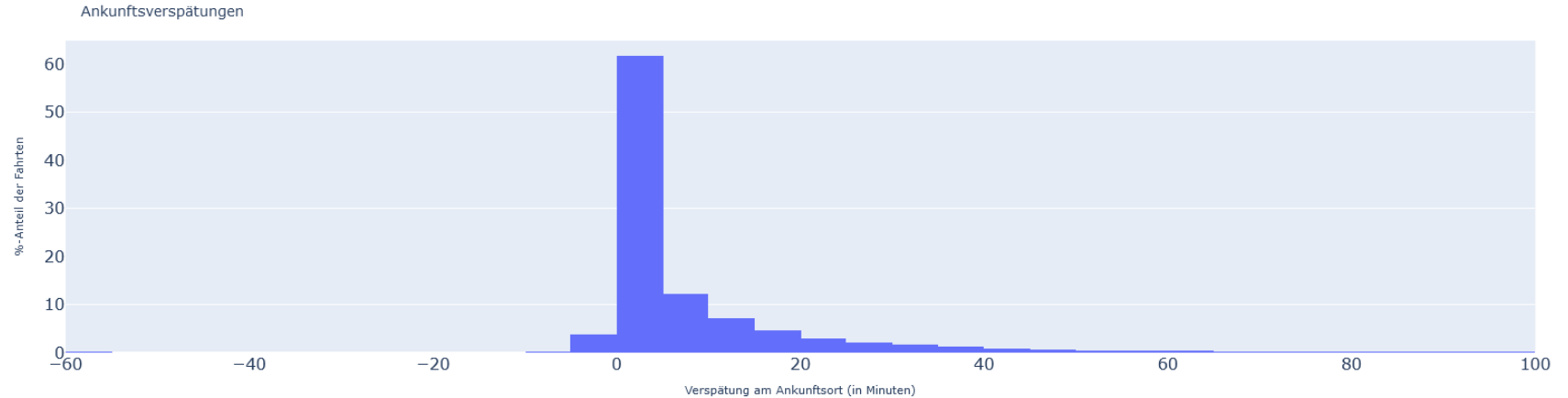
Erfasste Fahrten je Wochentag



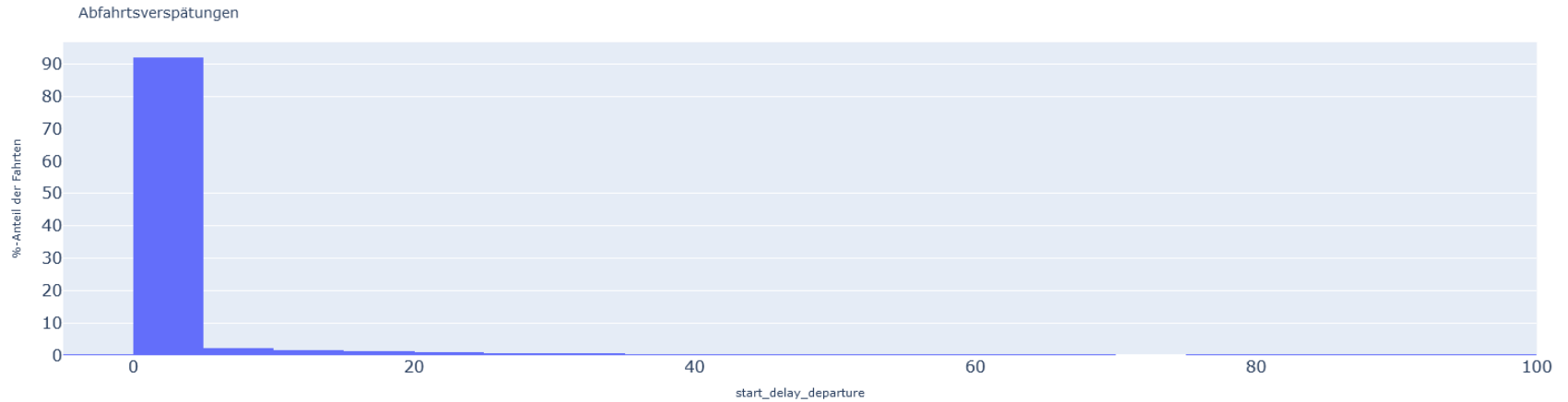
Erfasste Fahrten je Bahnlinie



Verteilung der Verspätungen



Verteilung der Verspätungen



Verspätungsnachrichten

* Innerhalb des Datensatzes Spalten: *messages.qos* & *messages.delay*

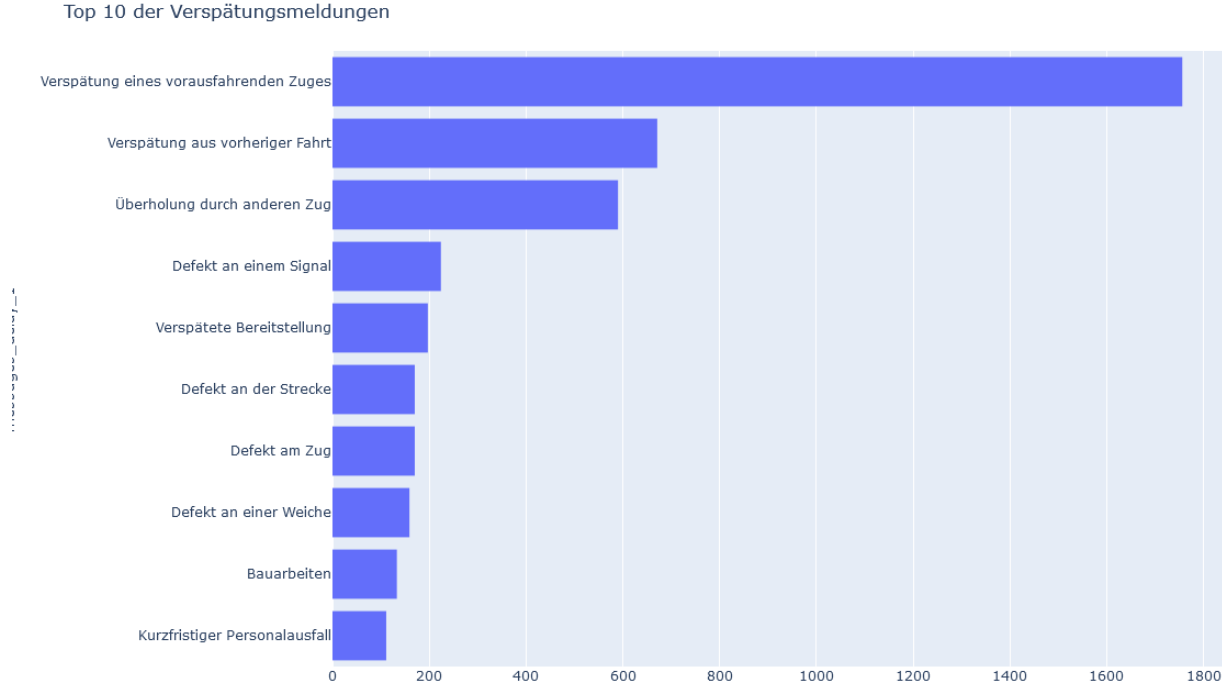
- ◆ *messages.qos*:
 - „Unannehmlichkeiten“, wie z.B. geänderte Wagenreihung
- ◆ *messages.delay*:
 - „Störungen“, wie z.B. Zugüberholungen, Defekte, Polizeieinsätze

* Vermutlich Zusammenhang zwischen Verspätungsmeldung & tatsächlicher Verspätung

Sample: Verspätungsnachrichten

	trainNumber	date	train	start_delay_departure	final_delay_arrival	messages_delay_1	messages_delay_2	messages_delay_3	messages_delay_4
5570	26410	2022-05-08	RE 4	0.0	20.0	Defekt an einem Signal	None	None	None
4255	10916	2022-04-22	RE 9	0.0	11.0	Defekt am Zug	None	None	None
3667	10909	2022-05-19	RE 9	0.0	1.0	None	None	None	None
8810	26830	2022-07-03	RE 1	0.0	34.0	None	None	None	None
11013	32568	2022-06-22	RE 7	13.0	31.0	Defekt an einem Bahnübergang	Verspätung aus vorheriger Fahrt	None	None

Welche Gründe gibt es für Verspätungen?

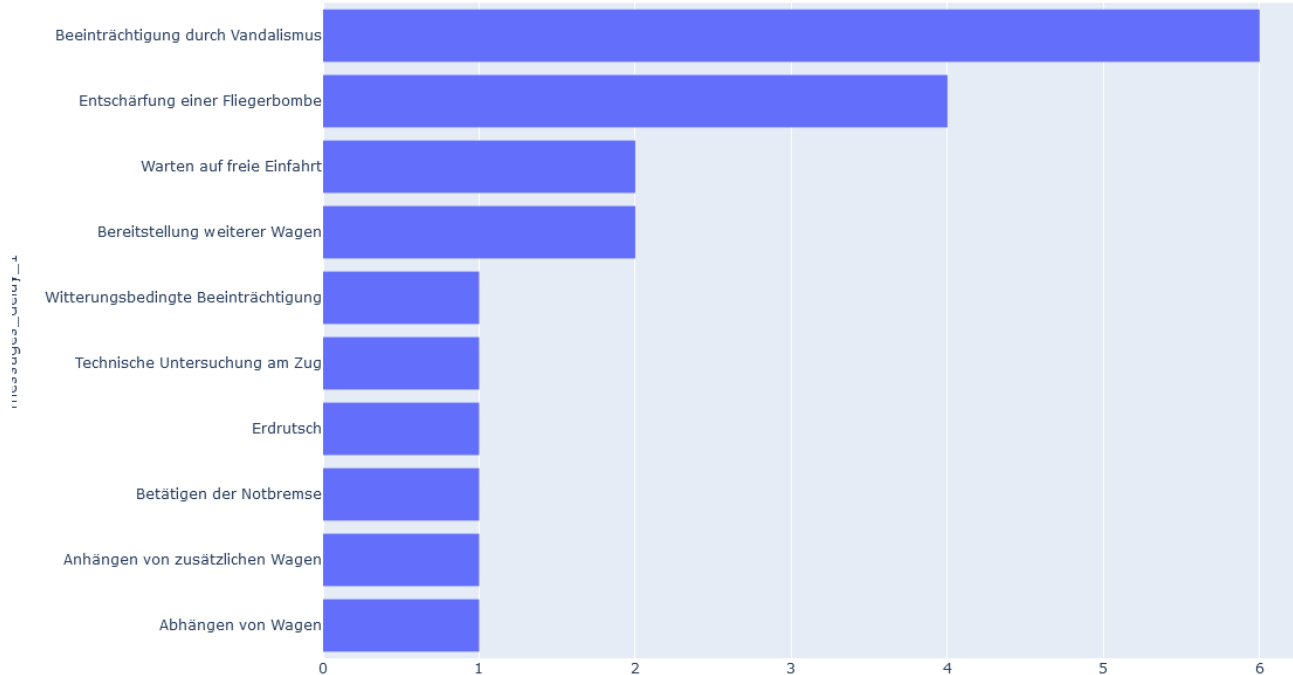


Deutsche Bahn in die Karten geschaut

Data Science & OSEMN-Pipeline | Die Daten & unser Ziel | OSEMN-Pipeline am Beispiel | Ende

Welche Gründe gibt es für Verspätungen?

Die 10 seltensten Verspätungsmeldungen

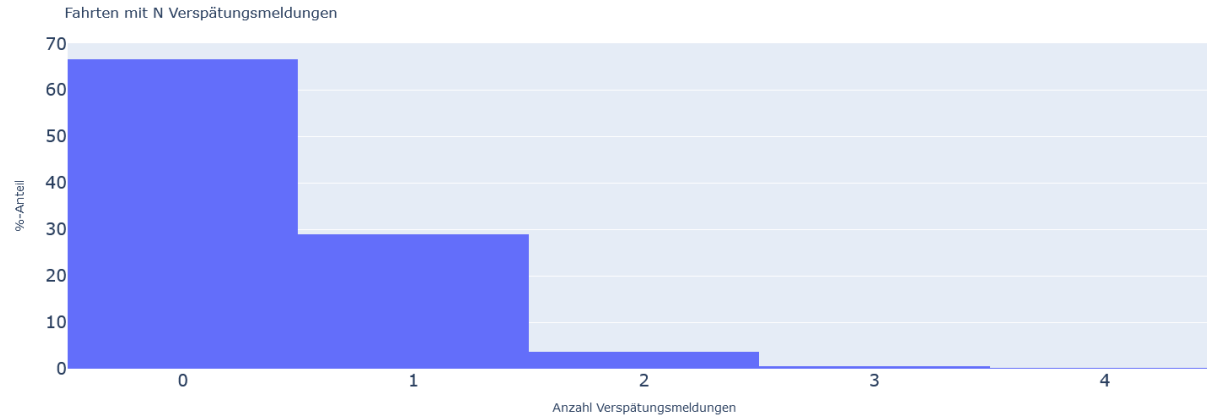


Deutsche Bahn in die Karten geschaut

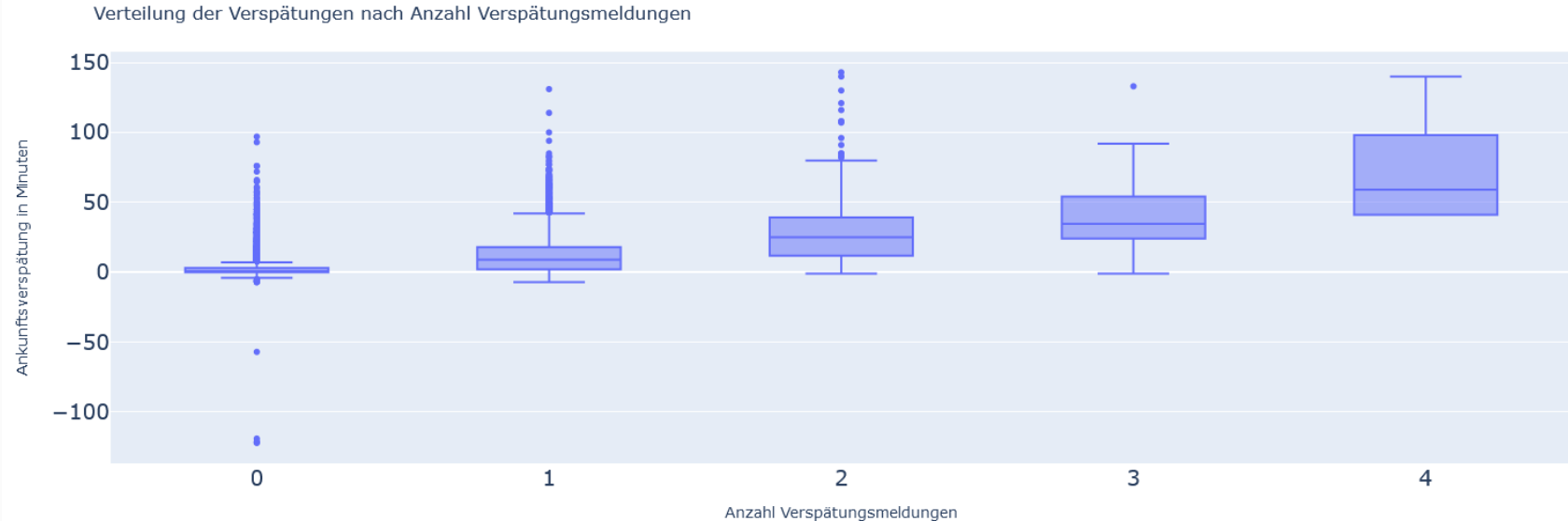
Data Science & OSEMN-Pipeline | Die Daten & unser Ziel | OSEMN-Pipeline am Beispiel | Ende

Anzahl Fahrten mit N Verspätungsmeldungen

❄ Vermutung: Mehr Verspätungsmeldungen → mehr Verspätung



Wie hängen Verspätungen und Anzahl Verspätungsmeldungen zusammen?



Agenda

- Data Science & die OSEMN-Pipeline
- Die Daten & unser Ziel
- OSEMN-Pipeline am Beispiel
 - ◆ Explorative Datenanalyse
 - ◆ ML-Modelle: Training & Ergebnisse

Wie trainiere ich ein ML-Modell?

1. Daten vorbereiten

- ◆ Aufteilung in Trainings-/Testdaten
- ◆ ggf. weitere Vorverarbeitung (z.B. One-Hot-Encoding)

2. Modell trainieren mit gewählten Algorithmen

3. Modell nutzen/Performance messen

1. Daten vorbereiten

* ML-Modelle benötigen numerische Daten

- ◆ Kategorische Daten müssen „übersetzt“ werden (z.B. durch One-Hot-Encoding)

* Aufteilung unserer Daten in Trainings-/Testdaten

2. Modell trainieren mit gewählten Algorithmen

- ✧ Wir trainieren zunächst ein Basismodell als „Benchmark“
- ✧ Anschließend trainieren wir weitere Algorithmen und vergleichen die Performance

Basismodell trainieren

- ✧ Um unsere trainierten Modelle beurteilen zu können, braucht es eine „Benchmark“
- ✧ Abhängig von der Problemstellung, z.B.:
 - ◆ bestehende Lösung (regelbasiert/ML-Modell)
 - ◆ Heuristiken
- ✧ Wir trainieren 2 Basismodelle
 - ◆ Modell 1: „Die Bahn ist immer pünktlich“
 - ◆ Modell 2: Modell sagt immer Durchschnittsverspätung voraus
- ✧ Basismodelle helfen Performance unserer trainierten Modelle zu beurteilen

Basismodell trainieren

```
from sklearn.dummy import DummyRegressor

dummy_mean = DummyRegressor(strategy="mean")
dummy_always_on_time = DummyRegressor(strategy="constant", constant=0)
dummy_mean.fit(X_train, y_train)
dummy_always_on_time.fit(X_train, y_train);
```

```
measure_performance({"Modell: Die Bahn ist immer pünktlich": dummy_always_on_time,
                    "Modell: Durchschnittliche Verspätung" : dummy_mean,},
                    X_test, y_test, round_digits=2)
```

	Modell	Root Mean Squared Error	Mean Absolute Error	R ² Score
0	Modell: Die Bahn ist immer pünktlich	14.88	7.36	-0.306992
1	Modell: Durchschnittliche Verspätung	14.42	7.11	-0.227737

Weitere Modelle trainieren

- ✱ Nachdem wir unsere Benchmark gesetzt haben, trainieren wir weitere Modelle

- ✱ Folgende Modelle betrachten wir:
 - ◆ LinearRegression
 - ◆ DecisionTreeRegressor
 - ◆ RandomForestRegressor

Weitere Modelle trainieren

```
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor

lin_model = LinearRegression() # Lineares Modell
lin_model.fit(X_train, y_train)

tree_reg = DecisionTreeRegressor() # Einfacher DecisionTree
tree_reg.fit(X_train, y_train)

forest_reg = RandomForestRegressor() # Random Forest
forest_reg.fit(X_train, np.ravel(y_train))
```

```
measure_performance({"Modell: Die Bahn ist immer pünktlich": dummy_always_on_time,
                    "Modell: Durchschnittliche Verspätung" : dummy_mean,
                    "LinearRegression": lin_model, "DecisionTreeRegressor": tree_reg,
                    "RandomForestRegressor": forest_reg},
                    X_test, y_test, round_digits=2)
```

	Modell	Root Mean Squared Error	Mean Absolute Error	R ² Score
0	Modell: Die Bahn ist immer pünktlich	14.88	7.36	-0.306992
1	Modell: Durchschnittliche Verspätung	14.42	7.11	-0.227737
2	LinearRegression	9.70	5.78	0.444747
3	DecisionTreeRegressor	11.18	6.13	0.262072
4	RandomForestRegressor	10.29	5.88	0.374205

Feature Engineering

- ✦ Modell verbessern, indem Daten um Features angereichert werden
 - ◆ z.B. Datumsinformationen: Wochentag, Werktag/kein Werktag, Ferientag/kein Ferientag, Tagesstunde..
 - ◆ Aus Routeninformationen: abweichende Route ja/nein

Beispiel: Feature Engineering

- ✱ Aus dem Datum erstellen wir drei weitere Spalten
 - ◆ `date_after_introduction_9_euro_ticket`:
 - Ist das Datum vor/nach Einführung des 9-Euro-Tickets?
 - ◆ `is_school_holiday`:
 - Ist das Datum ein Schulferientag in NRW?
 - ◆ `is_working_day`:
 - Ist das Datum ein Werktag?

Performance mit zusätzlichen Features

	Modell	Root Mean Squared Error	Mean Absolute Error	R² Score
0	Modell: Die Bahn ist immer pünktlich	14.88	7.36	-0.306992
1	Modell: Durchschnittliche Verspätung	14.42	7.11	-0.227737
2	LinearRegression	9.63	5.79	0.452077
3	DecisionTreeRegressor	11.39	6.17	0.233520
4	RandomForestRegressor	10.09	5.76	0.398597

Zusammenfassung der Ergebnisse

- ❖ Daten weisen einige Qualitätsprobleme auf
 - ◆ Negative Verspätungen, Verspätungen ohne Verspätungsmeldungen, ..

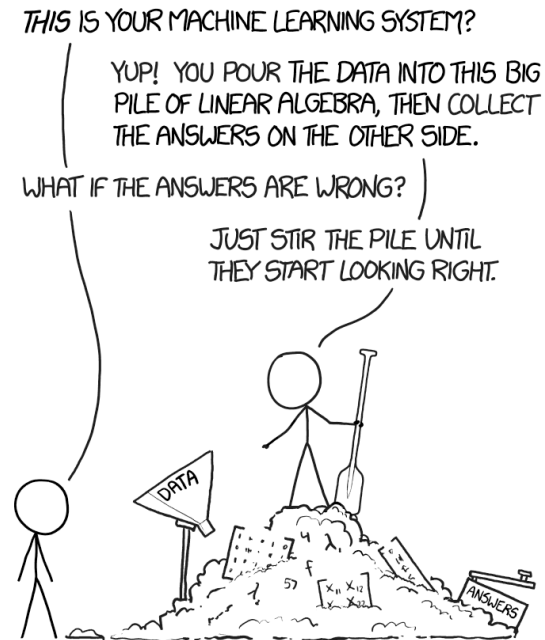
- ❖ Trotzdem: trainiertes Modell ist ~19% besser als Baseline

	Modell	Root Mean Squared Error	Mean Absolute Error	R ² Score
0	Modell: Die Bahn ist immer pünktlich	14.88	7.36	-0.306992
1	Modell: Durchschnittliche Verspätung	14.42	7.11	-0.227737
2	LinearRegression	9.63	5.79	0.452077
3	DecisionTreeRegressor	11.39	6.17	0.233520
4	RandomForestRegressor	10.09	5.76	0.398597

- ❖ Trainiertes Modell ist ein erster MVP, den wir nun in die Produktion bekommen wollen

Ausblick

- ❖ Diesen MVP gilt es stetig zu verbessern, z.B. durch:
 - ◆ weitere „Experimente“ -> Explorative Datenanalyse, Feature Engineering, ..
- ❖ Mehr und bessere Daten, z.B.
 - ◆ Wetterdaten
 - ◆ Signale von den Zwischenstationen (Verspätungsverläufe)
 - ◆ Gleis/Bahnhofsauslastung
 - ◆ Baustelleninformationen (abfragbar per API)
- ❖ Bei der Kommunikation der Ergebnisse hilft u.a. MLOps



Impulsvorträge für Ihr Unternehmen

✿ Überblick über das gesamte Angebot an Impulsvorträgen unter:
www.iks-gmbh.com/impulsvortraege

✿ Ihr Nutzen:

- ◆ Unabhängiges, aktuelles Expertenwissen.
- ◆ Individuell auf Ihr Publikum und Ihr Unternehmen zugeschnittene Vorträge.
- ◆ Referenten mit langjähriger und branchenübergreifender Expertise in der IT-Beratung.
- ◆ Praxisnahe Vorträge, die aus Projektarbeit entstanden sind, frei von Produktwerbung.
- ◆ Ideale Ergänzung für Ihre Führungskräftetreffen, Abteilungsm Meetings, Hausmessen, Innovation Days, Konferenzen, Open Spaces, Kick-off-Meetings oder Zukunftswerkshops.

WWW.IKS-GMBH.COM



Projekte. Beratung. Spezialisten.

