Digna

Datenqualitätsmanagement mit Künstlicher Intelligenz

Ein Jahr ohne Datenregeln im Data Warehouse

dext.ai Danijel Kivaranovic, PhD

Dipl.-Ing. Vojkan Radak itSV

Agenda

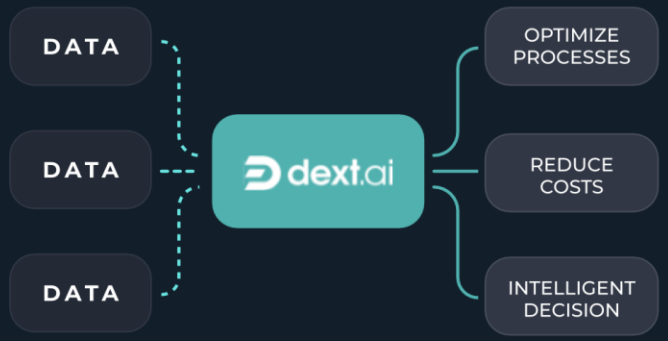
- Über DEXT.AI GmbH
- Warum brauchen wir Datenqualitätsmanagement?
- Anomalie-Erkennung mit Künstlicher Intelligenz
- Data Warehouse Landschaft der Sozialversicherungen
- Ein Tag eines engagierten DWH Produkt Managers

Über DEXT.AI GmbH

Dext.ai is an AI and ML software company based in Vienna

Our mission

Is to help customers unlock the powers of their data by building intelligent, user-friendly AI/ML-powered solutions. Since our inception in 2020, Dext.ai set out to help brands and companies unlock the powers of their data, optimize processes, reduce costs and make intelligent decisions through innovative, user-friendly AI/ML-powered solutions.



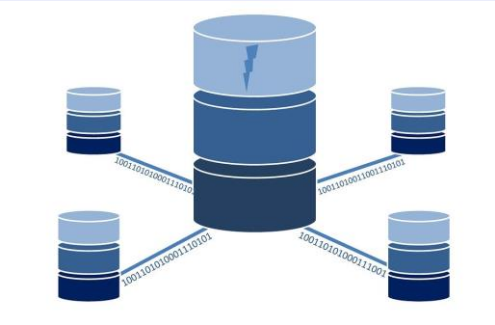
Warum brauchen wir Datenqualitätsmanagement?

Warum brauchen wir Datenqualitätsmanagement?

Können wir nicht einfach unsere Daten einer Künstlichen Intelligenz geben und dann wird alles gut?

Die Erwartung

Schlechte Daten

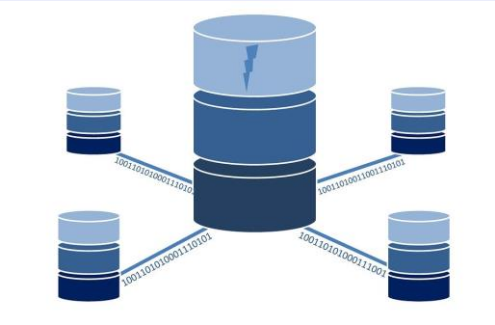


Künstliche Intelligenz

Ergebnis

Die Realität

Schlechte Daten



Künstliche Intelligenz

Ergebnis

Beispiel: ChatGPT

Der Algorithmus hinter ChatGPT wird auf Freitexten trainiert und kann per-se nicht zwischen „guten“ und “schlechten“ Texten unterscheiden.

Daher mussten Hunderte von Arbeitern pro Tag 150 bis 250 Textpassagen bereinigen und klassifizieren, damit ChatGPT lernt was “gut“ und was „schlecht“ ist.

https://time.com/6247678/openai-chatgpt-kenya-workers/

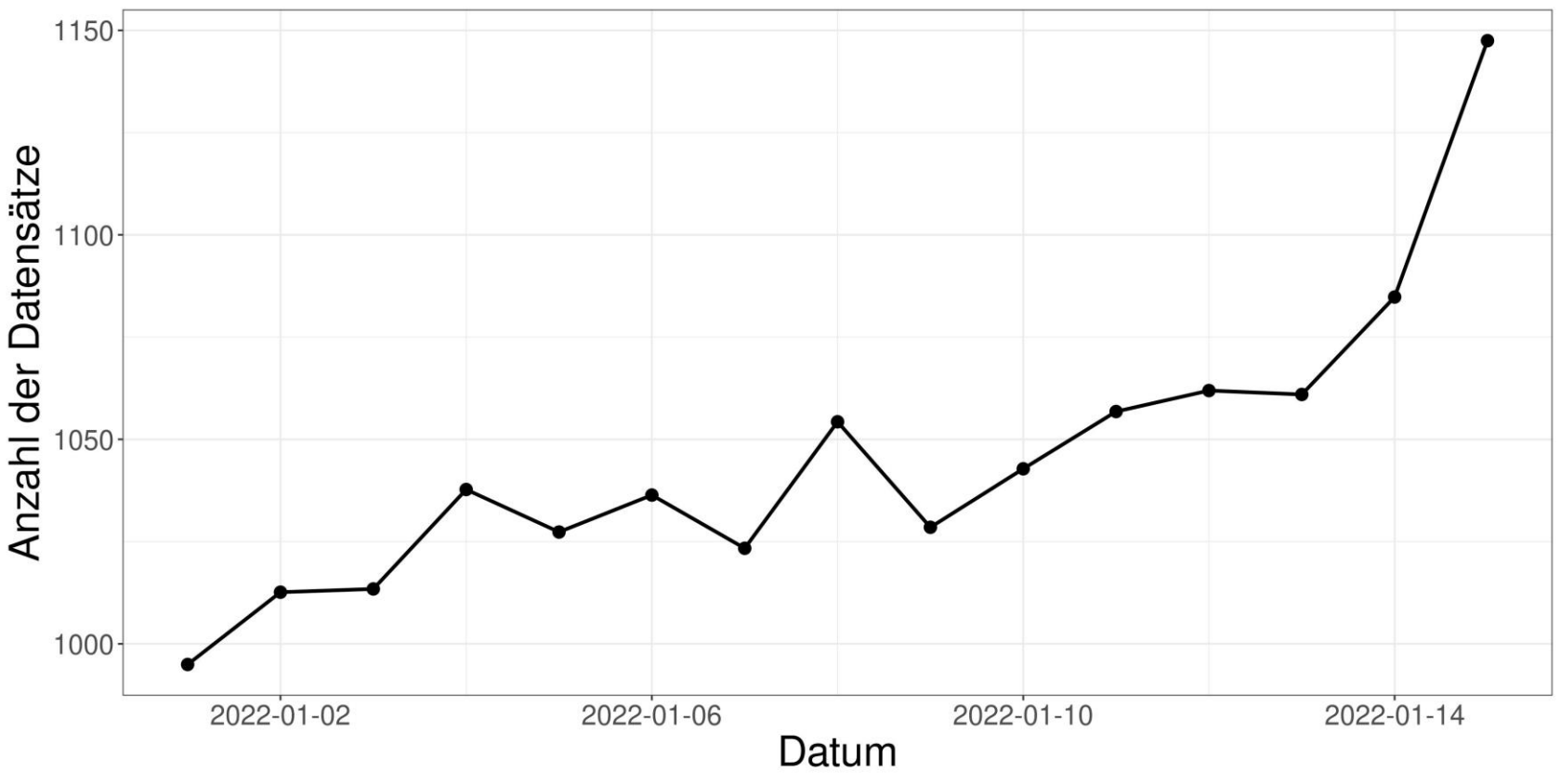
Das Fazit

Um ein erfolgreiches Daten-getriebenes Produkt zu entwickeln, musst Du

- einen Überblick über die Eigenschaften deiner Daten haben
- die Qualität deiner Daten ständig überwachen
- Prozesse haben, wie Du mit schlechten Daten umgehst

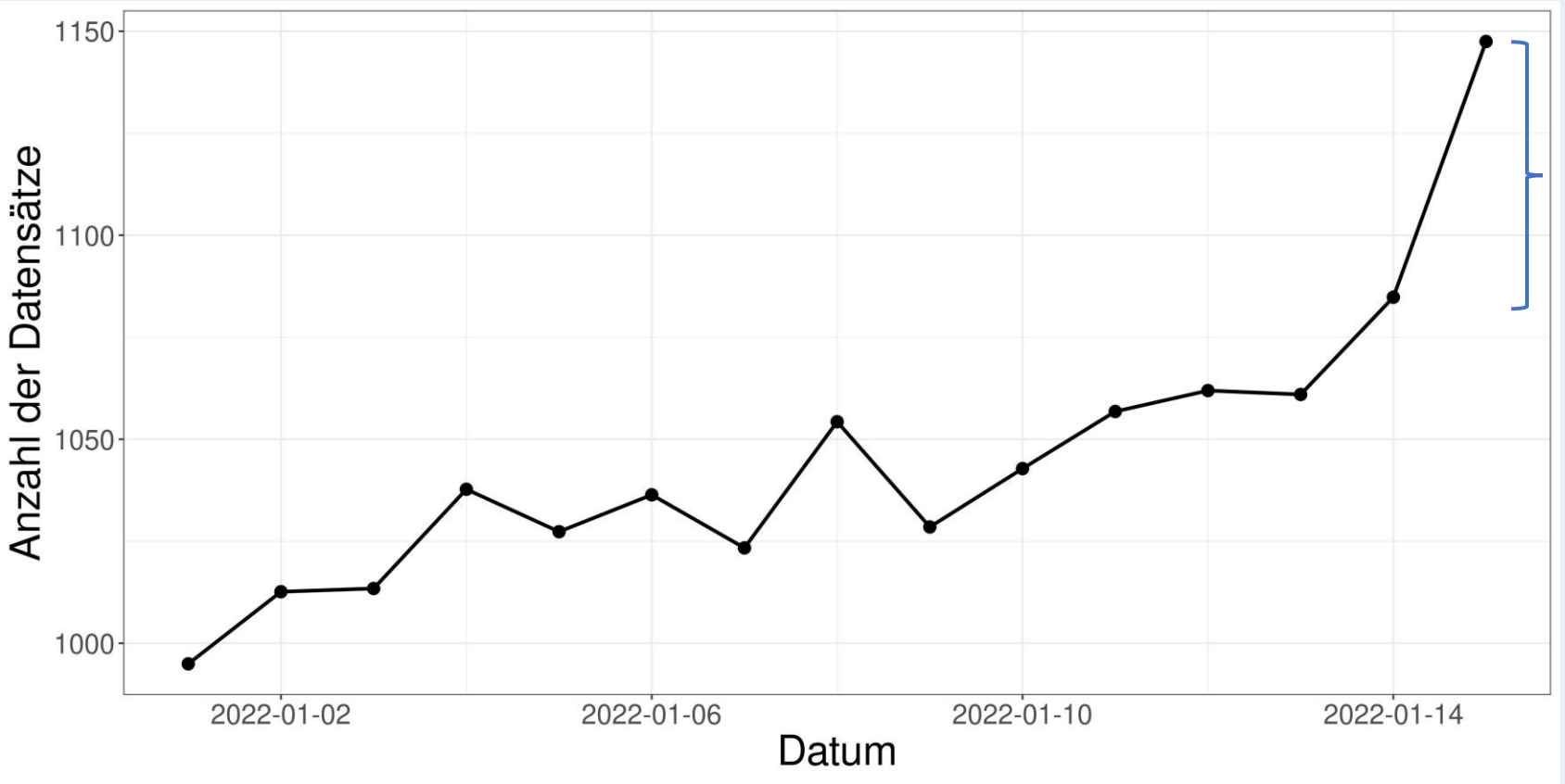
Anomalie-Erkennung mit Künstlicher Intelligenz

Was ist eine Anomalie?



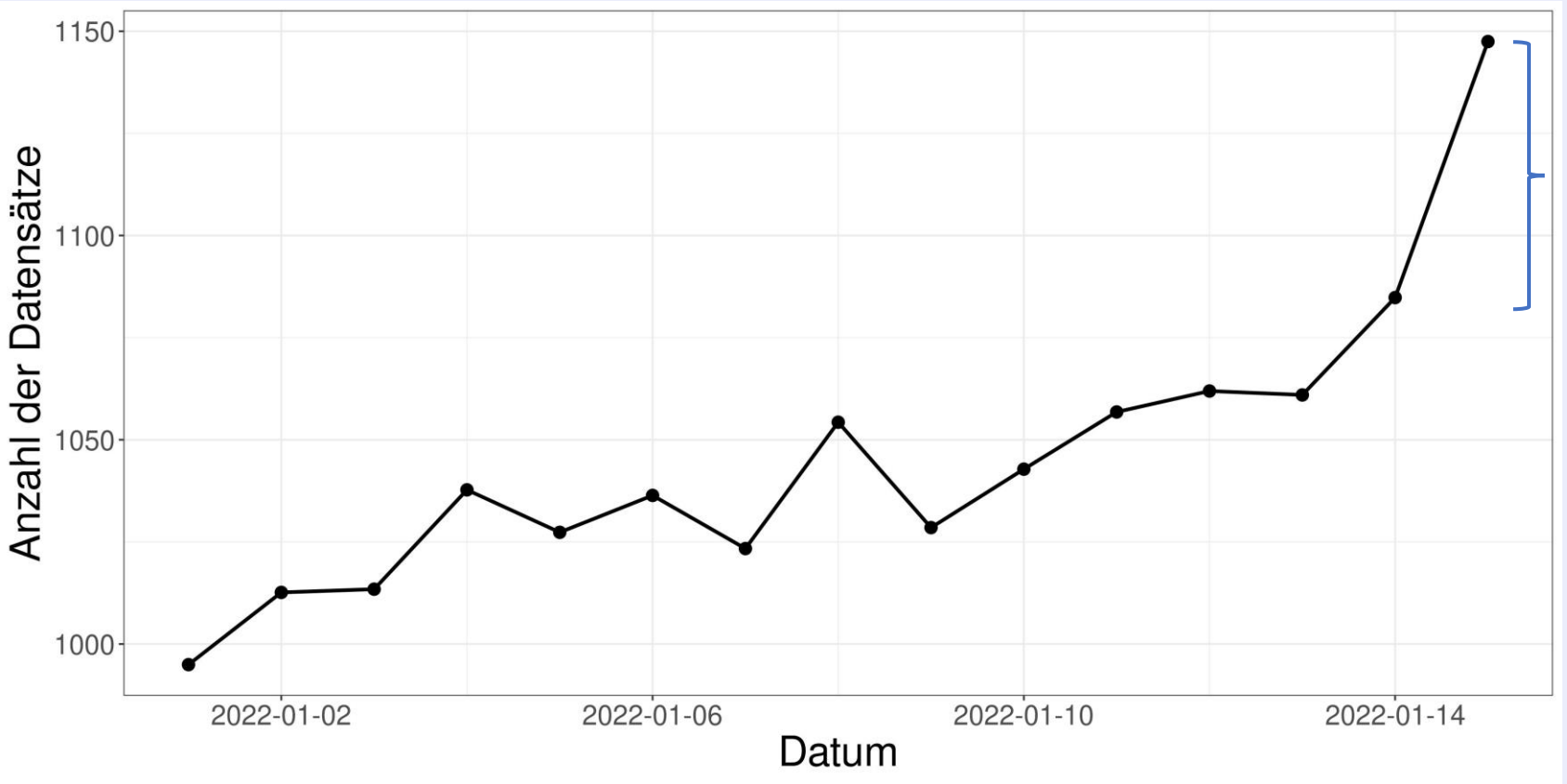
Anzahl der Datensätze in einer Tabelle vom 1.1.2022 bis 15.1.2022

Was ist eine Anomalie?



Anzahl der Datensätze in einer Tabelle vom 1.1.2022 bis 15.1.2022

Was ist eine Anomalie?



In der Praxis:

Der Analyst ist sich nicht sicher und erstellt eine Regal, dass er bei mehr als 10% Änderung informiert wird.

Anzahl der Datensätze in einer Tabelle vom 1.1.2022 bis 15.1.2022

Wie viele Daten-Qualitäts-Regeln bräuchtest du nun bei einer Tabelle mit 100 Spalten?

- Wie viele fehlende Werte sind normal in einer Spalte? 80%, 25%, 0%?
- Wie viele verschiedene Werte hast Du in einer Spalte? Erwartest du neue Werte?
- Wie sieht die Verteilung der numerischen Spalten aus? Ist es normal, dass der Mittelwert um 50 sich verändert an einem Tag?

Anomalie-Erkennung mit Digna

Anzahl der Datensätze in einer Tabelle vom 1.1.2022 bis 15.1.2022

Anomalie-Erkennung mit Digna

Unsere Künstliche Intelligenz definiert voll automatisiert die Bereiche:

GRÜN = Keine Anomalie
GELB = Grenzbereich
ROT = Anomalie

Anzahl der Datensätze in einer Tabelle vom 1.1.2022 bis 15.1.2022

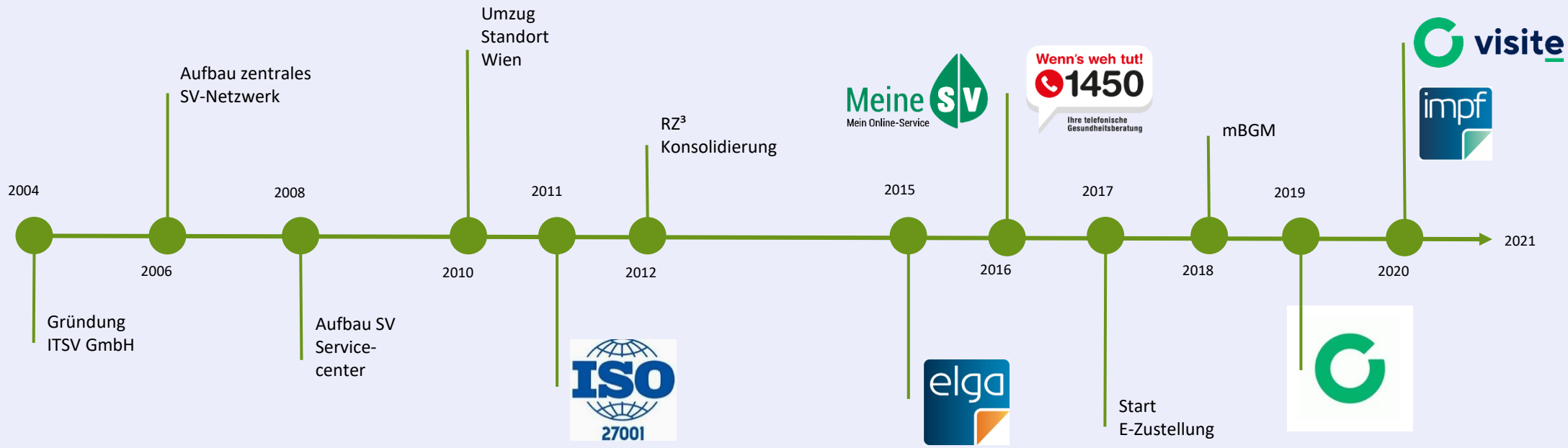
Data Warehouse Landschaft der Sozialversicherungen

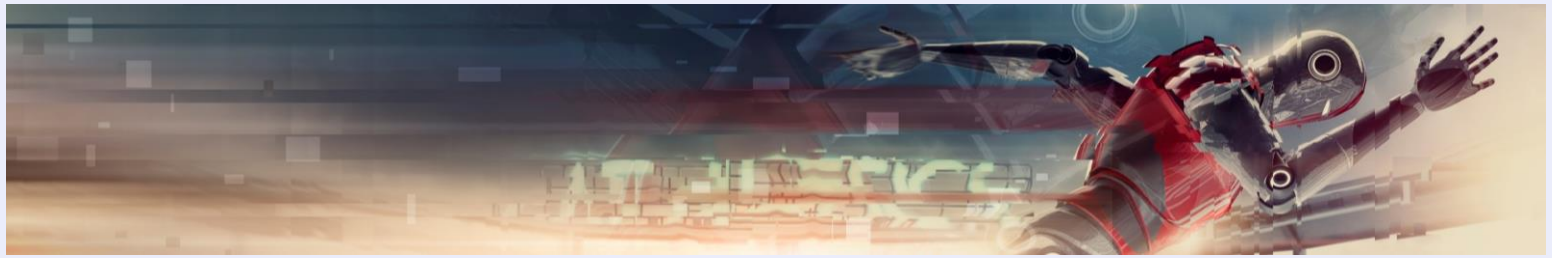
Vorstellung der ITSV GmbH

- 100%iges Tochterunternehmen der österreichischen Sozialversicherung
- Full-Service IT-Provider
- Wir begleiten 9 Millionen ÖsterreicherInnen ein Leben lang.
- Als Management-GmbH zur Entwicklung der IT-Strategie, übergreifenden Steuerung und Koordinierung ist unser primärer Auftrag Kundenorientierung, Datensicherheit und Kosteneffizienz.
- So konnten seit 2007 über 390 Mio. Euro IT-Kosten in der Sozialversicherung eingespart werden.

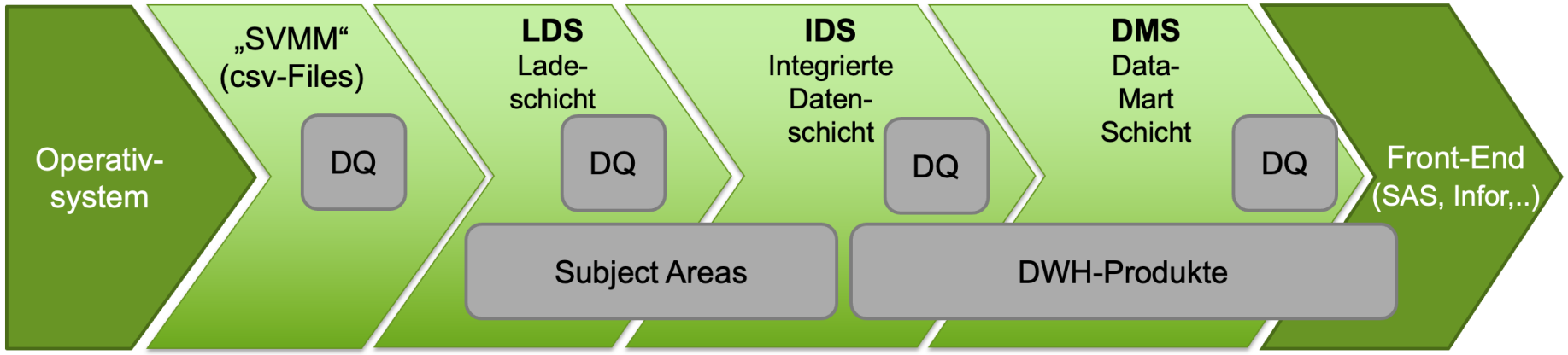
Wir digitalisieren das österreichische Gesundheitswesen. #herzderdigitalisierung

Vorstellung der ITSV GmbH





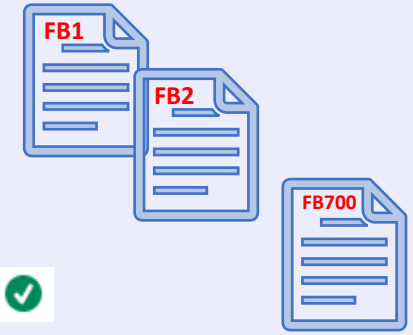
Grundarchitektur der DWH-Landschaft der SV



DWH in Zahlen

- ITSV DWH “Data Lake“
    - 40 Lieferanten der Quelldaten
    - 250.000 Lieferungen pro Jahr
    - 500 verschiedene Strukturen
    - **Tägliches „Delta“ ~ 50 Millionen** Datensätze
    - 7000 Felder
    - 55 Milliarden Datensätze
- Gesamte DWH/s
    - 100 Datenbanken
    - 400 Strukturen
    - 500 Milliarden Datensätze

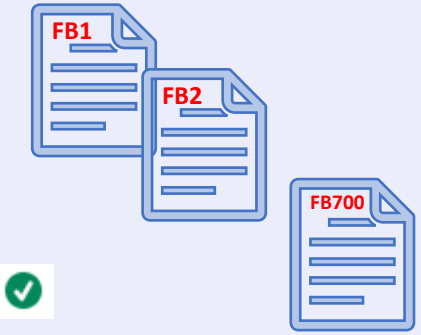
DWH in Zahlen



- Herausforderungen
  - Datumsfelder mit vielen verschiedenen Formaten geliefert ✅
  - Referential Integrity (PK, FK, …) wen nicht in Quellsystemen erledigt ✅
  - Numeric Felder mir CHAR Inhalten ✅
  - Unüblich andere Anzahl der Datensätze … mehrfache oder leere Lieferungen
  - Lieferungen verspäten sich …kommen nicht „wie üblich“ (täglich, alle 3/4 Tage, monatlich …) an …
  - Neue, unbekannte Werte, bei Dimensionen, die als „statisch“ betrachtet sind … Berichtsinkonsistenz
  - Datenmengen (hunderte Strukturen, tausende Felder, …)

- Was bisher geschah… Wir haben es wirklich versucht…
  - Checks: Felder: Datum, Numeric, Längen, …, PK,…, FK,… ✅
  - Bei Produkten, die als „kritisch“ gezeichnet sind, obere/untere Grenzen der Datensatzanzahl bei gewählten Felder, definieret und auf „Ausreißer“ geprüft
  - Verschiedene statistische Methoden: MAX, MIN, AVG, %, Quantille, K-Means, **~ 9000 Rules** … ausprobiert …

DWH in Zahlen



- Herausforderungen
  - Datumsfelder mit vielen verschiedenen Formaten geliefert ✅
  - Referential Integrity (PK, FK, …) wen nicht in Quellsystemen erledigt ✅
  - Numeric Felder mir CHAR Inhalten ✅
  - Unüblich andere Anzahl der Datensätze … mehrfache oder leere Lieferungen
  - Lieferungen verspäten sich …kommen nicht „wie üblich“ (täglich, alle 3/4 Tage, monatlich …) an …
  - Neue, unbekannte Werte, bei Dimensionen, die als „statisch“ betrachtet sind … Berichtsinkonsistenz
  - D…de Felder, …)

9000 Rules !

- Was b…cht…
  - Checks: Felder: Datum, Numeric, Längen, …, PK,…, FK,… ✅
  - Bei Produkten, die als „kritisch“ gezeichnet sind, obere/untere Grenzen der Datensatzanzahl bei gewählten Felder, definieret und auf „Ausreißer“ geprüft
  - Verschiedene statistische Methoden: MAX, MIN, AVG, %, Quantille, K-Means, **~ 9000 Rules** … ausprobiert …

Ein Tag eines engagierten DWH Produkt Managers

Tag eines DWH Produkt Mangers mit 9000 Rules…

Tägliche DWH Datenqualität Checks

- Am Arbeitsplatz:
  - Rote Punkte suchen in Excel Sammelbericht von 700 Excels Berichten, kritisch anschauen
  - 243 von 700 Berichten, sind rot
  - Erster Bericht > „weiß ich schon längst“ …müssen wir die Rules anpassen
  - Zweiter Bericht > „weiß ich schon längst“ …müssen wir die Rules anpassen
  - …
  - Elfter Bericht … > Kaffee
- Bei der Kaffee-Maschine
  - „Gleich passe ich die Rules an“
  - „Heute war doch der Release von den neuen DWH Produkt?“
  - „hoffentlich geht alles OK“
- Am Arbeitsplatz:
  - drrrriiiiinnngggg! Telefon abheben … Problem mit den neuen DWH Produkt ? …
  - OK! Schaue ich gleich an …

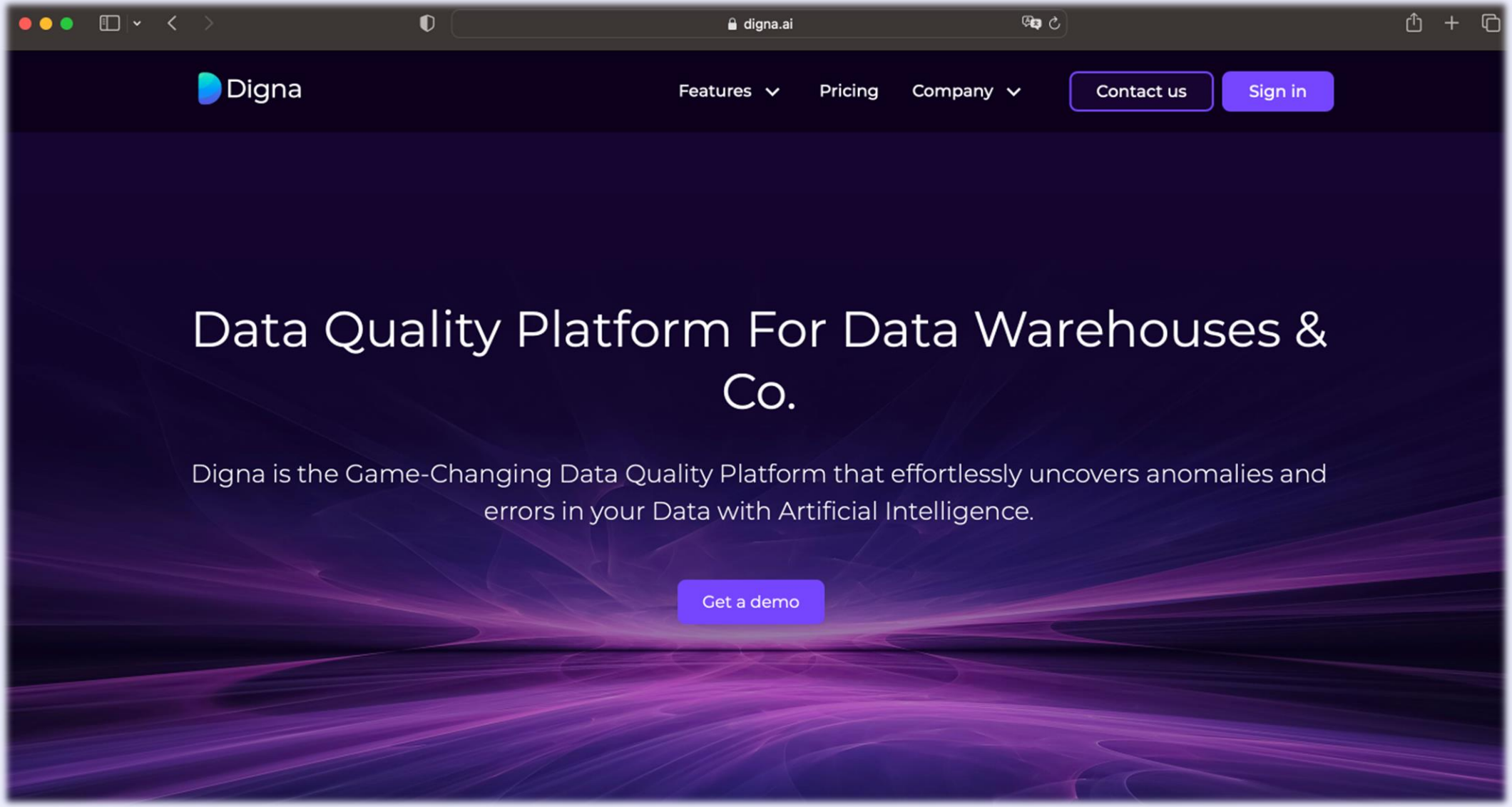
Tag eines DWH Produkt Mangers mit 9000 Rules…

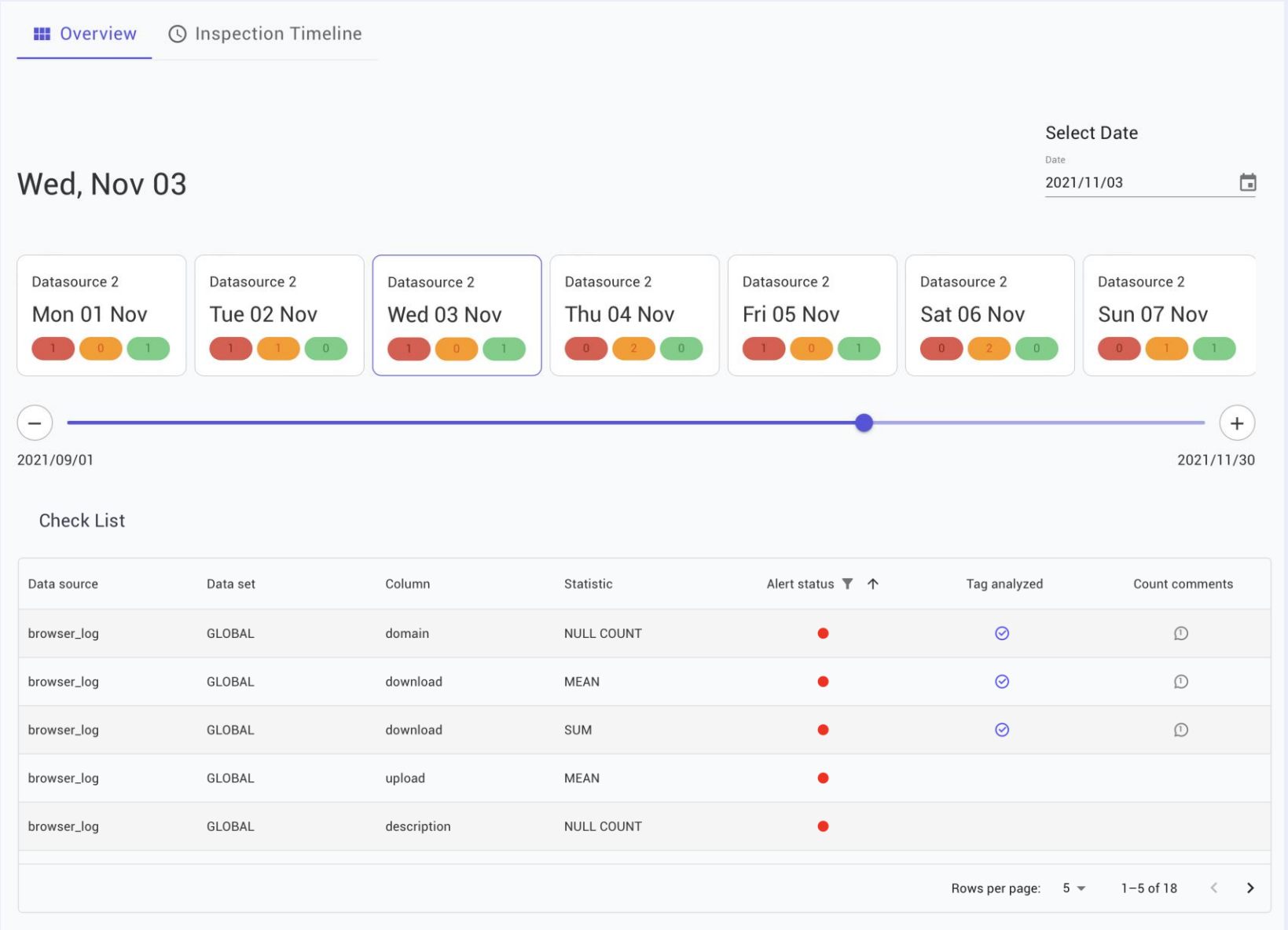
Tägliche DWH Datenqualität Checks

- Am Arbeitsplatz:
  - Rote Punkte suchen in Excel Sammelbericht von 700 Excels Berichten, kritisch anschauen
  - 243 von 700 Berichten, sind rot
  - Erster Bericht > „weiß ich schon längst“ …müssen wir die Rules anpassen
  - Zweiter Bericht > „weiß ich schon längst“ …müssen wir die Rules anpassen
  - …
  - Elfter Bericht … > Kaffee
- Bei der Kaffee-Maschine
  - „Gleich passe ich die Rules an“
  - „Heute war doch der Release von den neuen DWH Produkt?“
  - „hoffentlich geht alles OK“
- Am Arbeitsplatz:
  - drrrriiiiinnnggggg! Telefon abheben … Problem mit den neuen DWH Produkt ? …
  - OK! Schaue ich gleich an …

… und die Rules ?

… bleiben unverändert !!!


digna.ai
Digna
Features
Pricing
Company
Contact us
Sign in
Data Quality Platform For Data Warehouses & Co.
Digna is the Game-Changing Data Quality Platform that effortlessly uncovers anomalies and errors in your Data with Artificial Intelligence.
Get a demo


Overview
Inspection Timeline
Wed, Nov 03
Select Date
Date
2021/11/03
Datasource 2
Mon 01 Nov
1 0 1
Datasource 2
Tue 02 Nov
1 1 0
Datasource 2
Wed 03 Nov
1 0 1
Datasource 2
Thu 04 Nov
0 2 0
Datasource 2
Fri 05 Nov
1 0 1
Datasource 2
Sat 06 Nov
0 2 0
Datasource 2
Sun 07 Nov
0 1 1
2021/09/01
2021/11/30
Check List
Data source
Data set
Column
Statistic
Alert status
Tag analyzed
Count comments
browser_log GLOBAL domain NULL COUNT
browser_log GLOBAL download MEAN
browser_log GLOBAL download SUM
browser_log GLOBAL upload MEAN
browser_log GLOBAL description NULL COUNT
Rows per page: 5
1–5 of 18

Tag eines DWH Produkt Mangers ohne 9000 Rules…

Tägliche DWH Datenqualität Checks

- Am Arbeitsplatz:
  - Mails checken. Oh … DIGNA schreibt …: 11 rote Punkte
  - Erster Punkt: interessant… **neuer** Impfstoff wird verwendet
  - Zweiter Punkt: aha, **Verspätung** bei der Lieferung …vielleicht wieder Scheduler Probelme bei Lieferant: xy …den schicken wir ein Mail
  - Dritter Punkt: WOW, **der neue DWH Produkt** Bericht gibt echt Gas ! **So viele** Web Klicks hat **keiner bisher** geschafft.
  - Vierter Punkt: **Anzahl** der „long running queries“ erhöht – 90 ! Das muss ich mal schauen wer schon wieder nicht optimierte Queries in der Produktion laufen lässt.
  - …
  - Elfter Punkt: Aha „Anzahl der Minuten ohne Kaffee“ erhöht - 47 !
- Bei der Kaffee-Maschine
  - „und was machst du am Wochenende?“
  - „Heute war doch der Release von den neuen DWH Produkt ?“
  - „ja, … und es ist fantastisch **gut gegangen**, so viel Klicks haben wir …“