Die OTTO BI

Wie man hunderte von Data Scientist*innen mit Daten versorgt

Anissa Ayoub

Product Owner @OTTO BI

Alumni FH Wedel Master E-Commerce (Abschluss 2021)

https://www.linkedin.com/in/Anissa-ayoub

anissa.ayoub@otto.de

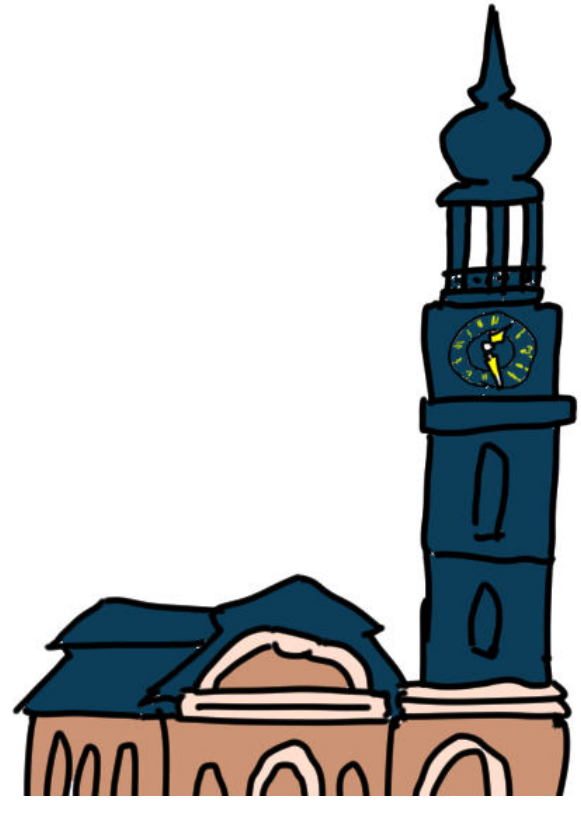
Bjørn Stachmann

https://kapitel26.github.io

@old_stachi

BjoernArno.Stachmann@otto.de

OTTO



business@OTTO

- 5,1 Milliarden € Umsatz
- 10 Millionen Produkte
- 19.000 Marken
- bis zu 10 Bestellungen pro Sekunde
- **otto.de**: 2,9 Millionen Visits pro Tag

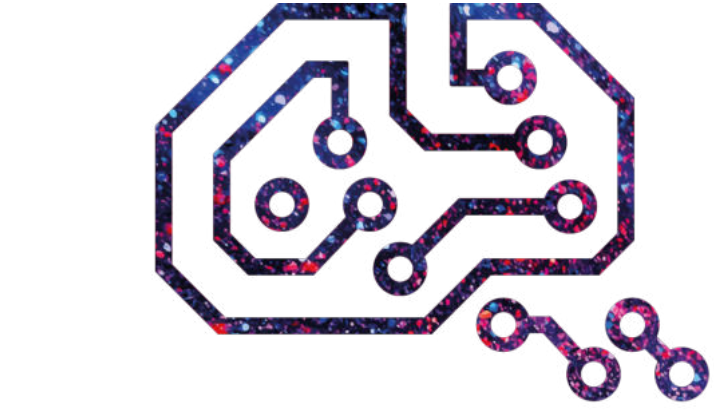

tech@OTTO
• 1000+ Techies
• 100+ Mio. Tech-Investionen

Und wir kommen aus der...

OTTO BI

INNOVATIONS AND EXCELLENT DECISIONS

Wir bauen...



BRAIN

OTTO BI PLATTFORM

Wir sind größer als einige Städte Deutschlands

OTTO BI

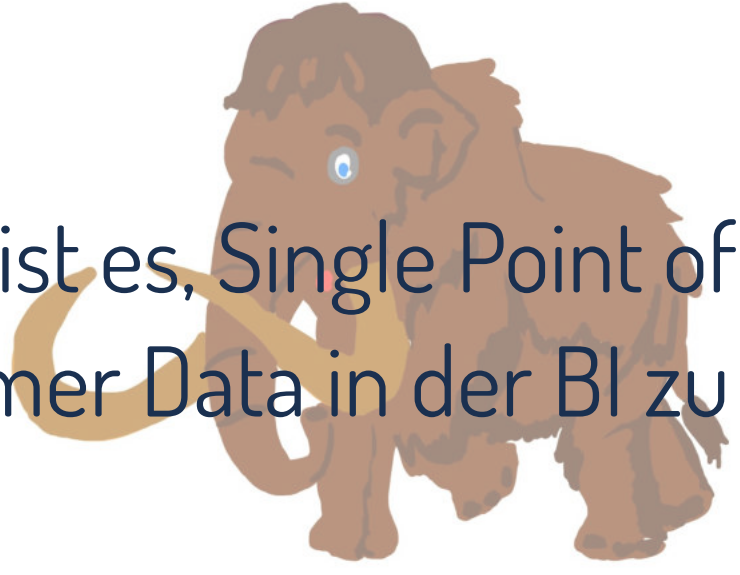
INNOVATIONS AND EXCELLENT DECISIONS

Orga@BI

- 15 Abteilungen
- 400+ Kolleg*innen
- 50+ Teams, agil organisiert

Team MAMMUT

- 4 Data Eng.
- 1 Business Analyst
- 1 Product Owner
- 1 Agile Master



Unsere Vision ist es, Single Point of Truth für Core-Customer Data in der BI zu werden

Es fing alles ganz konventionell an.


REAL TIME
ANALYTIC
BIG DATA
CLASSIC
REPORTING

Aber die Zeiten begannen sich zu ändern.

Sie sind hungrig!

Sehr hungrig!

Sie sind

Data Scientisten!

Wer soll helfen?

Mama BI

- User experience
- Plattform-Partner-Management
- Ausbau der Sortimentsbreite
- Dynamic Pricing
- Logistik

Deshalb wollen wir eine

Data Driven Company

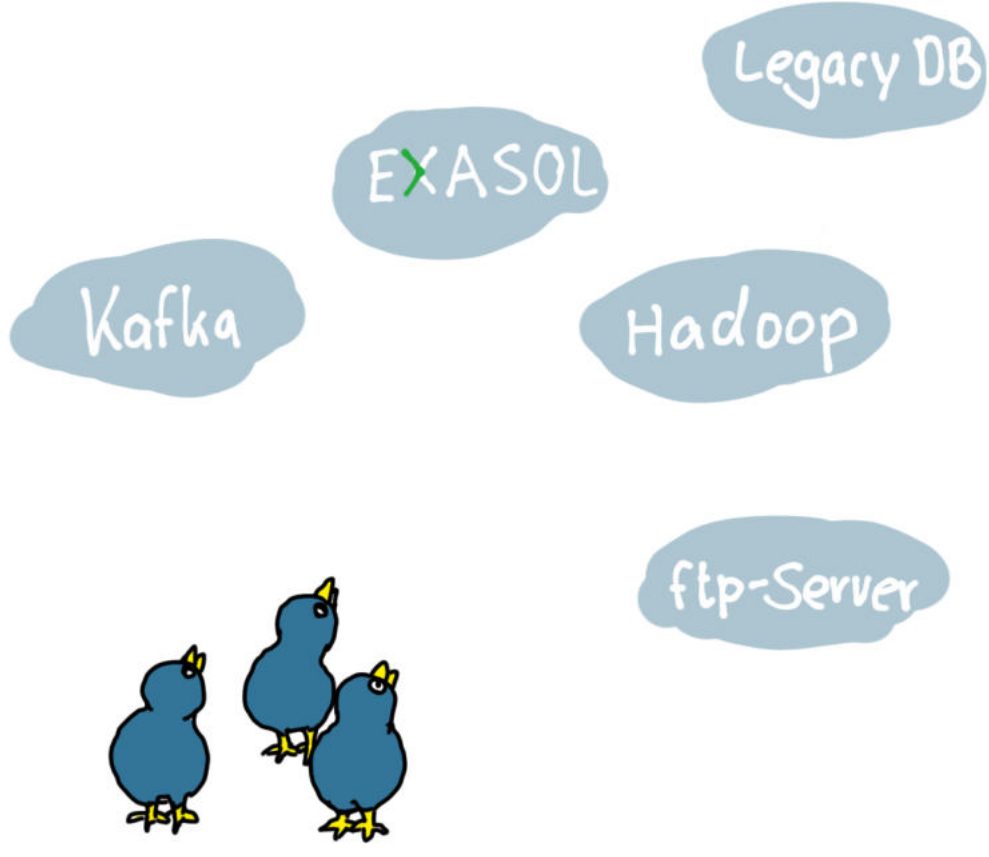
werden.

Unsere Vision

Keiner holt aus Daten & KI mehr raus als OTTO

Daten & KI überall

Aber, so einfach ist das nicht:


Legacy DB
EXASOL
Kafka
Hadoop
ftp-Server

Data Scientisten brauchen
mehr Zeit
für das **Abholen und Aufbereiten**
von Daten
als für Analysen.

Und wie löst man das?

wir bauen ein

Aber was genau ist ein

Data Warehouse?

Moment ... mal eben googeln ...

WIKIPEDIA
Die freie Enzyklopädie

Hauptseite
Themenportale
Zufälliger Artikel

Mitmachen

Artikel verbessern

Data Warehouse

Ein **Data Warehouse** (kurz **DWH** oder **DW**; wörtlich „Datenlager“, im Deutschen dominiert die englische Schreibweise, die Schreibweise *Datawarehouse* wird jedoch auch verwendet) ist eine für Analysezwecke optimierte zentrale Datenbank, die Daten aus mehreren, in der Regel heterogenen Quellen zusammenführt.[1]

Business Intelligence Warehouse

Alle Daten

- aus *verschiedenen Quellen*
- auf *einer Plattform*,
- *bereinigt*, *interpretiert*, ggf. *aggregiert*
- leicht *auffindbar*, *nutzbar* und verknüpfbar

Wie bringen wir Daten aus vielen Quellen in einen Topf?

Our High-Performance Databases

Exasol	**BigQuery**
In Memory dutzende TB	Cloud Storage nahezy unbegrenzt
Super effiziente Joins und Views	Beliebig viele Clients Trennung von Storage & Compute
Skalierbbar derzeit 40 Nodes	Automatische Skalierung fully managed

Wie erreichen wir Verknüpfbarkeit?
Wie schaffen wir es Dutzende von Teams zu koordinieren?
Und Dutzende von Fachbereichen zu versorgen?


Warehouse
Core
Domain
Access

Also alles gut, oder?

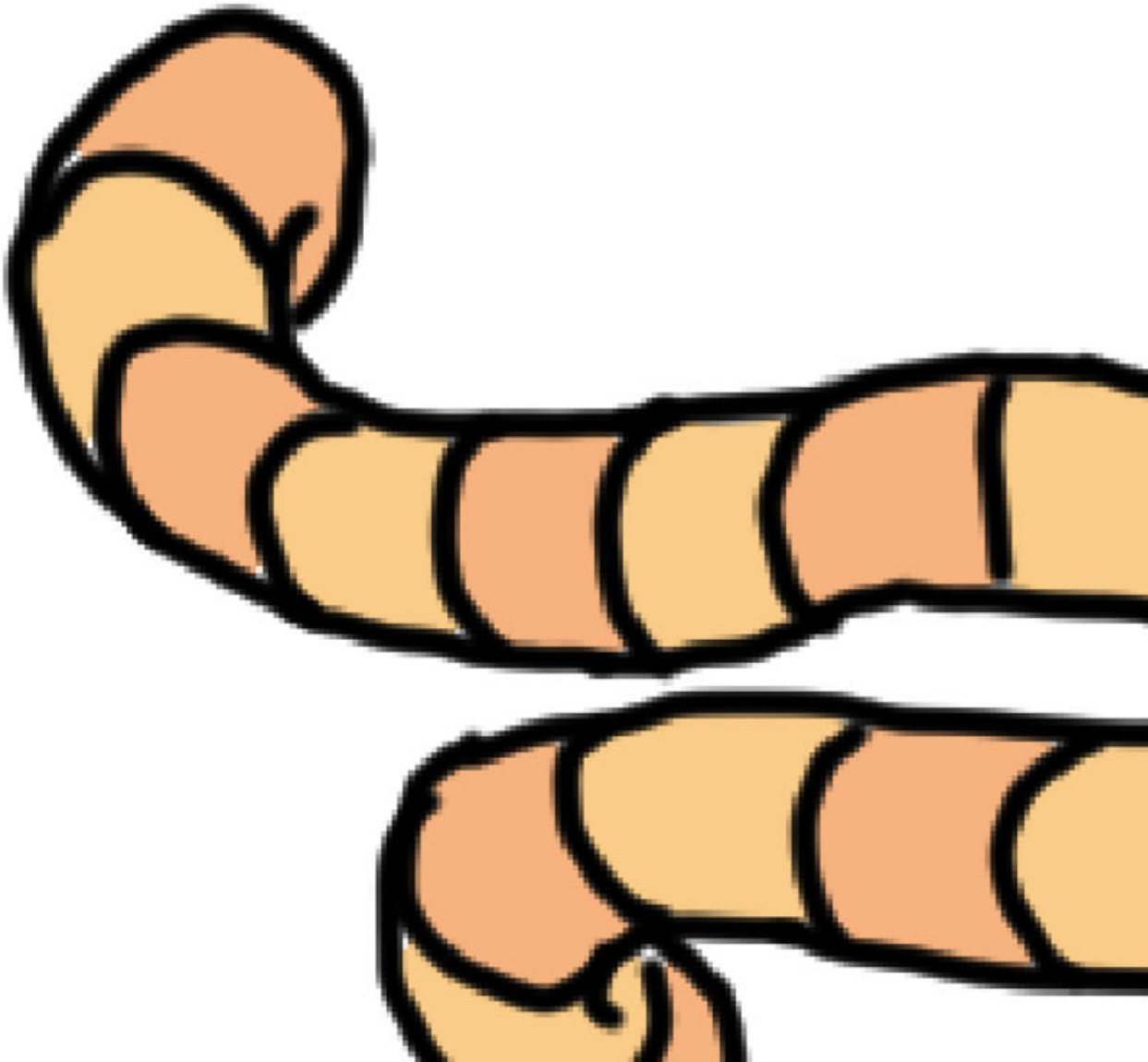
Aber ganz so einfach ist das auch wieder nicht …

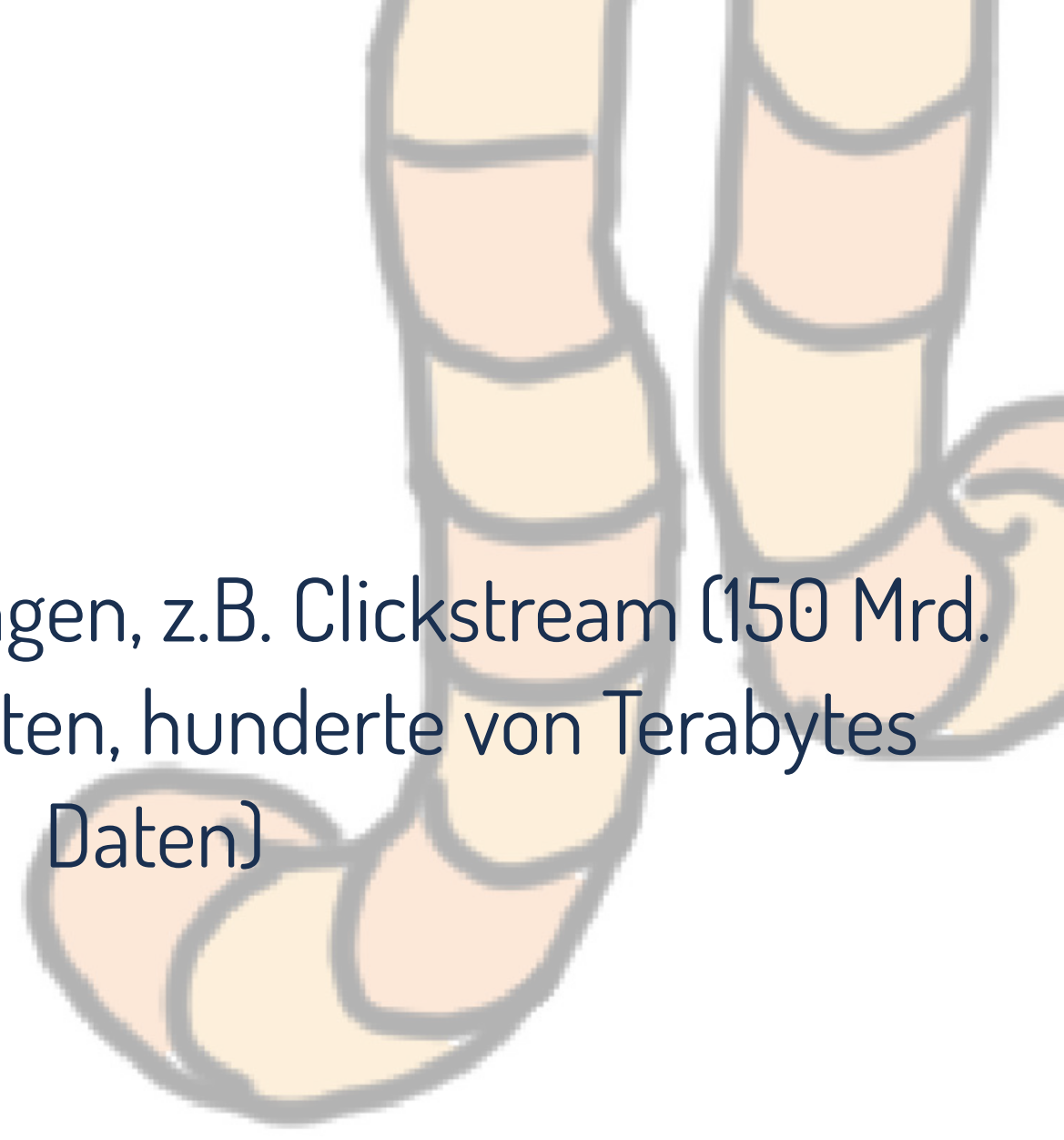
700+ Quellsystemen, zehntausende von Tabellen

lädt man nicht "mal eben" in ein Warehouse.

Werden aber ASAP benötigt.

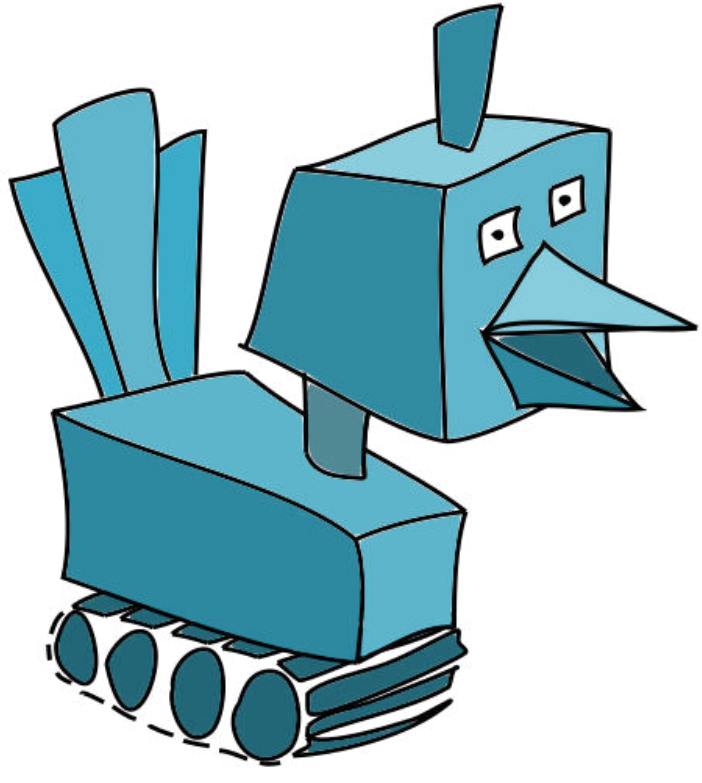
⇝ Bereitstellung zu träge

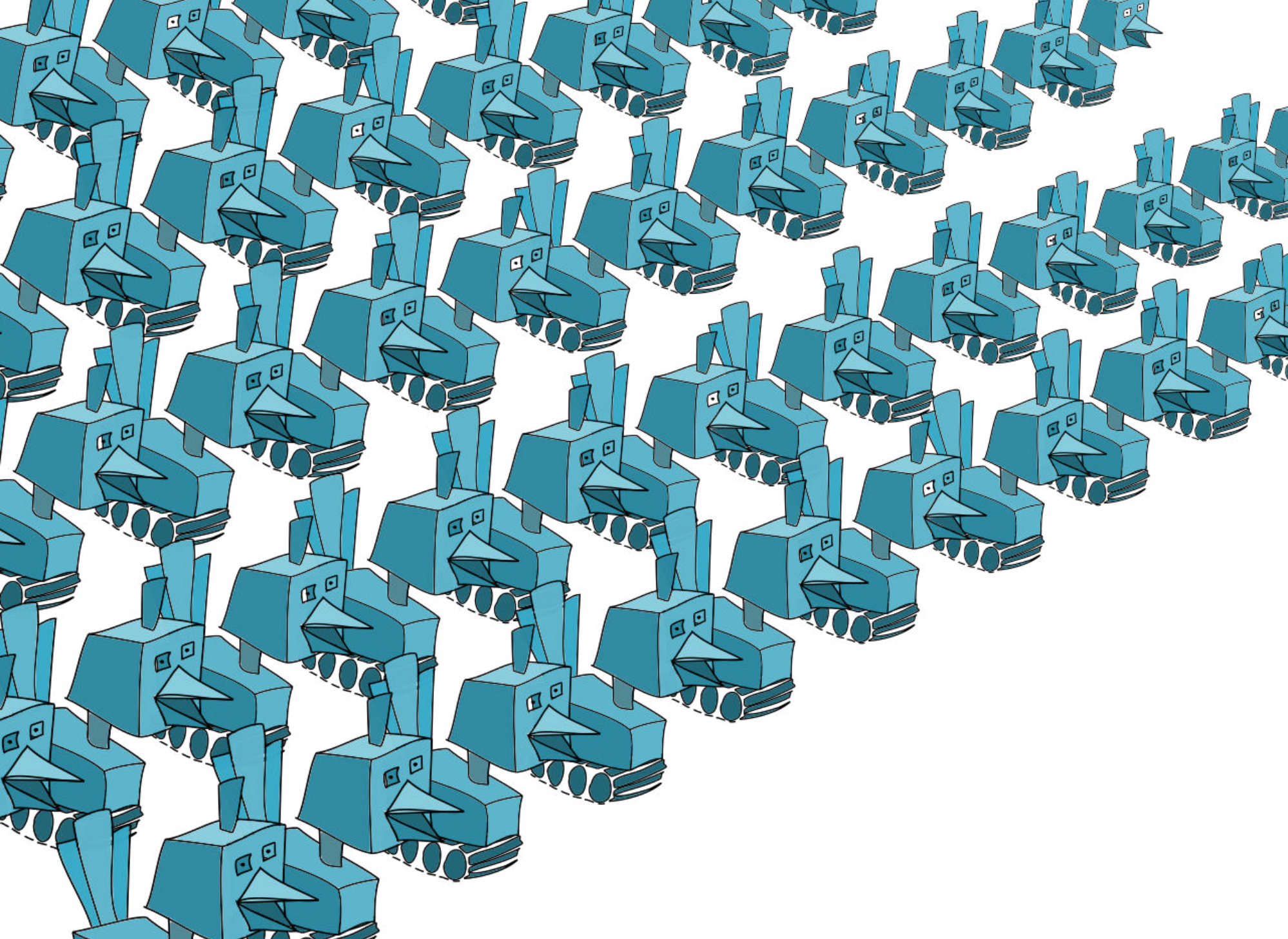




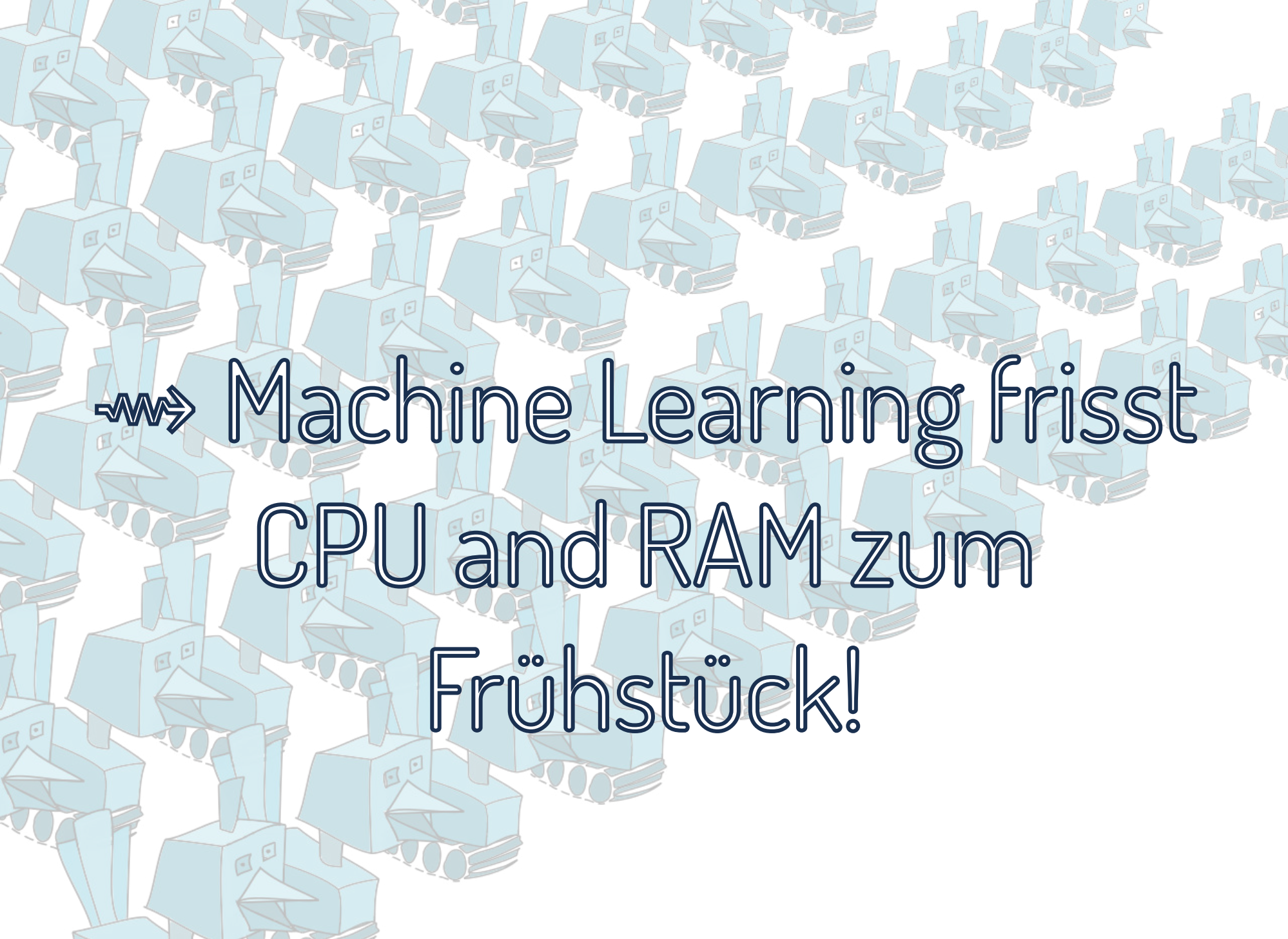
Gewaltige Datenmengen, z.B. Clickstream (150 Mrd. Zeilen, 2000+ Spalten, hunderte von Terabytes Daten)

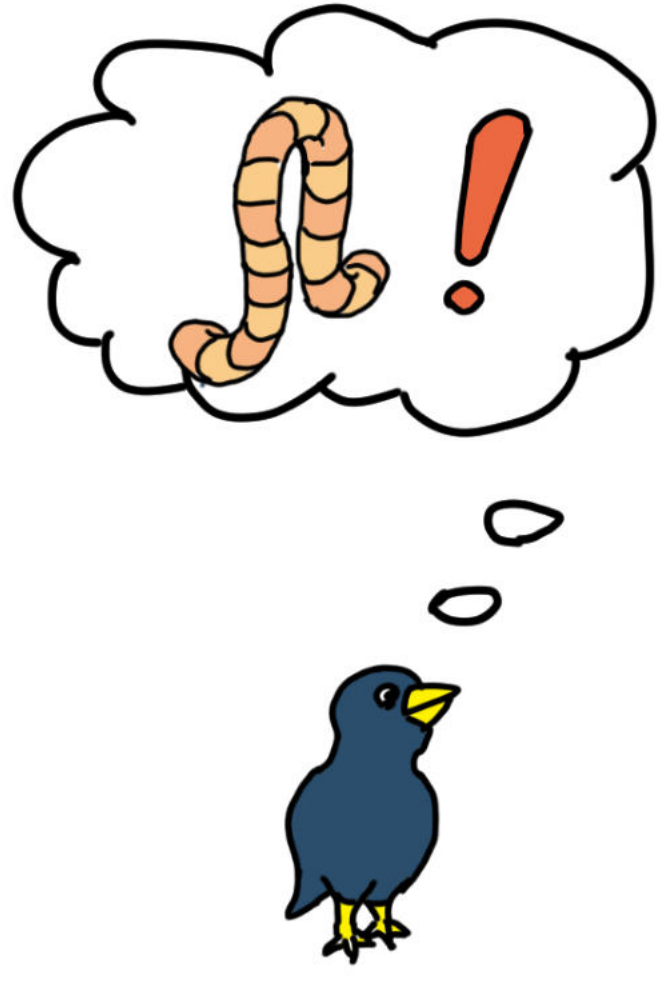

⇝ nicht 1:1 in der in-memory
Exasol abbildbar






Machine Learning
liebt Scans über
alle Zeilen & alle Spalten.


Machine Learning frisst
CPU and RAM zum
Frühstück!



Manche mögen's roh!

Insbesondere Data Scientisten

⇝ Rohdaten erforderlich

Warum kann man nicht einfach **alles** in das große **Warehouse** tun?

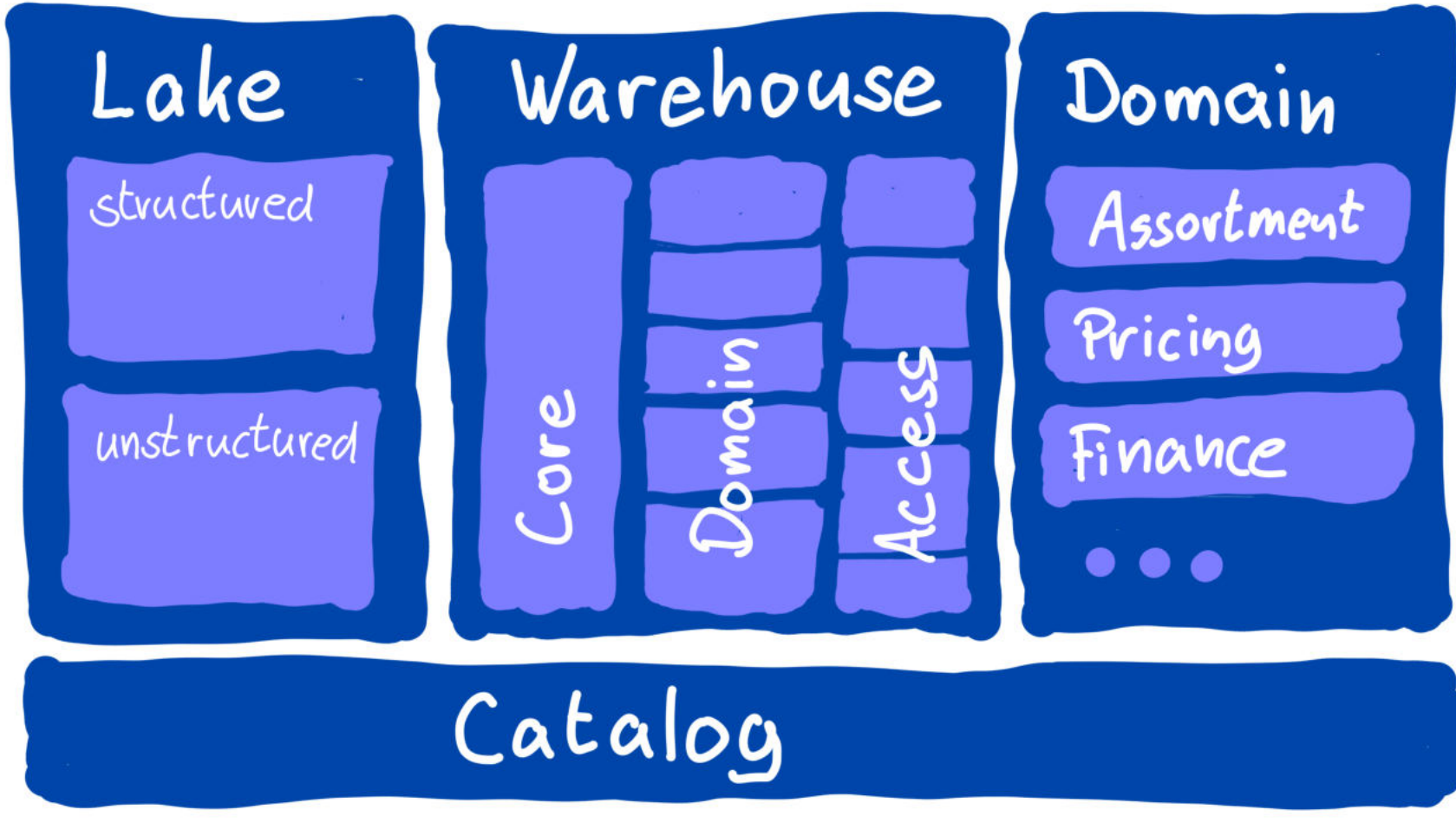
- Bereitstellung zu träge
- nicht 1:1 im Warehouse abbildbar
- ML frisst CPU and RAM
- Rohdaten erforderlich
- Archiv & Back-Up

Wir haben gelernt,

Data Lake und **Warehouse**

zu kombinieren.

Data Lake	Data Warehouse
quellnah	einheitlich modelliert
roh	bereinigt
heterogen	verknüpfbar
semi-structured, binär	columnar, SQL
big data	universell, verknüpfbar
redundanz	single point of truth
Archiv & Backup	


Lake
structured
unstructured
Warehouse
Core
Domain
Access
Domain
Assortment
Pricing
Finance
Catalog

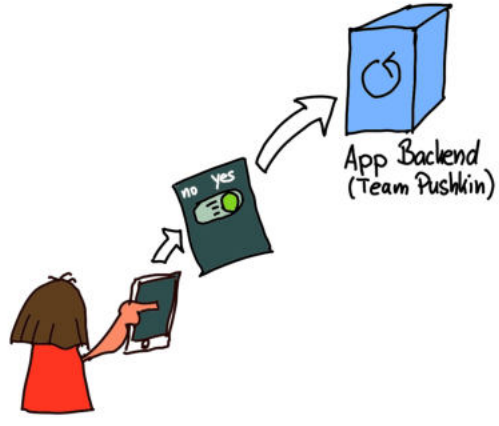
... und dadurch können wir

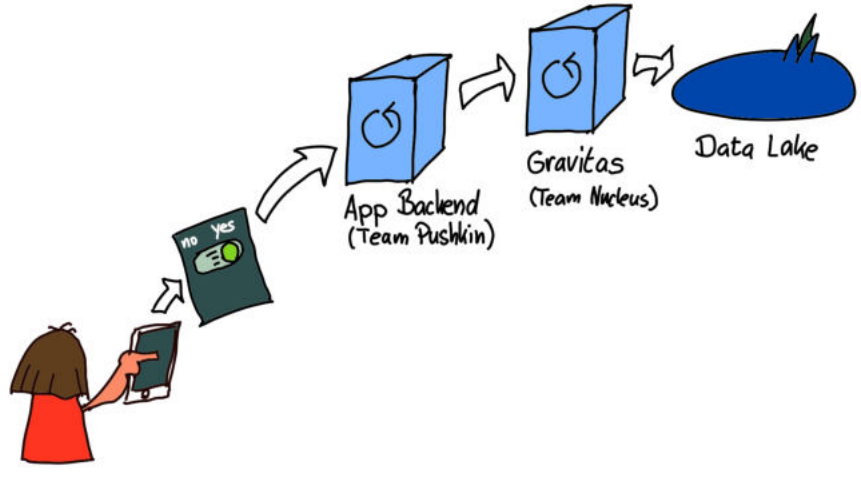
- interactive Queries
- Reporting
- Batch Processing
- Big Data
- Machine Learning

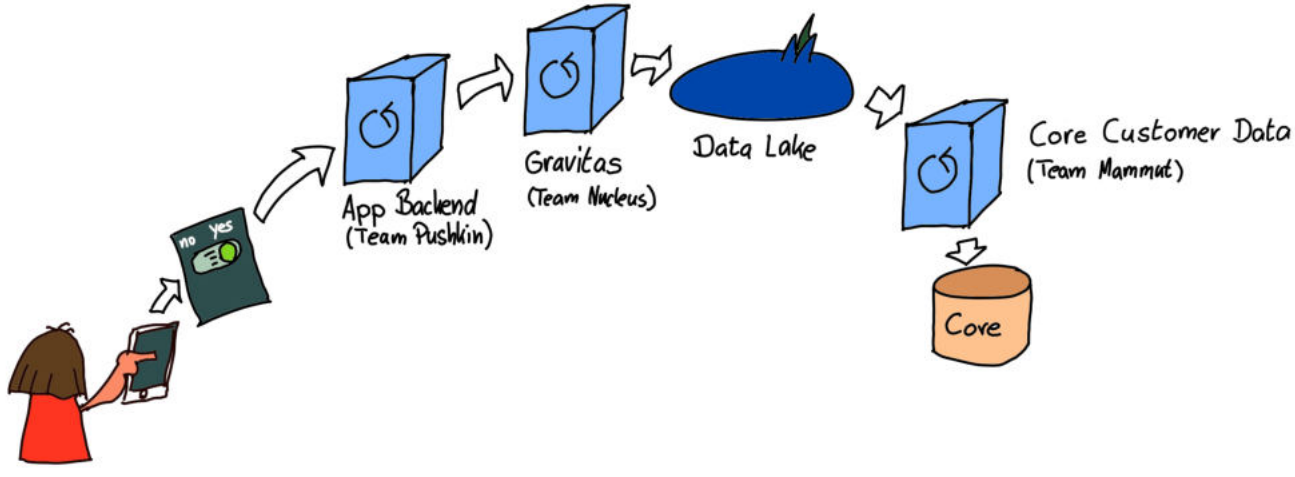
unterstützen.

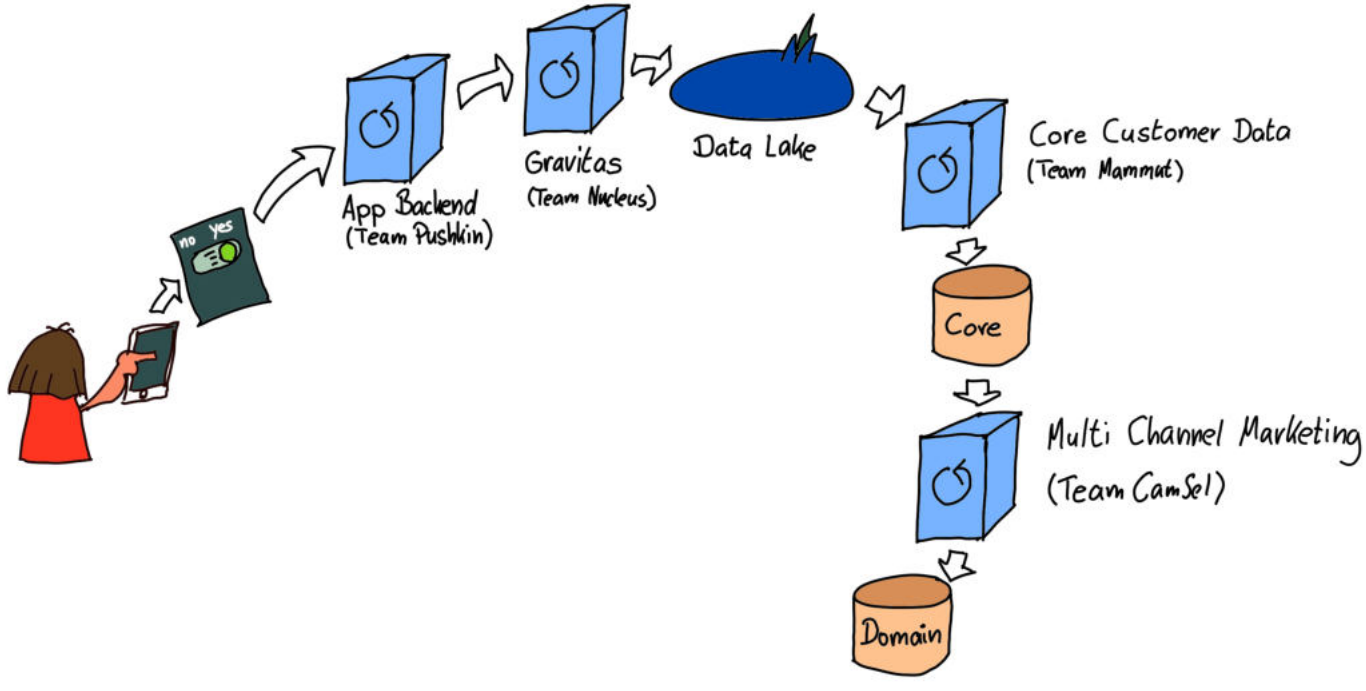
… und was heißt das jetzt in der Praxis?

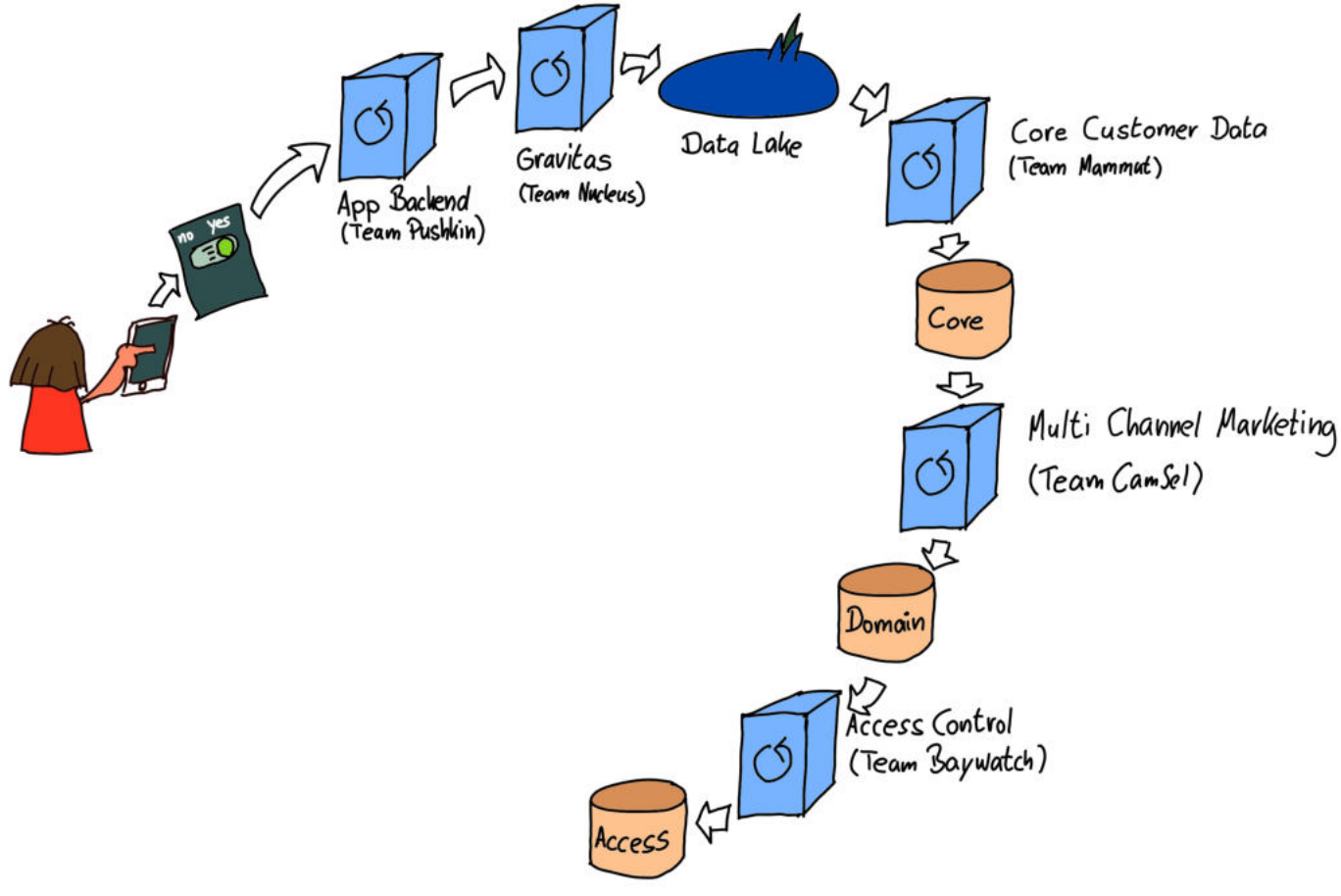

no
yes

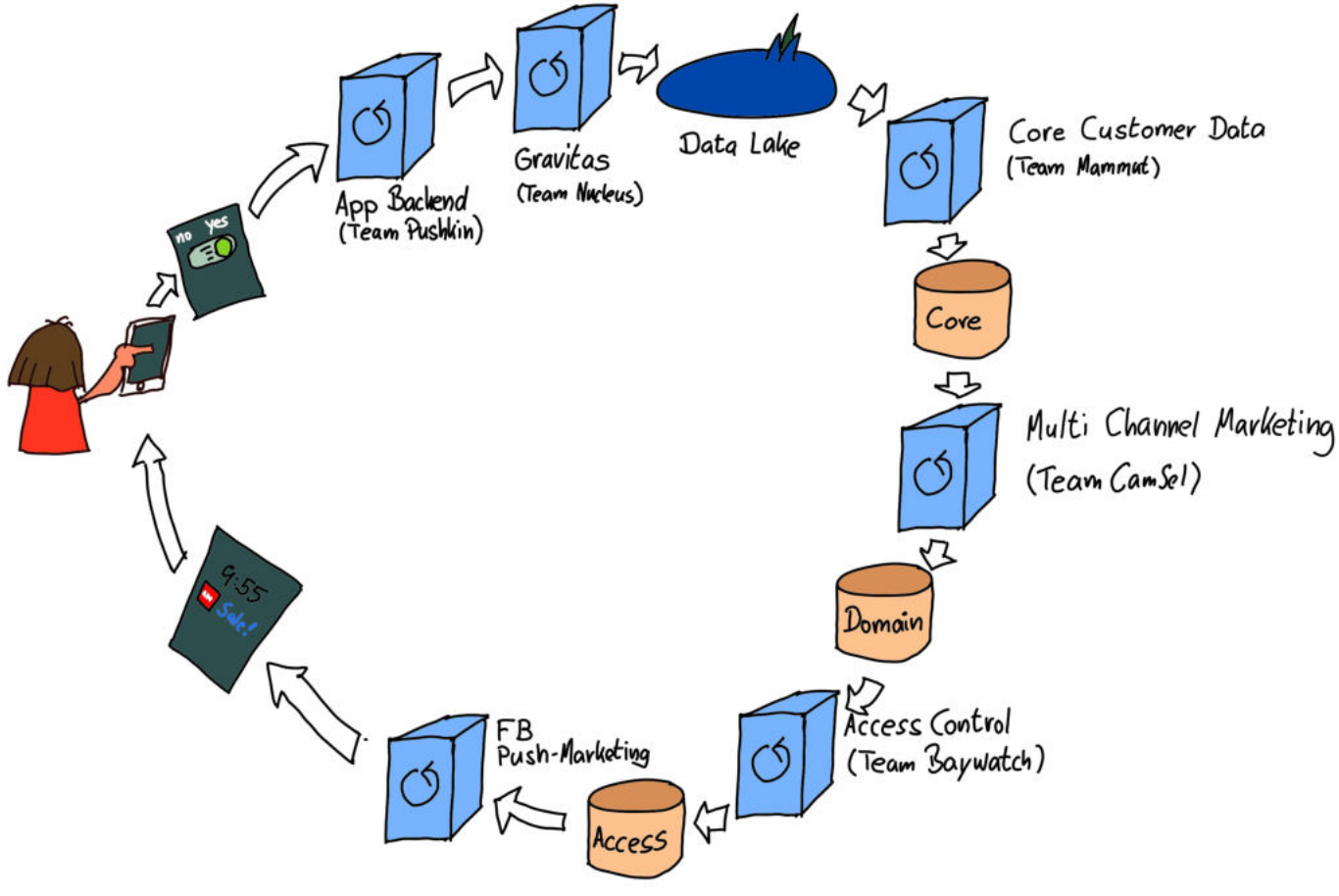

no yes
App Backend
(Team Pushkin)

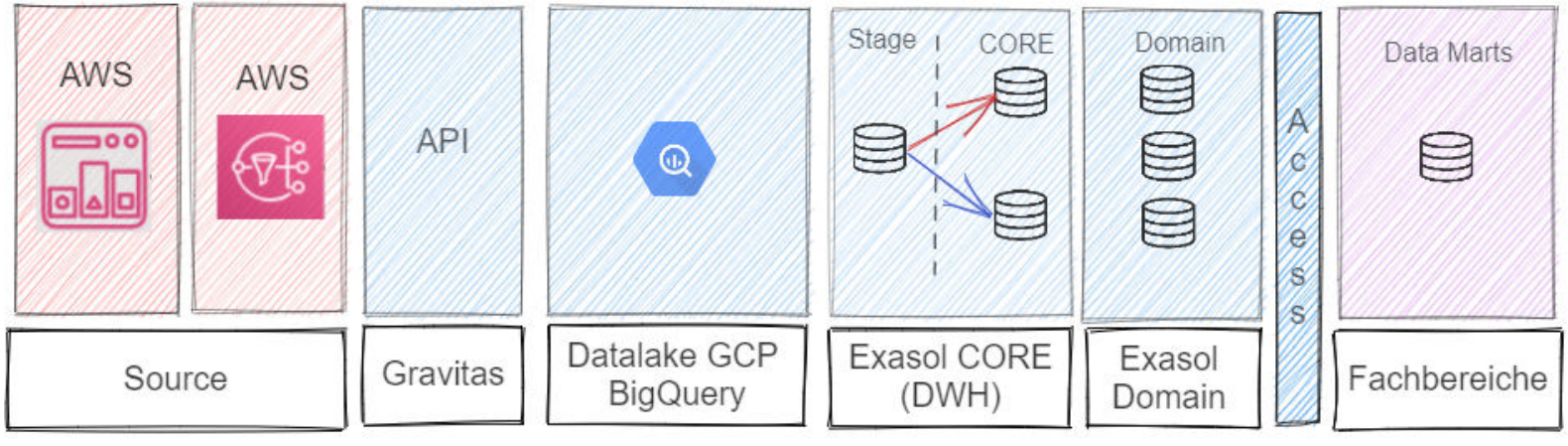

no
yes
App Backend
(Team Pushkin)
Gravitas
(Team Nucleus)
Data Lake


no yes
App Backend
(Team Pushkin)
Gravitas
(Team Nucleus)
Data Lake
Core Customer Data
(Team Mammut)
Core


no
yes
App Backend
(Team Pushkin)
Gravitas
(Team Nucleus)
Data Lake
Core Customer Data
(Team Mammut)
Core
Multi Channel Marketing
(Team CamSel)
Domain

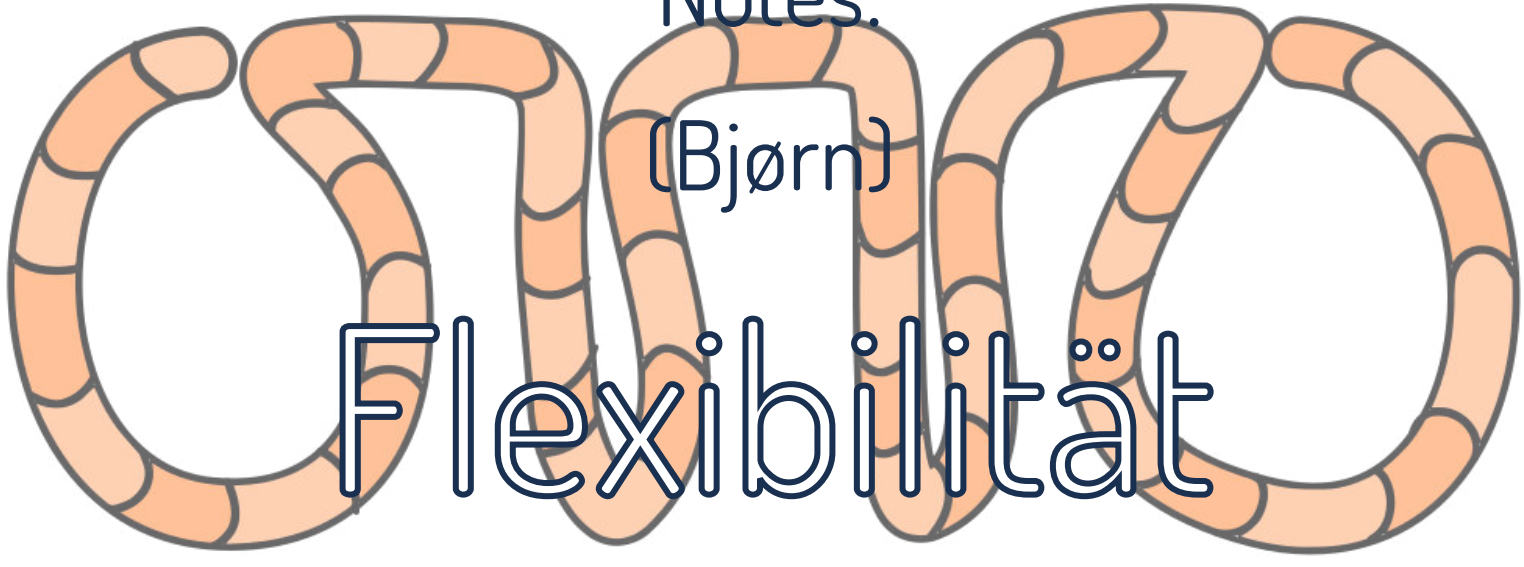

no
yes
App Backend
(Team Pushkin)
Gravitas
(Team Nucleus)
Data Lake
Core Customer Data
(Team Mammut)
Core
Multi Channel Marketing
(Team CamSel)
Domain
Access Control
(Team Baywatch)
Access


App Backend
(Team Pushkin)
Gravitas
(Team Nucleus)
Data Lake
Core Customer Data
(Team Mammut)
Core
Multi Channel Marketing
(Team CamSel)
Domain
Access Control
(Team Baywatch)
Access
FB
Push-Marketing
no
yes
9:55
Sale!


AWS
AWS
API
Stage
CORE
Domain
Access
Data Marts
Source
Gravitas
Datalake GCP
BigQuery
Exasol CORE
(DWH)
Exasol
Domain
Fachbereiche

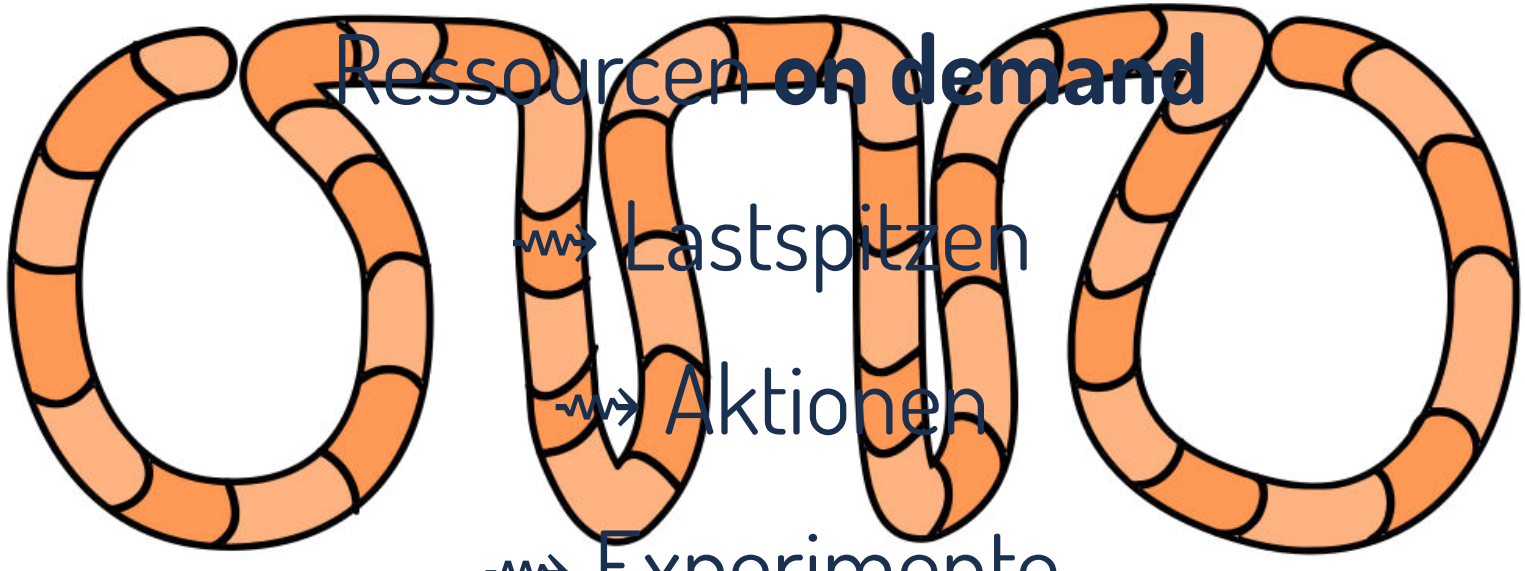
Notes:

(Bjørn)



Notes: (Bjørn)

Flexibilität, Unabhängigkeit



Ressourcen **on demand**

- Lastspitzen
- Aktionen
- Experimente
- Konkurrierende Lösungen

Empowerment

Es muss für jeden bei OTTO leichter werden, Daten zu nutzen.

Empowerment

- großes Ökosystem mit einsatzbereiten Produkten
- Managed Services ⇝ einfach
- On-Demand, Kostentransparenz