Modul 2: Datenprodukt konzipieren, Infrastruktur und Software

Angewandte Datenanalyse für die öffentliche Verwaltung in Bayern (ADA Bayern)
www.ada-oeffentliche-verwaltung.de

BERD
@NFDI

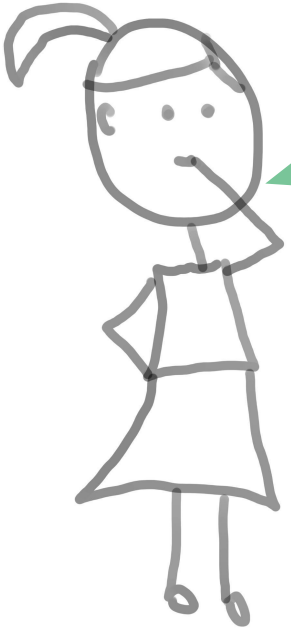
LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Bayerisches Staatsministerium
für Digitales

Die drei Workshop-Tage

1. Tag: Gemeinsame Probleme verstehen
2. **Tag: Best-Practice: Zielvorstellungen entwickeln**
3. Tag: Infrastruktur (Fokus: Einzelbaumerkennung)

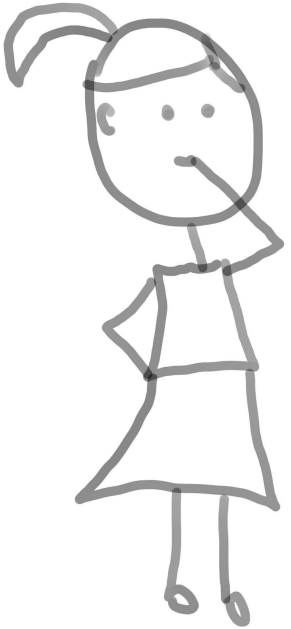
Gemeinsames Ziel festlegen



*Wie können wir **gemeinsam…***

- *effektiver Einzelbaumerkennung durchführen und nutzen,*
- *Infrastruktur nutzen (Cloud?),*
- *Daten, Modelle und Code austauschen / gemeinsam nutzen?*
- *Datentypen überdenken / standardisieren, harmonisieren in die Vergangenheit, Robustheit gegenüber z.B. Auflösung*
- *Weitere Partner identifizieren und einbeziehen -> Skaleneffekte (Bayern-weit?)*

Rückblick



Was ist euch von gestern besonders in Erinnerung geblieben?

Übersicht: Was ist die Cloud?	09:35 - 10:00
ADA Bayern und die Cloud	10:00 - 10:30
Pause	10:30 - 10:45
Recap: Arbeitsabläufe	10:45 - 10:55
Methoden zur Nutzung von Fernerkundungsdaten	10:55 - 12:00
Mittagspause	12:00 - 12:50
Herausforderungen mit der IT-Infrastruktur	12:50 - 13:25
Brainstorming: Mögliche Anwendungen & Datenprodukte (inkl. Pause)	13:25 - 14:45
Pause	14:45 - 14:55
Wünsche an die Infrastruktur	14:55 - 15:20
Wrap-Up und optional Parkspaziergang	ab 15:20

Was ist die Cloud?

- ein Cluster!
- ein Supercomputer!
- ein Datenspeicher!
- ein Superman!

- Nichts davon?
- Alles davon?

In einfachen Worten:

- Cloud = Massenhaft Speicherplatz + Rechenpower nahebei + Netzwerkbandbreite

Diese Slides sind inspiriert durch eine Vorlesung von Mohamed Farag (2024): Cloud Infrastructure: Design, Analysis and Implementation. URL: https://www.andrew.cmu.edu/user/mfarag/14848/

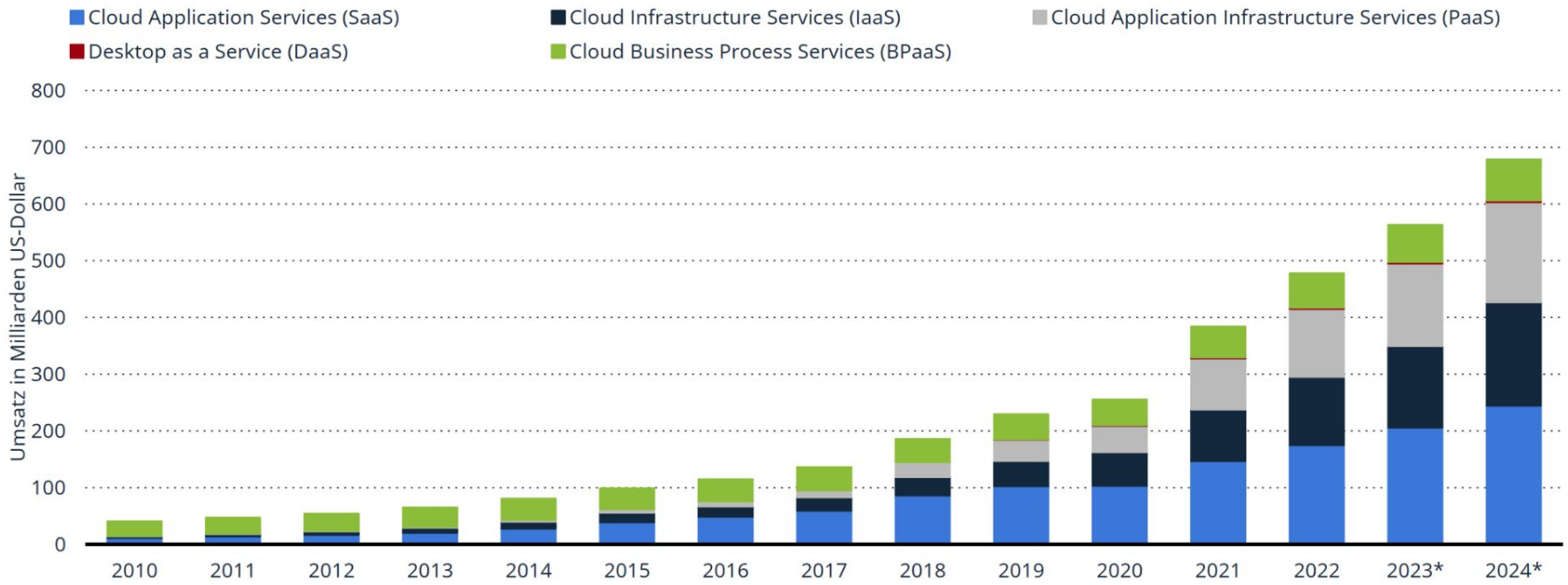
Beispielhafte Cloud Services

- Dropbox
- Google Drive
- Microsoft OneDrive
- Apple iCloud

- Netflix - hosted on AWS
- Google search - Google Cloud
- Google Docs, Sheets, and Slides
- Facebook - AWS
- Der Spiegel - Google Cloud & AWS

Umsatz mit Cloud Computing weltweit von 2010 bis 2022 und Prognose bis 2024 nach Segment (in Milliarden US-Dollar)

Prognose zum Umsatz mit Cloud Computing weltweit nach Segment bis 2024



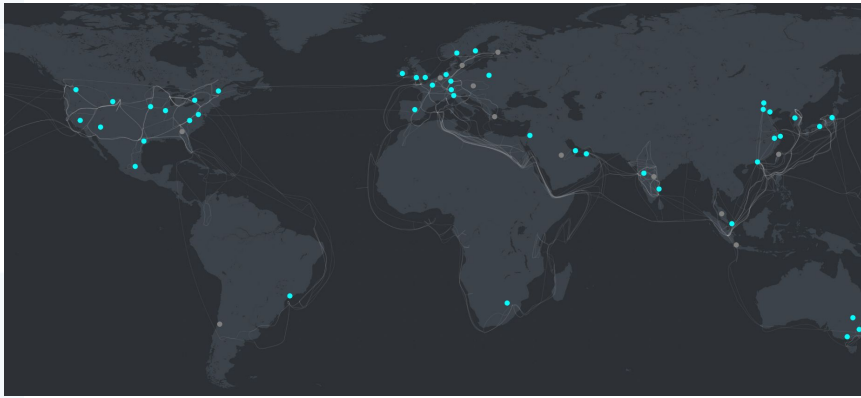
5

Beschreibung: Die Statistik veranschaulicht die weltweiten Ausgaben für Cloud Computing Services von 2010 bis 2022 und gibt eine Prognose bis 2024. Laut Quelle wurden im Jahr 2022 mit Cloud Business Process Services (BPaaS) in der Public Cloud rund 61,56 Milliarden US-Dollar umgesetzt. Mehr
Hinweis(e): Weltweit; * Prognose. Die Zahlen vor 2019 stammen aus früheren Veröffentlichungen und sind aufgrund einer möglicherweise geänderten Datenbasis eventuell nicht ohne Weiteres vergleichbar. Mehr
Quelle(n): Gartner

statista

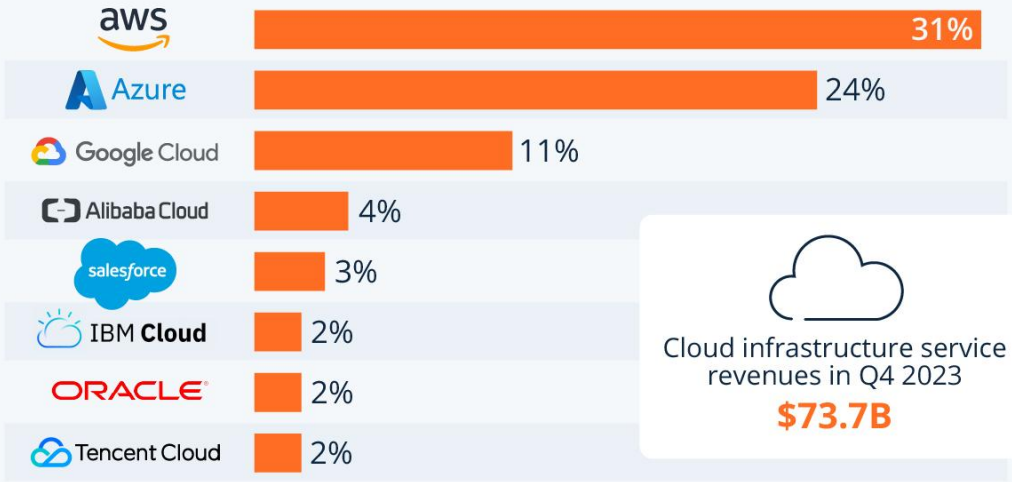
Cloud Computing

Microsoft führt Azure auf rund 300 miteinander verbundenen Rechenzentren weltweit aus



Amazon Maintains Cloud Lead as Microsoft Edges Closer

Worldwide market share of leading cloud infrastructure service providers in Q4 2023*



Cloud infrastructure service revenues in Q4 2023
$73.7B

* Includes platform as a service (PaaS) and infrastructure as a service (IaaS) as well as hosted private cloud services

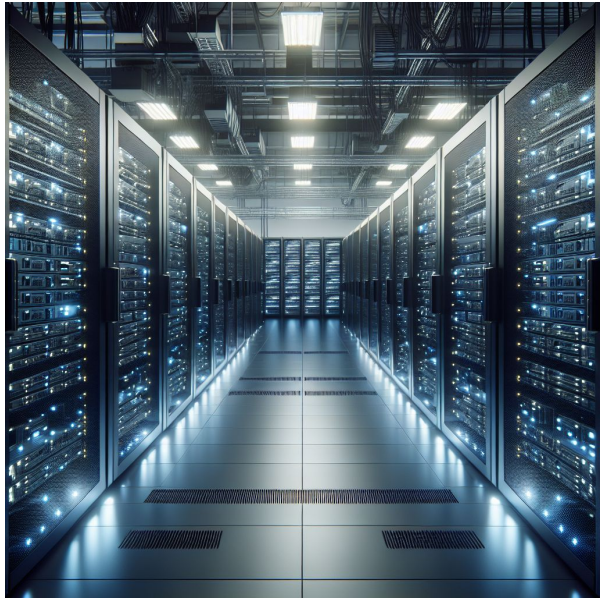
Source: Synergy Research Group

statista

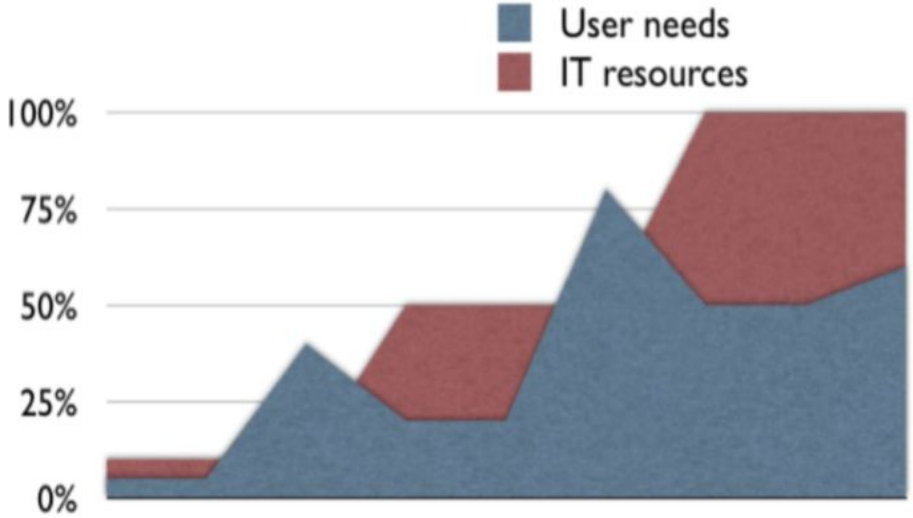
Was ist die Cloud und wofür ist sie gut?

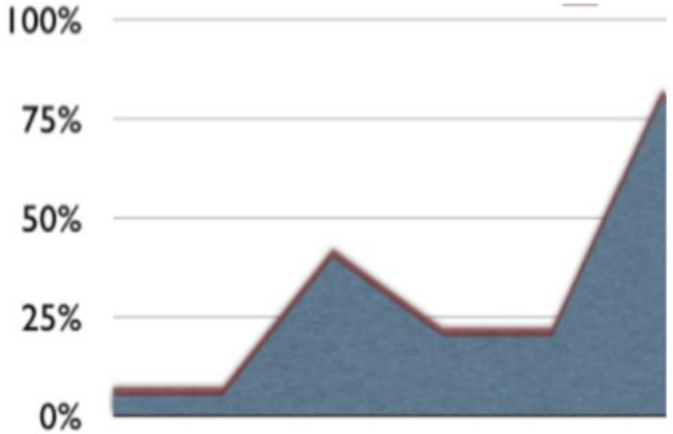
- Bis ca. 2004 mussten **Server gekauft** bzw. gemietet werden
  - **Wenig flexibel**: Was falls plötzlich mehr Rechenkapazität erforderlich ist?
  - **Aufwendige Administrierung**: Betriebssystem und Basis-Software installieren, Updates bereitstellen, …

- Die Cloud heute: **Pay-as-you-go für nötige Ressourcen**
  - Bereitgestellte Rechenkapazität lässt sich automatisch anpassen
    ⟶ Kosteneinsparungen
  - Administration erfolgt (teilweise) durch Cloudanbieter
  - Leistungssteigerung mittels redundanter Ressourcen



IT-Ressourcen vs Nutzerbedarfe einer IT-Abteilung mit und ohne Cloud





Welches Diagramm beschreibt die Infrastruktur in der Cloud? Warum?

Eine wissenschaftliche Definition der Cloud

Cloud Computing ist ein Modell mit den folgenden Charakteristika:

Peter Mell, Timothy Grance (2011): *The NIST Definition of Cloud Computing.* (PDF) National Institute of Standards and Technology (NIST)

Konventionelle Computing Infrastruktur vs Cloud Computing Infrastruktur

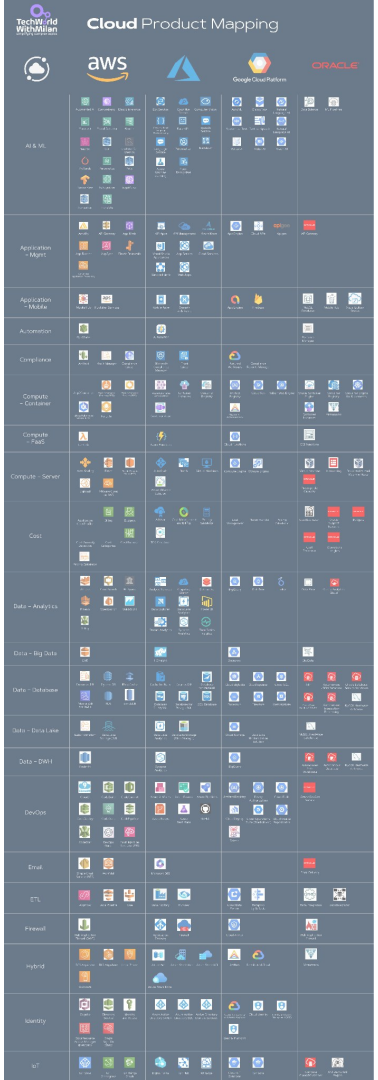
Konventionell	Cloud
Dedizierte Hardware	Gemeinsam genutzte Hardware
Feste Kapazität	Elastische Kapazität
Für Kapazität bezahlen	Bezahlung pro Nutzung
Kapital- und Betriebskosten	Betriebliche Kosten
Manuell bereitgestellt	Selbstbereitstellung
Verwaltet durch Systemadministratoren	Über APIs verwaltet

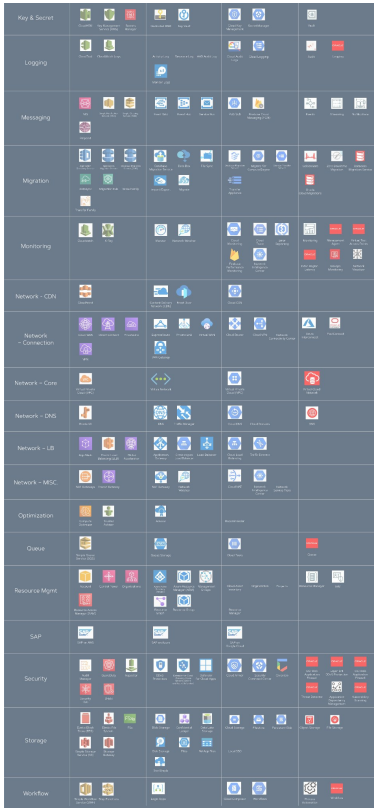
TechWorld WithMilan
simplifying complex topics

Cloud Product Mapping

	aws	Azure	Google Cloud Platform
Compute – Container	App2Container Elastic Container Registry (ECR) Elastic Container Service (ECS) Elastic Kubernetes Service (EKS) Fargate	Azure Kubernetes Service (AKS) Container Instances Container Registry Container Apps	Artifact Registry Cloud Run Kubernetes Engine Migrate for Anthos and GKE
Compute – FaaS	Lambda	Azure Functions	Cloud Functions
Compute – Server	Auto Scaling Batch Elastic Compute Cloud (EC2) Lightsail VMware Cloud on AWS	AutoScale Batch Virtual Machines Azure VMware Solution	Compute Engine VMware Engine
Storage	Elastic Block Store (EBS) Elastic File System FSx Simple Storage Service (S3) Storage Gateway	Blob Storage Confidential Ledger Data Lake Storage Disk Storage Files NetApp Files StorSimple	Cloud Storage Filestore Persistent Disk Local SSD

Quelle: https://github.com/milanm/Cloud-Product-Mapping





TechWorld WithMilan
simplifying complex topics

Cloud Product Mapping

aws

Google Cloud Platform

Data – Analytics	Athena, CloudSearch, FinSpace, Kinesis, OpenSearch, QuickSight, X-Ray	Analysis Services, Cognitive Search, Databricks, Data Explorer, Data Lake Analytics, Power BI, Stream Analytics, Synapse Analytics, Time Series Insights	BigQuery, Dataflow, Looker
Data – Big Data	EMR	HDInsight	Dataproc
Data – Database	DocumentDB, DynamoDB, ElastiCache, MemoryDB for Redis, RDS, SimpleDB	Cache for Redis, Cosmos DB, Database for MariaDB, Database for MySQL, Database for PostgreSQL, SQL Database	Cloud Bigtable, Cloud Spanner, Cloud SQL, Datastore, Firestore, Memorystore
Data – Data Lake	Lake Formation, Data Lake Storage (S3)	Data Lake Analytics, Data Lake Storage (Blob Storage)	Cloud Storage, Data Lake Modernization Solution
Data – DWH	Redshift	Synapse Analytics	BigQuery

AI & ML	Augmented AI, Comprehend, Elastic Inference, Forecast, Fraud Detector, Kendra, Neuron, Lex, Lookout for Metrics, PyTorch, Personalize, Polly, TensorFlow, Rekognition, SageMaker	Bot Service, Cognitive Search, Computer Vision, Conversational Language Understanding, Face API, Speech Service, Language Service, Personaliser, Translator, Azure Machine Learning, Form Recognizer	AutoML, DialogFlow, Natural Language AI, Speech-to-Text, Text-to-Speech, Natural Language AI, Vertex AI, Video AI, Vision AI
Cost	Application Cost Profiler, Billing, Budgets, Cost Anomaly Detection, Cost Categories, Cost Explorer, Pricing Calculator	Advisor, Cost Management and Billing, Pricing Calculator, TCO Calculator	Cost Management, Recommender, Pricing Calculator
Network – Connection	Cloud WAN, Direct Connect, PrivateLink, VPN	ExpressRoute, Private Link, Virtual WAN, VPN Gateway	Cloud Router, Cloud VPN, Network Connectivity Center
Network – Core	Virtual Private Cloud (VPC)	Virtual Network	Virtual Private Cloud (VPC)

Quelle: https://github.com/milanm/Cloud-Product-Mapping

Beispielhafte Preise für Speicherplatz

Azure Blob Storage in DE:

- Datenvolumen: €0.00162 - €0.17521/GB/Monat (abhängig von Archivierungsgrad/Zugriffshäufigkeit)
  - Zum Speichern von 1 Terabyte: €1 - 175€ pro Monat
- Datenzugriffe:
  - Schreiben: €0.02 - €0.06 pro 10 000 Schreibzugriffe (teurer für archivierte Daten)
  - Lesen: €0.002 - €0.005 pro 10 000 Lesezugriffe (teurer für archivierte Daten)
- Datenübermittlung über das Internet (egress mittels NAT Gateway)
  - 30€ / Monat und
  - 40€ / Terabyte

Rabatte bei mehrjährigen Sparplänen verfügbar

Beispielhafte Preise für Virtual Machines

Ausgewählte Azure Virtual Machines in DE mit Linux-Betriebssystem:

Größe	Typ	Name	vCPUs	RAM	€ / Stunde	€ / Monat
Winzig	General Purpose	B1ls	1	0,5GiB	€0,0047	€4
Klein	Computeoptimiert	F4s v2	4	8GiB	€0,1518	€111
Sehr Groß	General Purpose	DC48s v3	48	384GiB	€2,4798	€1 810
Riesig	Computeoptimiert	FX48mds	48	1008GiB	€4,0108	€2 927
1x V100 GPU	GPU	NC6s v3	6	112GiB	€2,7493	€2 007
8x H100 GPU	GPU	ND96isr H100 v5	96	1900GiB	€88,3378	€64 486

Sekundengenaue Abrechnung!

Enorme Rabatte möglich bei 1-3 jährigen Sparplänen oder bei flexiblem Nutzungszeitpunkt.

Everything-as-a-Service. Was bietet die Cloud?

Die Cloud bietet verschiedene Services. Aber was heißt was genau?

Alles, was sonst ein Produkt war:

- Software
- Compute
- Speicherplatz
- Andere Peripheriegeräte (Scanner, Drucker, Microcontroller, Sensoren …)
- Plattformen, z.B. APIs und dazugehörige Systeme

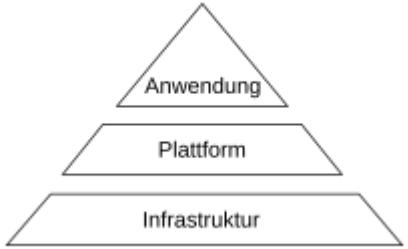
SaaS: Software-as-a-Service

- Bietet **Anwendungssoftware** in der Cloud
  - On-demand Software
  - Viele Anwendungen laufen direkt im Webbrowser

- SaaS ist größter Cloudmarkt

- Beispiele: Google Apps, Microsoft Office 365, Oracle’s Netsuite, SAP’s Concur, Cisco WebEx, GoTo Meeting

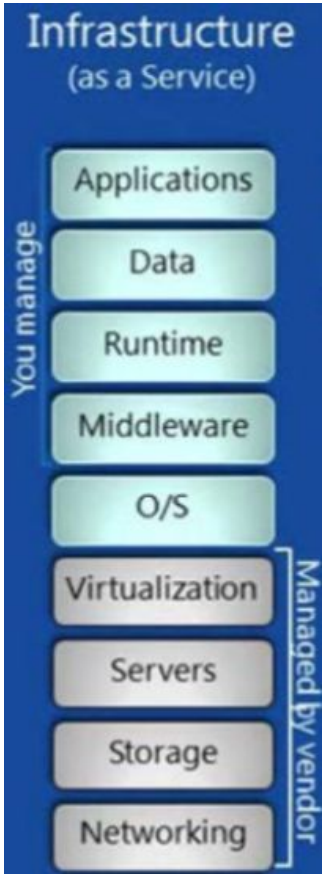
PaaS: Platform as a Service



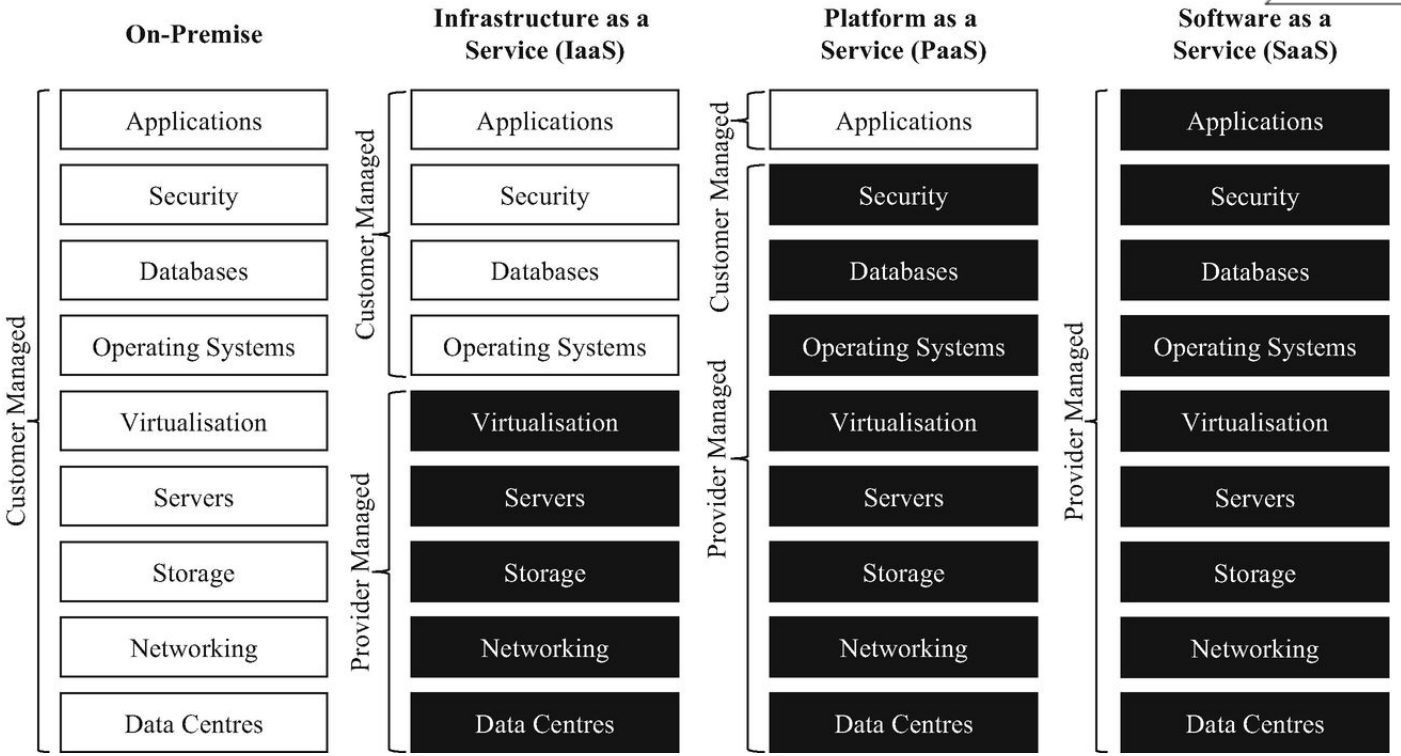
- Ziel: **Einfache Erstellung von Cloud-Anwendungen** für EntwicklerInnen
- Bietet **Computing-Plattformen**, die typischerweise Betriebssystem, Programmiersprache, Ausführungs-umgebung, Datenbank, Webserver usw. umfassen
- Anwendungen, die PaaS verwenden, erben Cloud-Merkmale wie Skalierbarkeit, hohe Verfügbarkeit, Mehrinstanzfähigkeit, SaaS-Ermöglichung und mehr.
- Beispiele:
  - für Webanwendungen: Salesforce Platform, Heroku, Google App Engine, AWS Elastic Beanstalk, Vercel, Cloudflare Workers, …
  - für Datenanwendungen: GCP Dataproc, Microsoft Fabric, Databricks, Google Earth Engine Terrabyte (by LRZ and DLR), …

IaaS: Infrastructure as a Service

- Bietet Speicher- und Computerressourcen für Entwickler
  - komplett flexibel programmierbar
- Beispiele:
  - Compute: Azure Virtual Machines Amazon EC2, GCP Compute Engine, VMWare vCloud
  - Speicher: Azure Blob Storage, Amazon S3, …
  - Network: Azure Virtual Network, Amazon Virtual Private Network, …



Cloud Computing Service Modelle



Quelle: Rosati, P., Lynn, T. (2020). Measuring the Business Value of Infrastructure Migration to the Cloud. In: Lynn, T., Mooney, J., Rosati, P., Fox, G. (eds) Measuring the Business Value of Cloud Computing. Palgrave Studies in Digital Business & Enabling Technologies. Palgrave Macmillan, Cham.

PaaS or IaaS?

Vorteile von PaaS

Kosteneffizienz

Schnellere Markteinführungszeit

Fokus auf Entwicklung

Einfache Integration mit anderen Services

Herausforderungen von PaaS

Mögliches Vendor Lock-In

Begrenzte Kontrolle über Infrastruktur

Plattform schränkt ggf. Softwarefeatures ein

Sicherheit

Vendor Lock-In: die Möglichkeit, „was Sie verwalten“ in eine andere Cloud-Umgebung umzuziehen. Ein Anbieterwechsel kann im PaaS-Model schwierig sein, wenn …

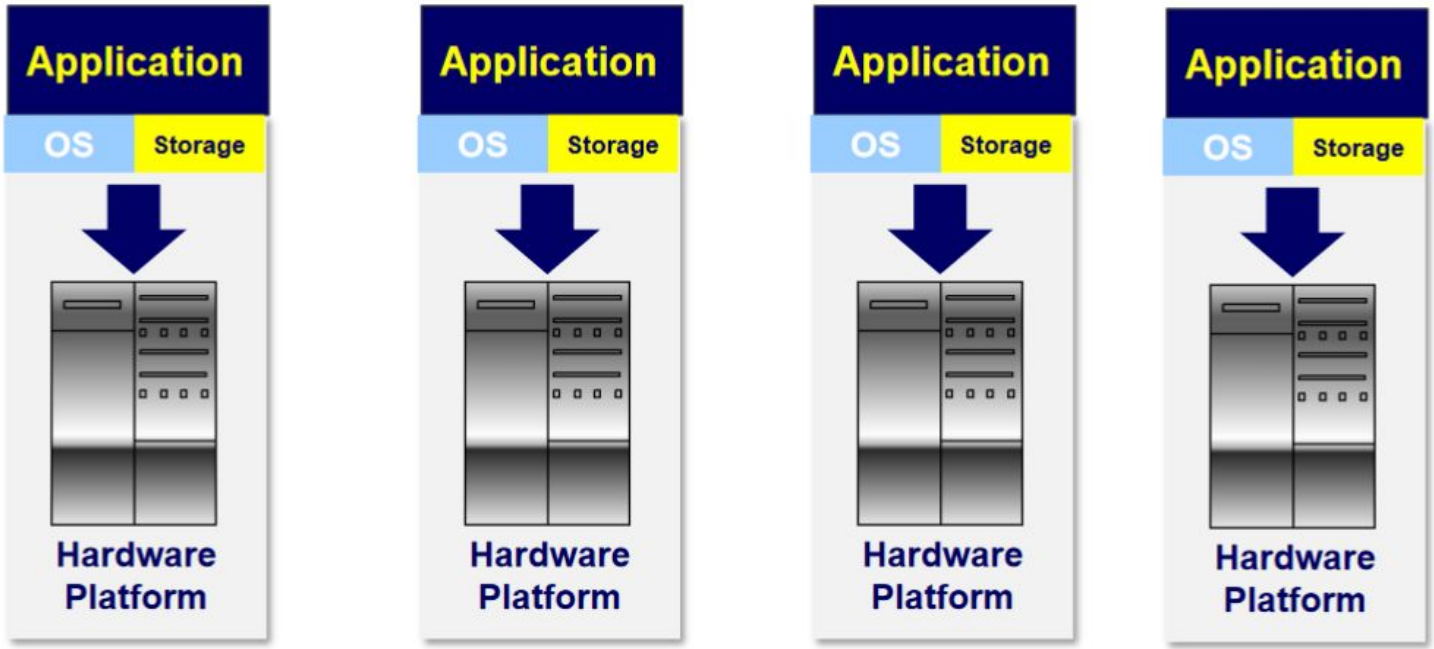
- … PaaS von Entwicklern verlangt, Apps auf der Grundlage ihrer spezifischen APIs zu entwickeln.
- … Entwickler plattformabhängige Entwicklungstools nutzen um ihre Entwicklungszeit zu beschleunigen

Schlüsseltechnologie: Virtualisierung ermöglicht Cloud Computing

- Virtualisierung, *def.*: Der Akt, eine virtuelle (anstelle einer physischen) Instanz von etwas zu erstellen
  - Virtuelle Maschinen verhalten sich so wie physische Computer, sind es aber nicht

- Virtualisierung ermöglicht die gemeinsame Nutzung von Ressourcen
  - Dadurch werden die bereitgestellten Services entkoppelt von der Hardware und den für die Hardware zuständigen IT-Technikern
  - Bessere Ressourcenausnutzung: 60-80% mit Virtualisierung verglichen mit 5-15% davor

Traditionelle Serverinfrastruktur



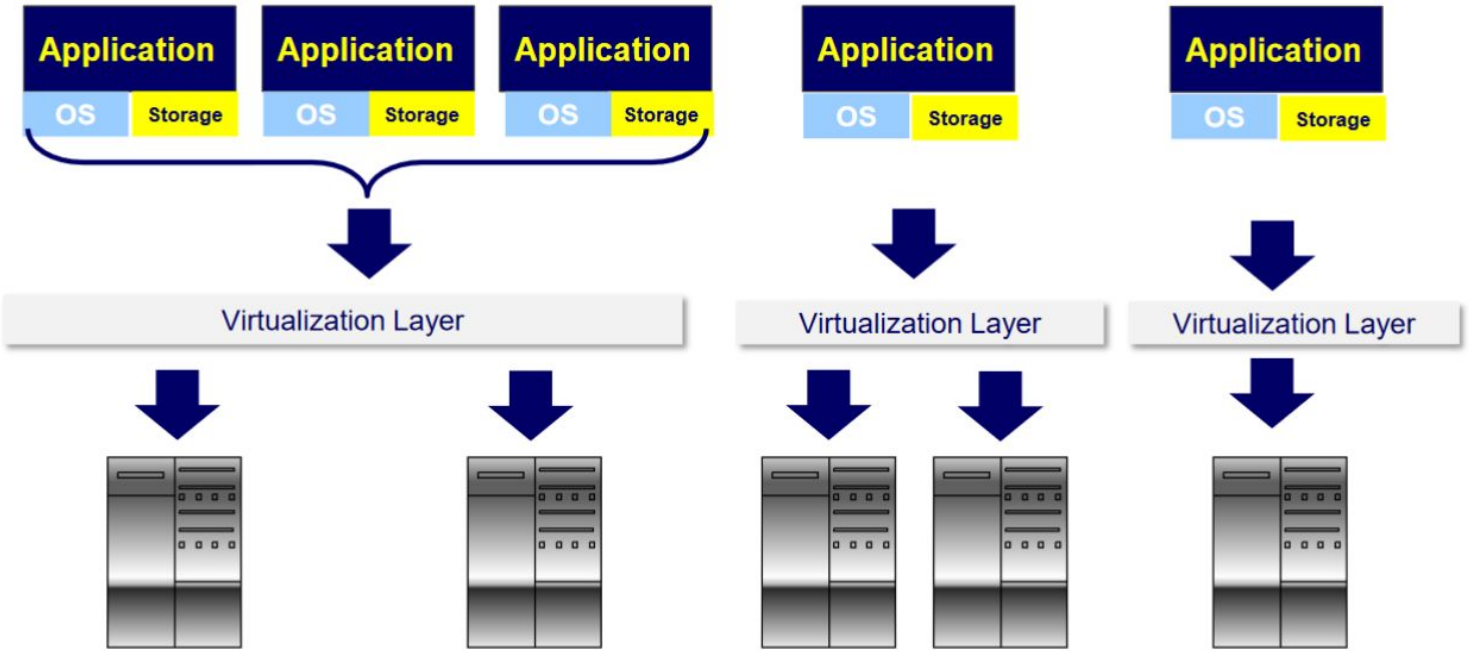
Email Exchange Server

Anwendungs-Server 1

Anwendungs-Server 2

Datenbank Server

Virtual Server Infrastruktur



Ein virtueller Server kann von einem oder mehreren Hosts bedient werden und ein Host kann mehr als einen virtuellen Server beherbergen.

Wichtige Virtualisierungstechnologien

- **Virtual machine (VM)**, *def.*: Eine virtuelle Maschine ist ein Computersystem, das mithilfe von Software auf einem physischen Computer erstellt wird, um die Funktionalität eines anderen separaten physischen Computers nachzuahmen.
  - Jede VM führt ihr eigenes Betriebssystem aus
- **Container**, *def.*: Ein Container ist ein portables Softwarepaket, das alle für die Ausführung erforderlichen Ressourcen enthält. Eigenschaften:
  - Isolation: Prozesse des Containers A stören nicht die Prozesse des Containers B
  - Replizierbarkeit: Derselbe Prozess aus demselben Container-Image sollte auf jedem Host-Rechner/Betriebssystem/Konfiguration gleich ausgeführt werden.
  - Ein Container verhält sich wie eine eigene isolierte Maschine, teilt aber sein OS mit einer Host-Maschine

Milestones:

- 2006: AWS EC2 startet Virtuelle Maschinen als Service zu vermieten
- 2014: Docker popularisiert Container

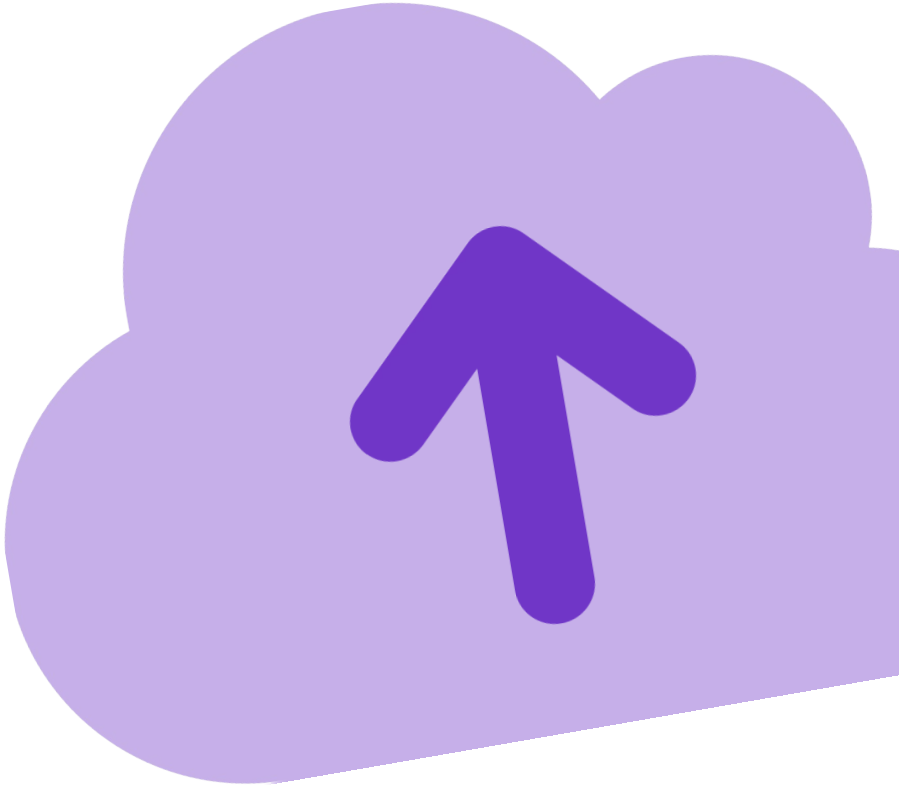
Zentrale Herausforderungen des Cloud Computing

- Isolieren und gleichzeitig teilen
  - Datenschutz muss gewährleistet sein
    - Daten, Methoden, Organisation usw.
  - Starke Ressourcenverwendung durch Anwendung A darf nicht zu Leistungseinbußen bei Anwendung B führen
  - Dies muss auf dynamische, aber vertrauenswürdige Weise erfolgen

- Elastische Bereitstellung
  - Sekundengenauer Abrechnung der verbrauchten Ressourcen
  - Ermöglicht Energieeinsparungen usw.
  - Erfordert Virtualisierung aller Aspekte, z. B. Computing, Speicher, Netzwerk usw.

Pause

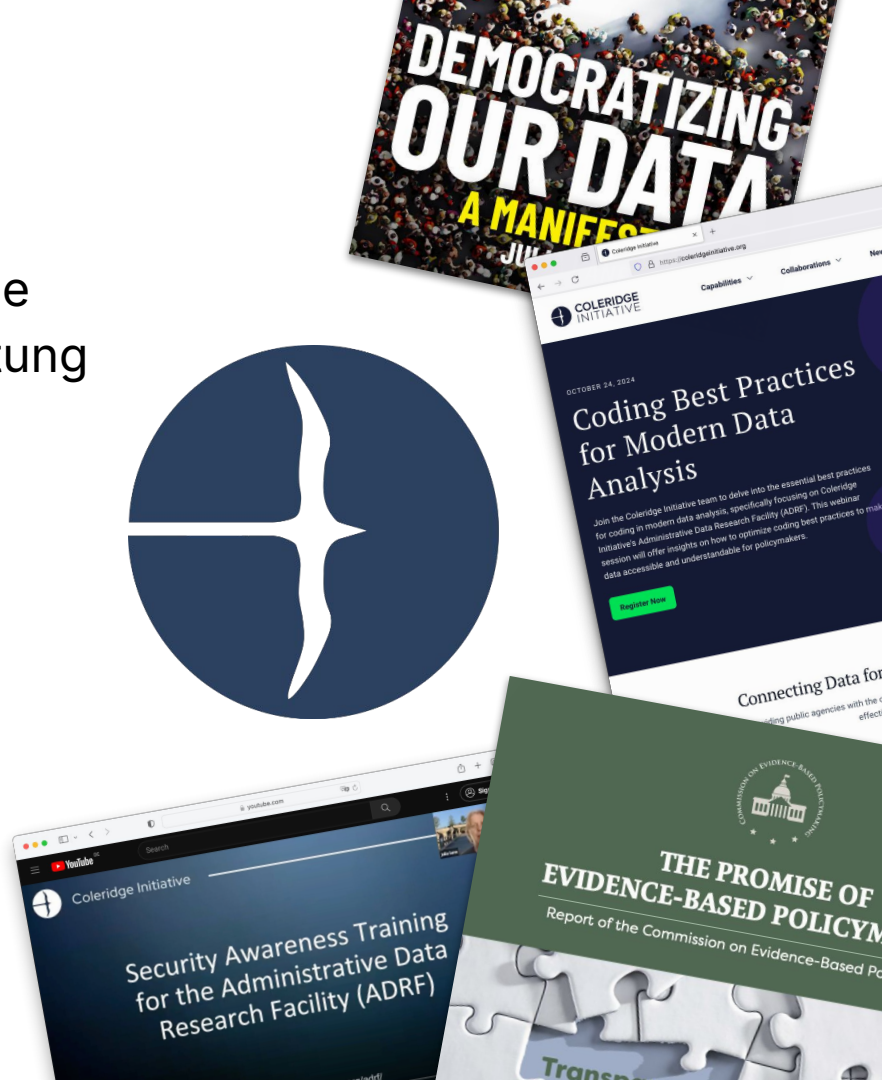
ADA Bayern in der Cloud



Coleridge Initiative & ADRF

Die Coleridge Initiative bemüht sich um die Verbreitung von evidenzbasierter Verwaltung

- Vermittelt Zugang zu **mehr als 400 vertraulichen, fein aufgelösten Datensätzen** von 50+ US-Institutionen



Rahmenkonzept: **Five Safes**

Coleridge in Bayern

Infrastruktur

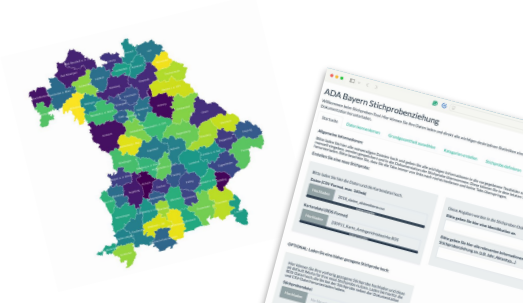
Bereitstellung und Provisionierung von **Speicher- und Rechenkapazität**, vor Ort und in der “Cloud”

Platform

Geschützter Zugang zu Daten sowie Kollaborations- und Analyseumgebungen

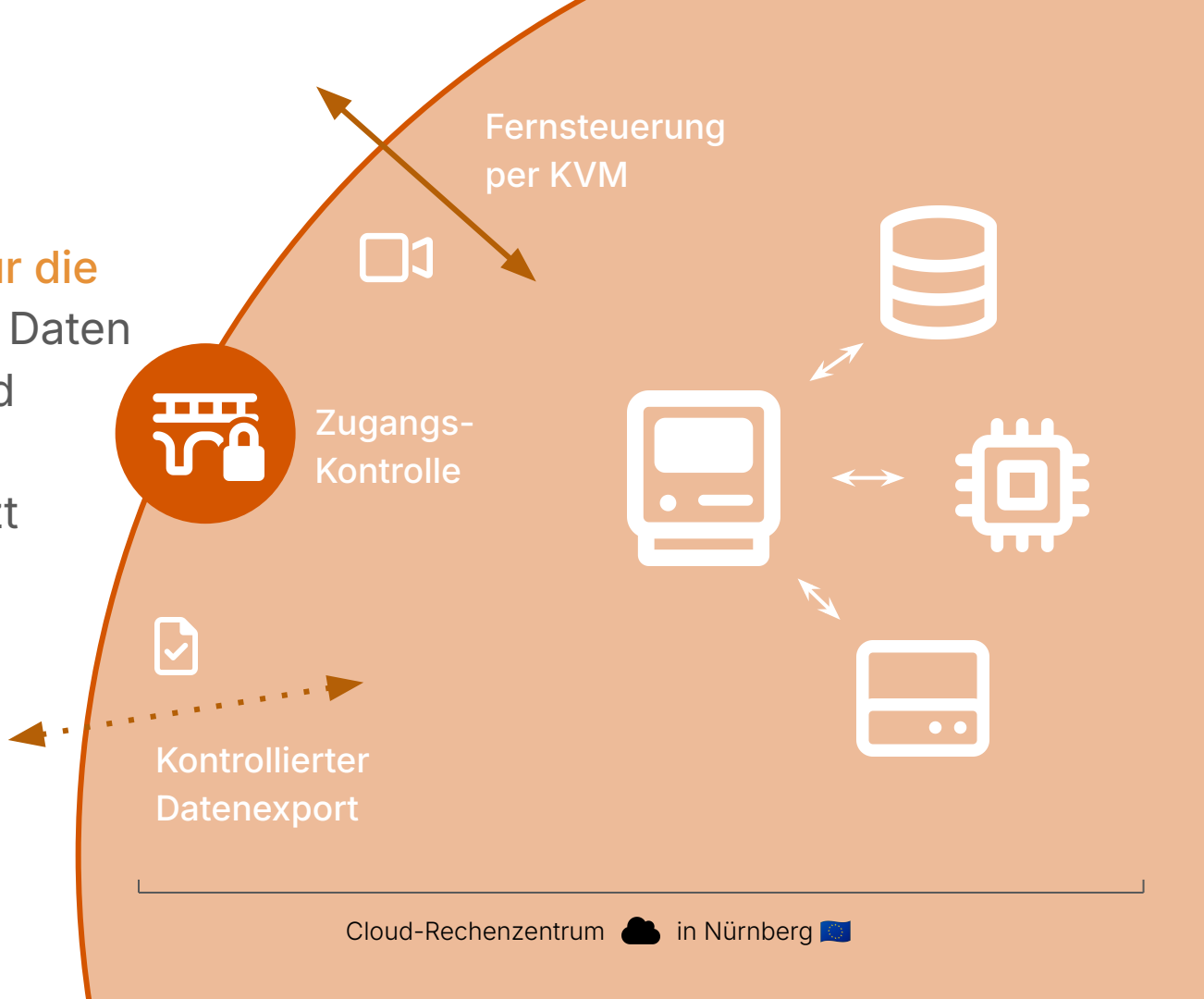
Software

Fertige **Tools** spezifisch für konkrete Anwendungen

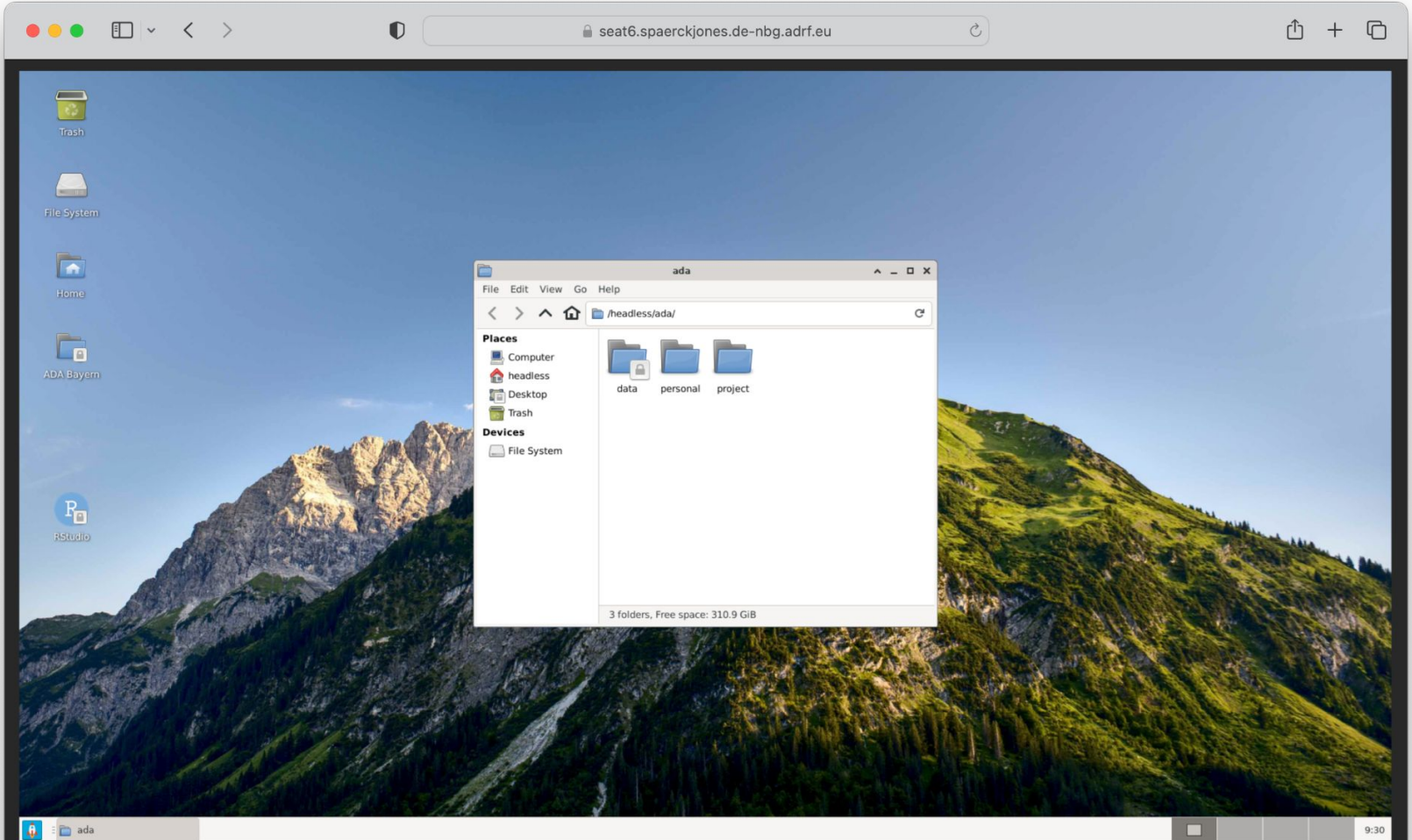


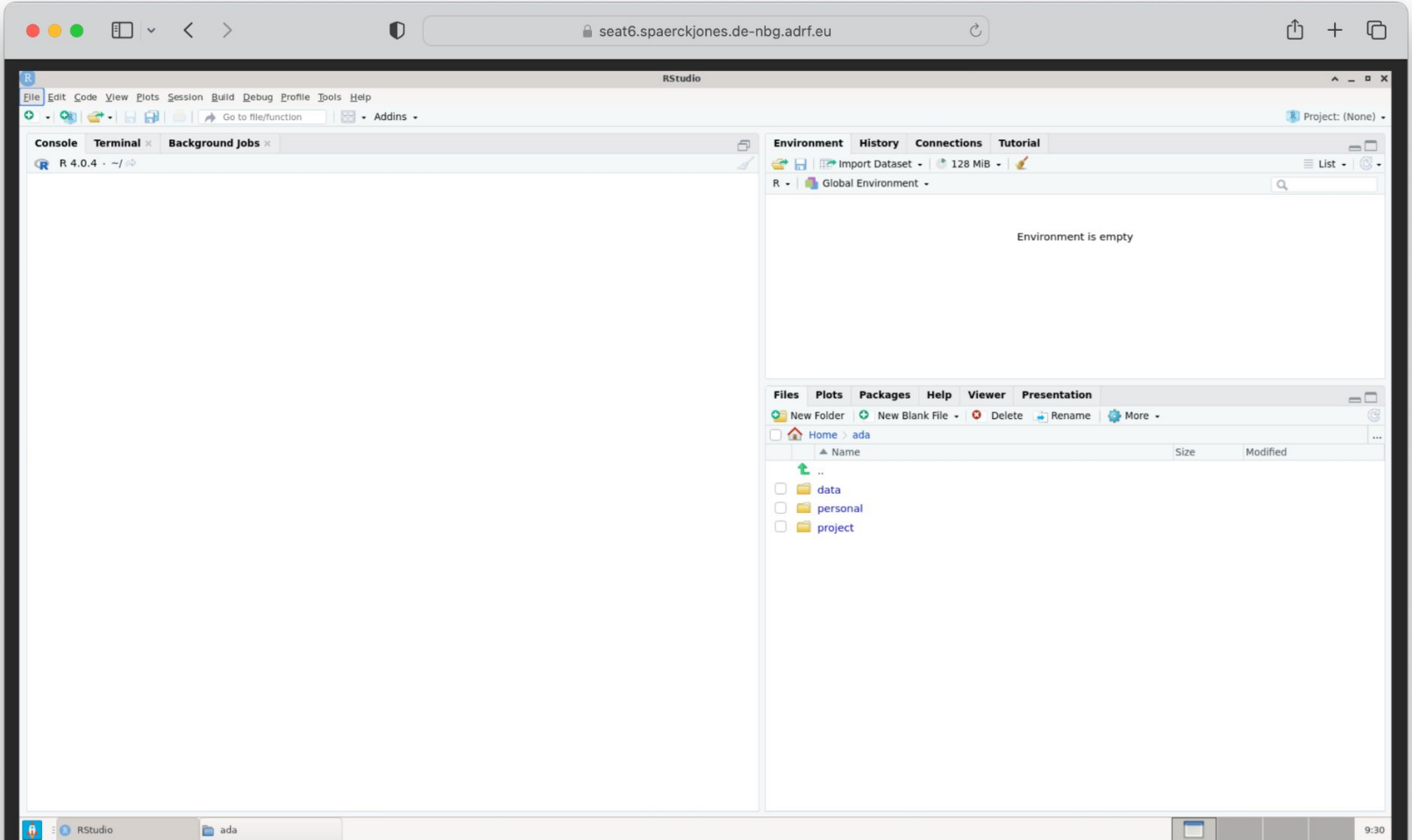
Safe settings

Geschützter Rahmen für die Datenverarbeitung, der Daten vor äußerem Zugriff und unintendierter Veröffentlichung schützt



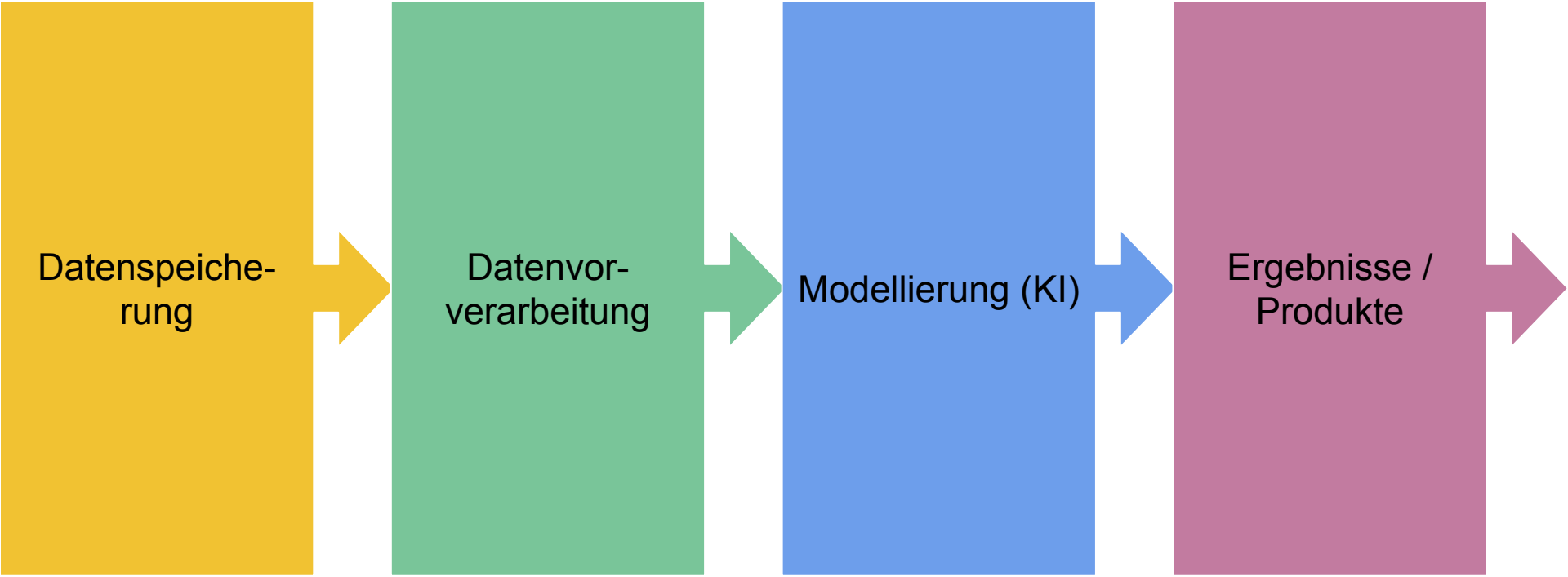

seat6.spaerckjones.de-nbg.adrf.eu
Trash
File System
Home
ADA Bayern
RStudio
Virtuelle Analyseumgebung
Die **Cloudplattform** können Sie nutzen wie Ihren eigenen PC
9:29


seat6.spaerckjones.de-nbg.adrf.eu
Trash
File System
Home
ADA Bayern
RStudio
ada
File Edit View Go Help
/headless/ada/
Places
Computer
headless
Desktop
Trash
Devices
File System
data
personal
project
3 folders, Free space: 310.9 GiB
ada
9:30


seat6.spaerckjones.de-nbg.adrf.eu
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console
Terminal
Background Jobs
R 4.0.4 · ~/
Environment
History
Connections
Tutorial
Import Dataset
128 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
New Folder
New Blank File
Delete
Rename
More
Home
ada
Name
Size
Modified
..
data
personal
project
RStudio
ada
9:30

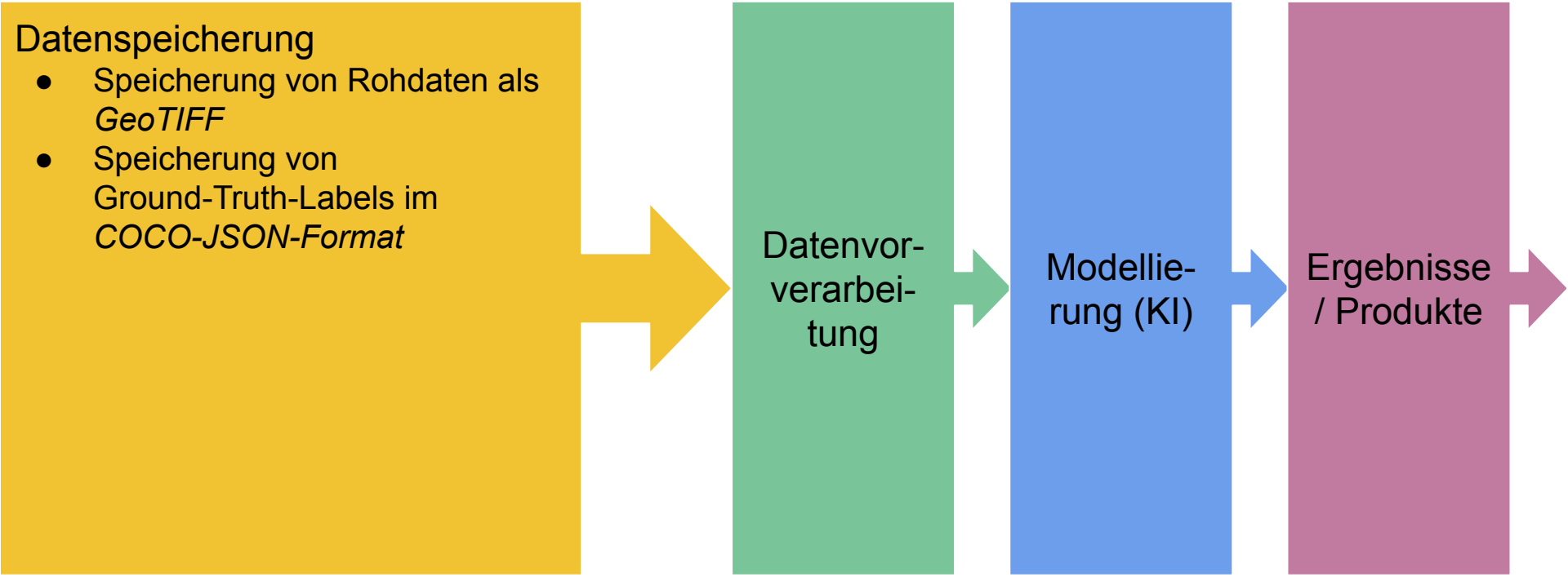
Arbeitsschritte

Welche Infrastruktur brauchen wir hierfür?



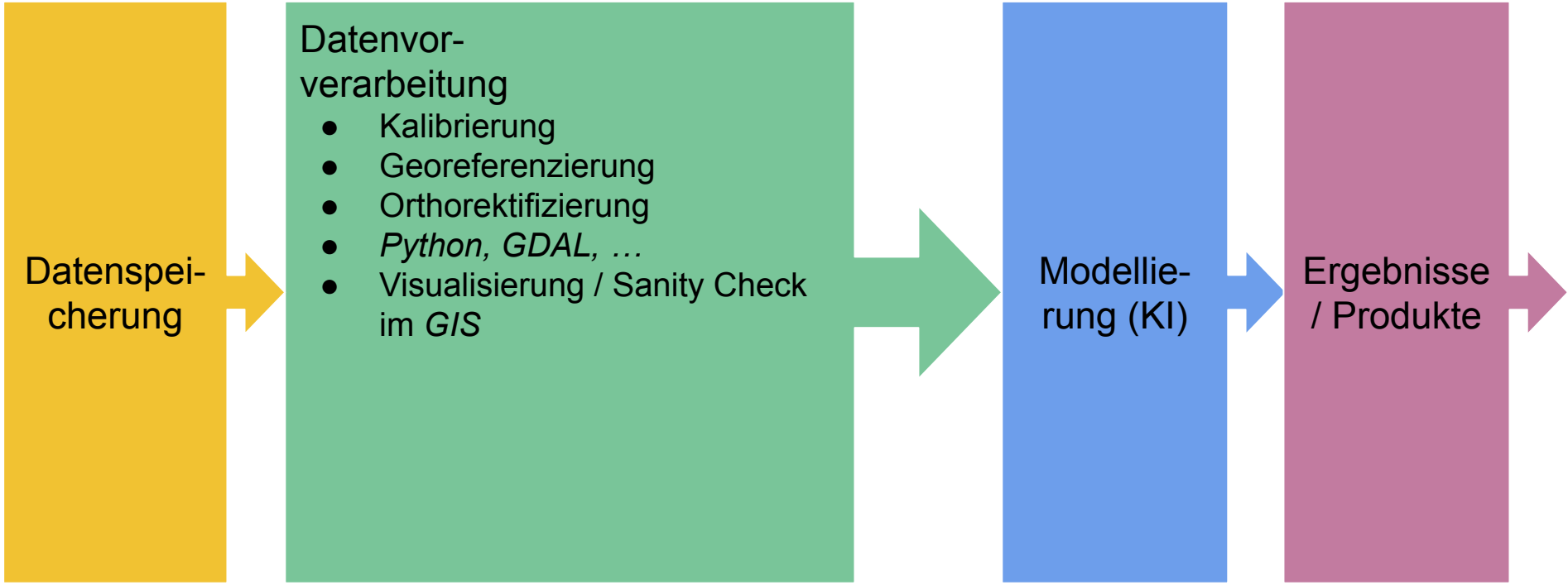
Arbeitsschritte

Welche Infrastruktur brauchen wir hierfür?



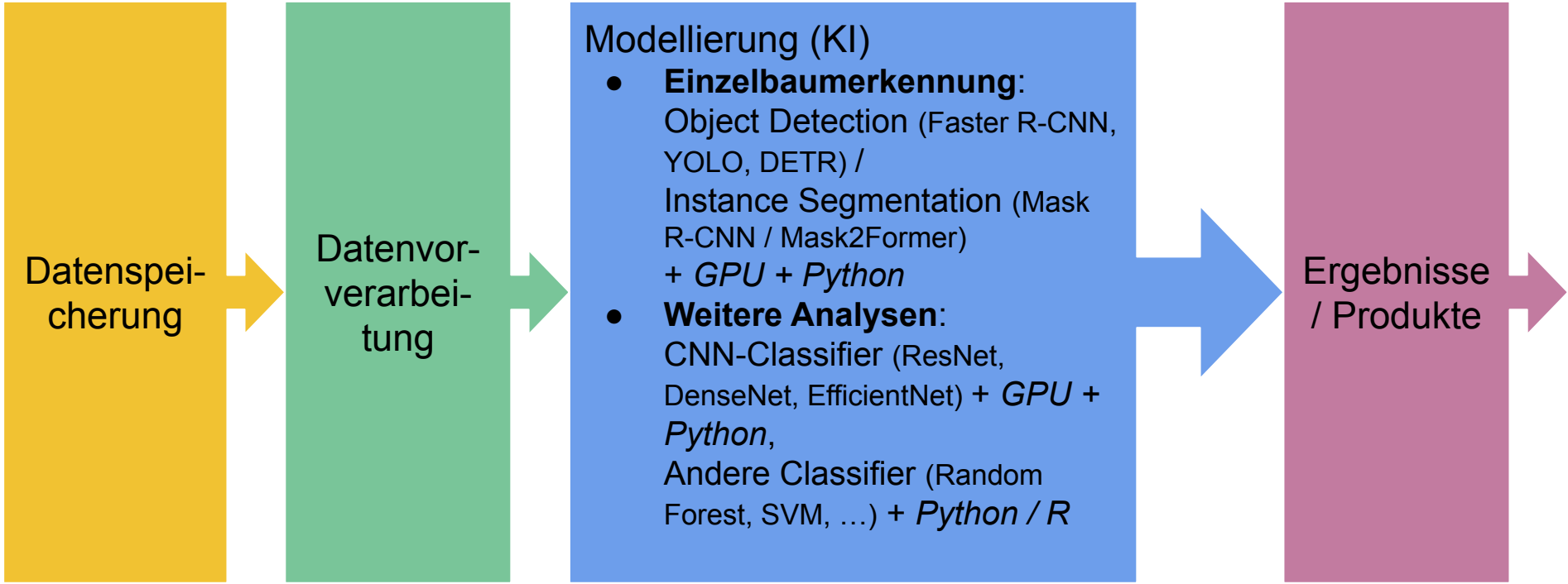
Arbeitsschritte

Welche Infrastruktur brauchen wir hierfür?



Arbeitsschritte

Welche Infrastruktur brauchen wir hierfür?



Arbeitsschritte

Welche Infrastruktur brauchen wir hierfür?

Datenspei-
cherung

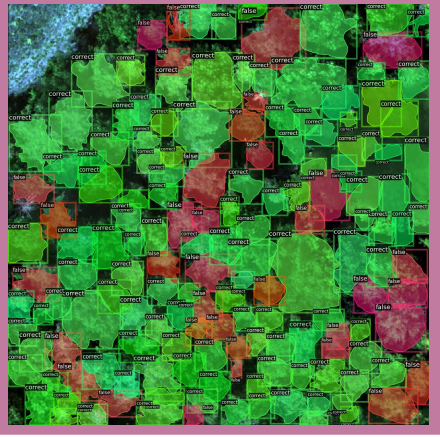
Datenvor-
verarbei-
tung

Modellie-
rung (KI)

Ergebnisse / Produkte

- Binäre Maske pro Baum
- Weitere Labels/Vorhersagen

→Optimalerweise visualisierbar als *GIS-Layer*



Pause

Mittagspause

Aktuelle Herausforderungen mit der IT-Infrastruktur

1-2-4-Alle

Was sind aktuelle Herausforderungen bei der Einzelbaumerkennung?
Zum Beispiel mit der IT-Infrastruktur?

1 - Jede/r macht sich alleine Gedanken zur Fragestellung. (1 Minute)

2 - Teilt eure Gedanken in Paaren und entwickelt sie weiter. (2 Minuten)

4 - Jeweils zwei Paare bilden eine Vierergruppe und entwickeln die Gedanken weiter. Einigt euch auf die wichtigsten Aspekte. (4 Minuten)

Alle - Jede Vierergruppe teilt die Ergebnisse im Plenum. (5 Minuten)

Mögliche Anwendungen & Datenprodukte

Mögliche Anwendungen & Datenprodukte

Brainstorming

Welche Endprodukte der Einzelbaumerkennung wären für eure Nutzenden (Förster:innen, Stadtplaner:innen, …) besonders nützlich?

1. Besprich mit deinem/n Nachbarn mögliche Ideen (2-3 Personen).
2. Schreibt eure 3 Favoriten auf die 3 Zettel.
3. Hängt die Zettel auf und diskutiert die Vorschläge der anderen.
4. Kommen noch weitere Ideen auf, fügt gerne weitere Zettel hinzu.

Mögliche Anwendungen & Datenprodukte

Gruppenarbeit

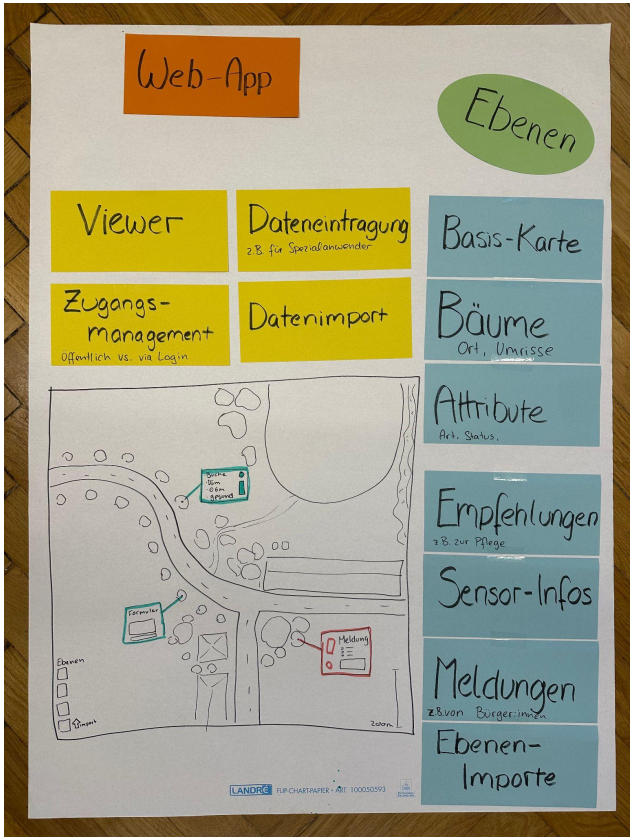
5-6 Personen pro Gruppe
Zeit bis 14:45

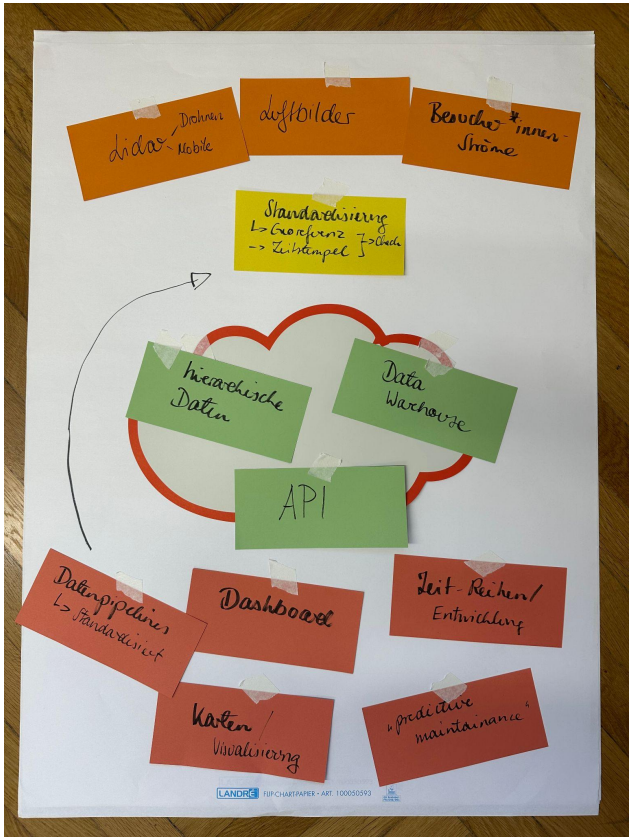
Cluster aus den Zetteln = Gruppe

1. Konsolidiert die Ideen zu eurem Top-Favoriten
2. Erstellt ein Poster für eure Idee

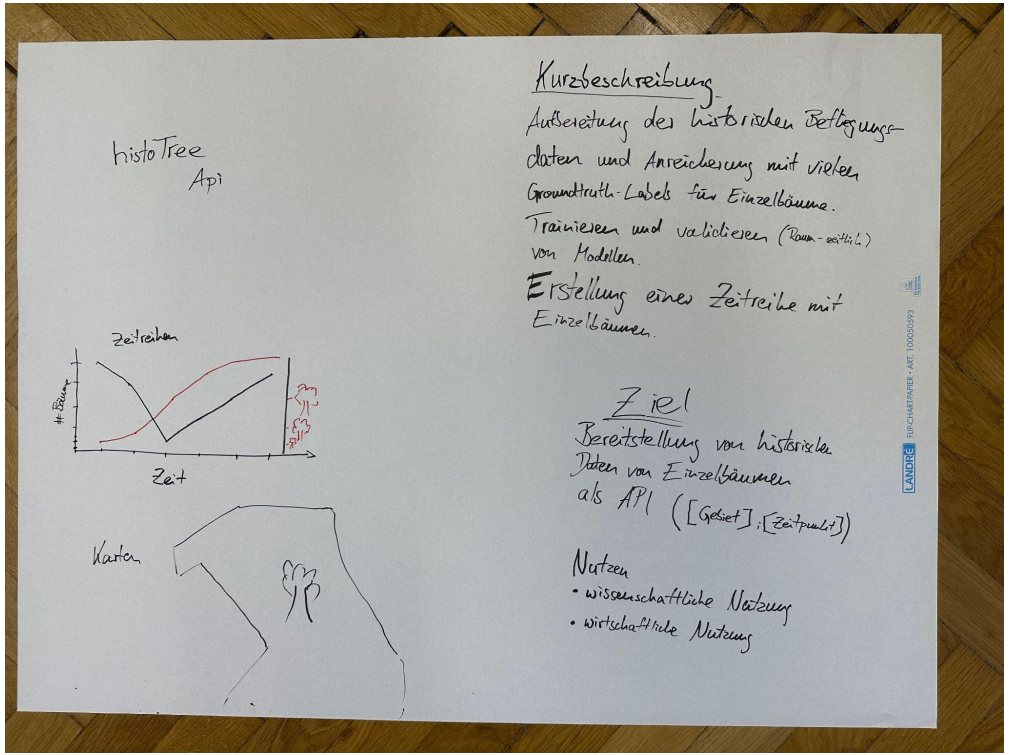
Produktname	Kurzbeschreibung
Bild	Ziel & Nutzen

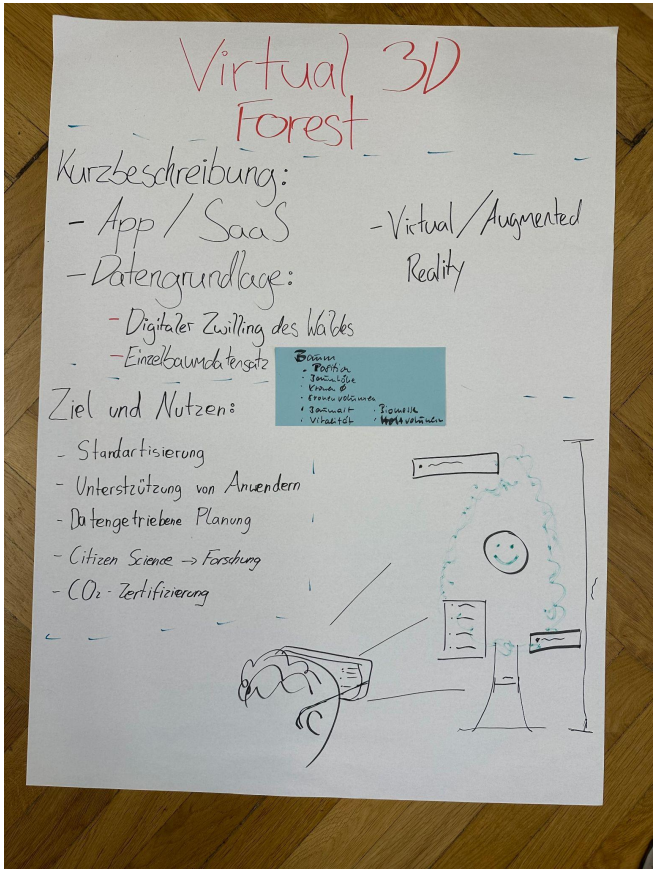
Vorstellung der Ergebnisse





Vorstellung der Ergebnisse



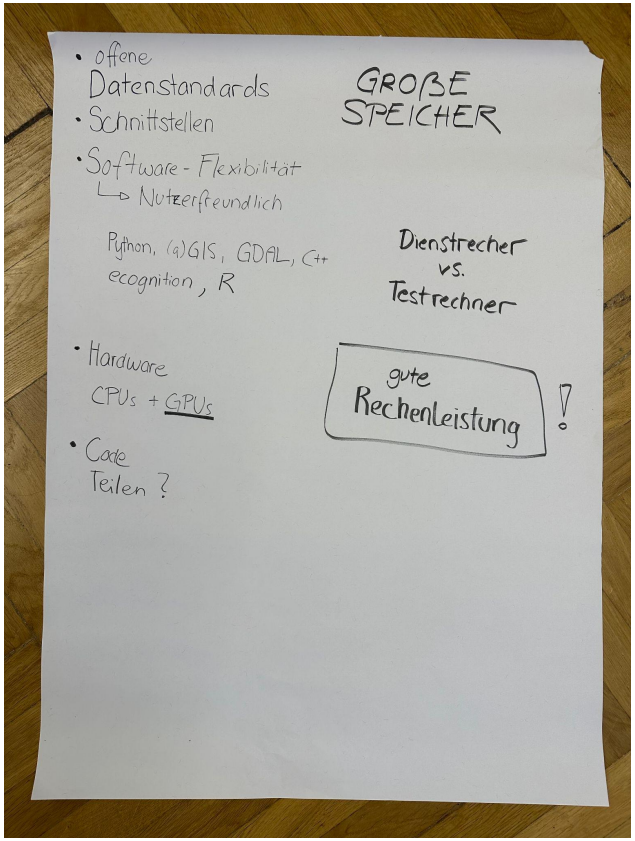


Pause

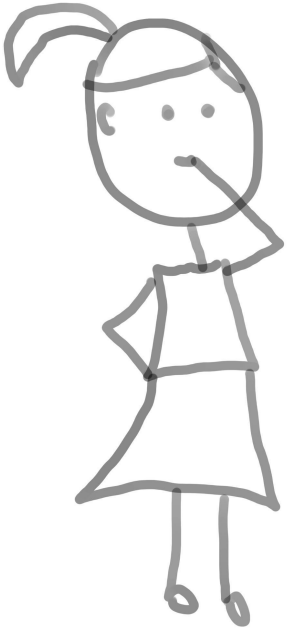
Wünsche an die IT-Infrastruktur

Wünsche an die IT-Infrastruktur

Vorstellung der Ergebnisse



Abschluss



Was war heute besonders ”merk-würdig”?

Welche Fragen gibt es?

Die drei Workshop-Tage

1. Tag: Gemeinsame Probleme verstehen
2. Tag: Best-Practice: Zielvorstellungen entwickeln
3. **Tag: Infrastruktur (Fokus: Einzelbaumerkennung)**

Zusammenfassung der Wünsche an die Infrastruktur	09:35 - 09:45
Erstellung des Inception Decks	09:45 - 09:50
Pause	09:50 - 09:55
Erstellung des Inception Decks	09:55 - 10:15
Pause	10:55 - 11:10
Erstellung des Inception Decks	11:10 - 12:00
Mittagessen	12:00 - 13:00
Erstellung des Inception Decks	13:00 - 13:30
Pause	14:20 - 14:35
Erstellung des Inception Decks	14:35 - 14:50
Nächste Schritte	14:50 - 15:30

Parkspaziergang