

Hospitation: Auswertungen in R mit fhircrackr

Jonathan Mang



Friedrich-Alexander-Universität
Medizinische Fakultät

**Uniklinikum
Erlangen**



Wo wir stehen

1. Gestern: **Tutorial R & fhircrackr** ([Folien](#), [Material](#))

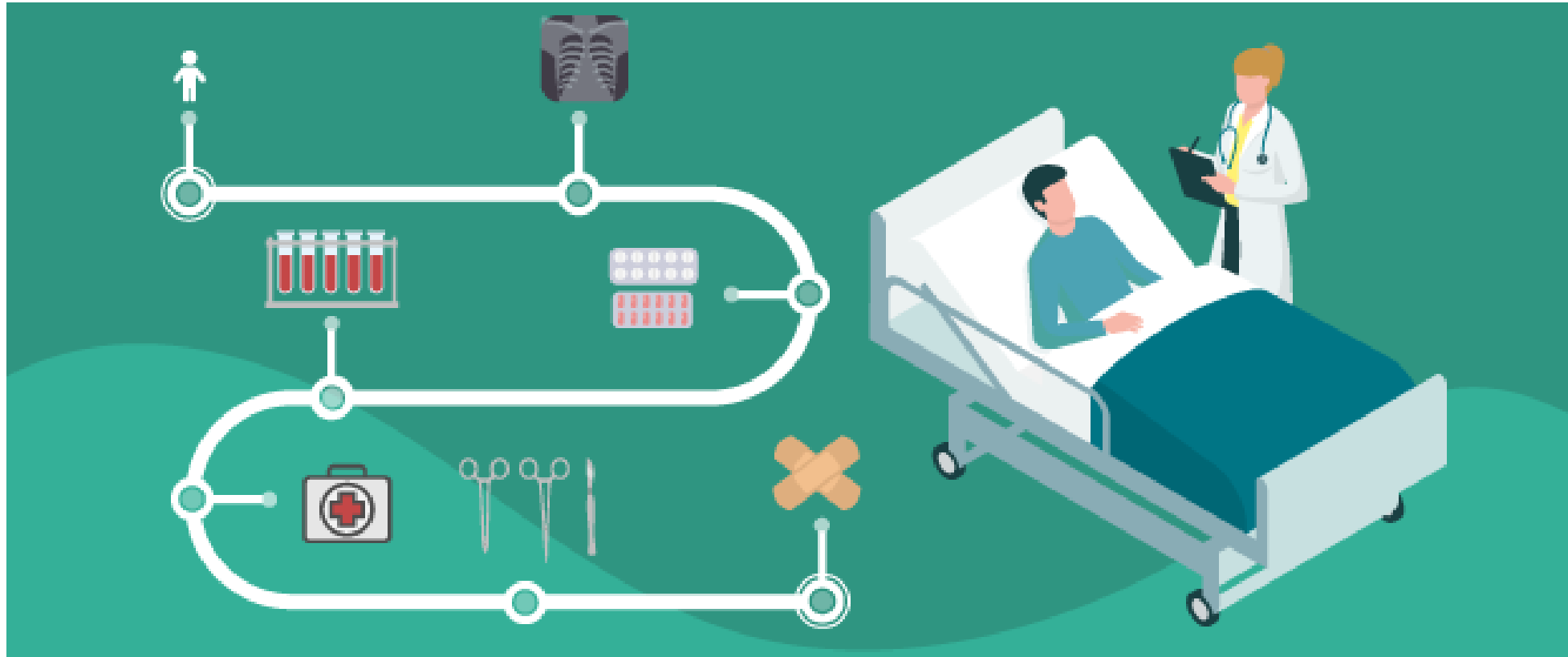
1. R / Rstudio / Datentypen / Daten einlesen / Tabellenmanipulation / Vektorrechnung
2. fhircrackr / FHIR-Ressourcen selektieren und herunterladen / Bundles / Melting / Cracking
3. Beispielanalyse „BMI/Bluthochdruck“

2. Heute: **Reproduzierbarkeit und Regulatorisches**

1. fhircrackr Übung #2
2. Was ist Reproduzierbarkeit? Wozu? Wie? Praktisch!



Warum?



Genfer (Ärzte-)Gelöbnis

- (...)
- **Ich werde meinen Beruf mit Gewissenhaftigkeit und Würde ausüben. Die Erhaltung u. Wiederherstellung der Gesundheit meiner Patienten soll oberstes Gebot meines Handelns sein.**
- (...)

- Quelle: Genfer (Ärzte-)Gelöbnis - Vom Weltärztebund 1948 in Genf beschlossene Neufassung der ärztlichen Berufspflichten (in Anlehnung an den Hippokratischen Eid; vom Dtsch. Ärztetag modifiziert)



Evidenzbasierte Medizin (Nachweisorientierte Medizin)

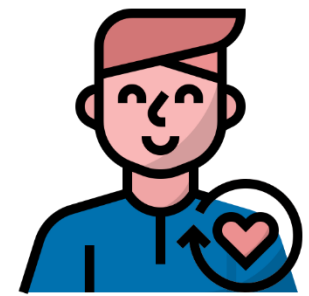
- Ziel der EbM ist der „gewissenhafte, ausdrückliche und umsichtige Gebrauch **der aktuell besten Beweise für Entscheidungen in der Versorgung** eines individuellen Patienten“ [1]



Krankheit



Therapie

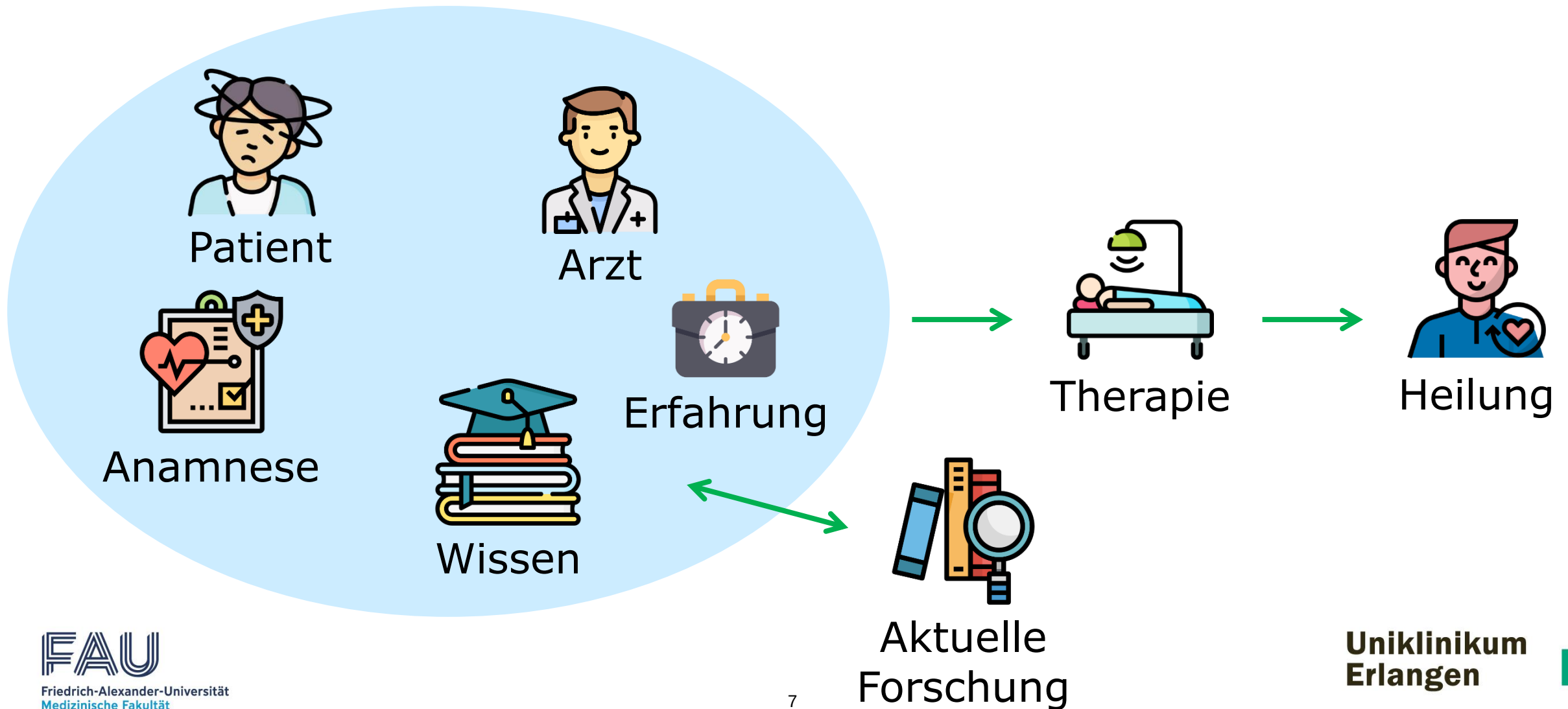


Heilung

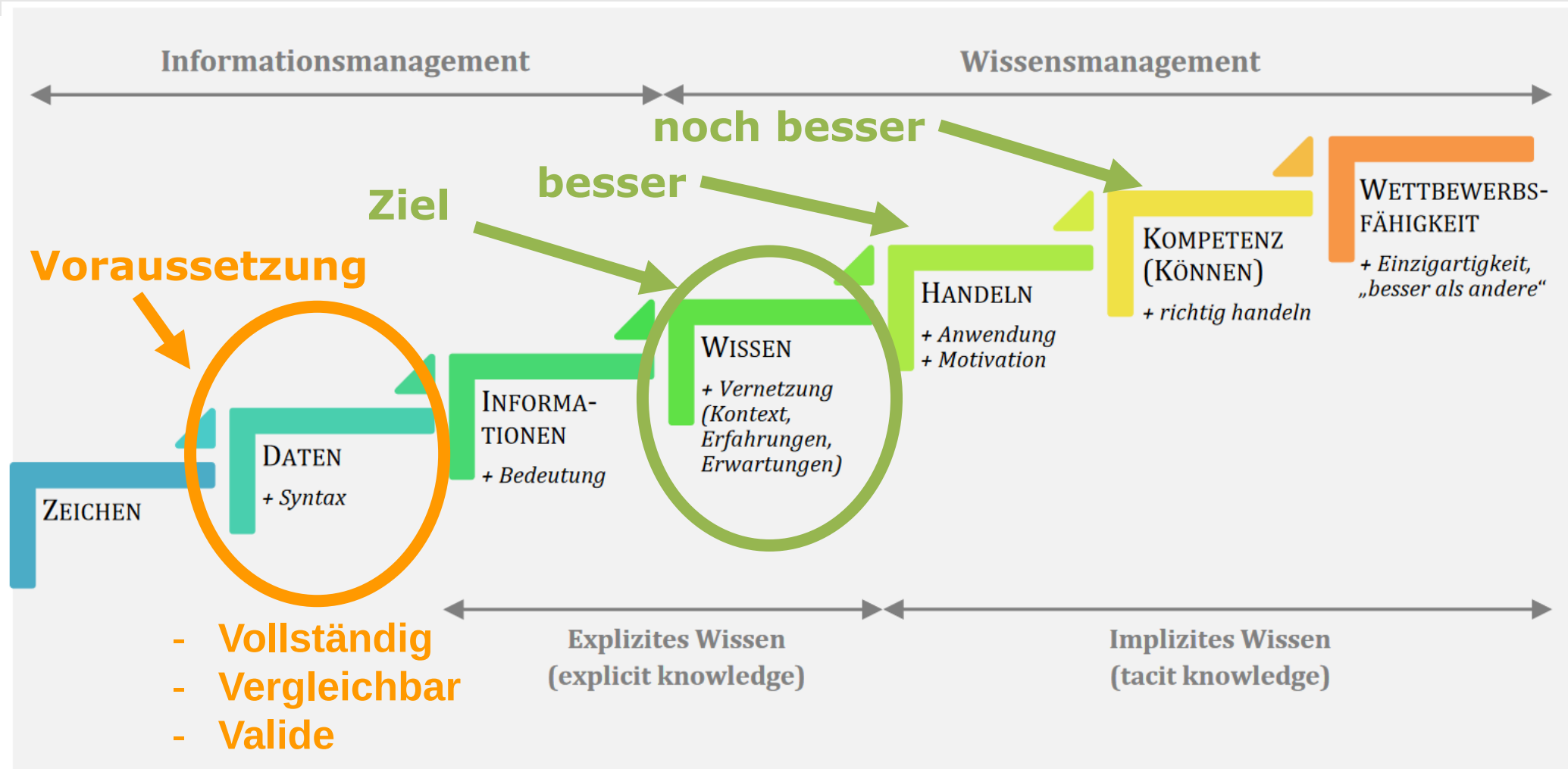
[1] <https://doi.org/10.1136/bmj.312.7023.71>



Evidenzbasierte Medizin

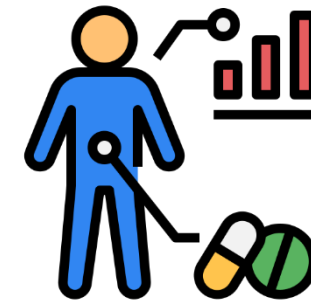
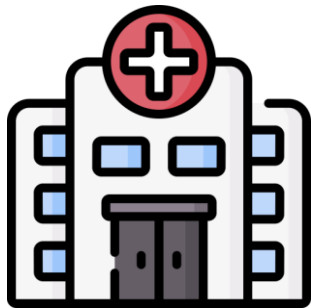


Die Wissenspyramide



Forschung braucht Daten! Aber woher?

- Vorhandene Daten direkt aus der Versorgung (= Routinedaten)
- Nachfragen (= Fragebogen), probieren (= klinische Studie) oder messen



Wo wir stehen

- Ziel: Hilfe für Patienten
 - → Historie der Wissensbeschaffung
 - Evidenzbasierte Medizin
 - Beweise für Entscheidungen
 - Wissen aus Daten
 - Routinedaten / klinische Studien
- 



Was ist Reproduzierbarkeit?

■ Replikation

- Wiederholung einer Studie
- Ziel: Kontrolle, Überprüfung der Ergebnisse
- Konsequenz: Akzeptanz steigern, Betrug & Fälschung reduzieren

■ Reproduktion

- Ausgehend von Daten: Analyse überprüfen

→ Reproduzierbarkeit = Replikation + Reproduktion



Factors contributing to the lack of reproducibility

1. A **lack of access** to methodological details, raw data, and research materials.
2. **Inability** to manage complex datasets
3. **Poor research practices** and experimental design
4. **Cognitive bias**
5. A **competitive culture** that rewards novel findings and undervalues negative results



3 Schritte zur Reproduzierbarkeit

1. before data analysis,
2. during the analysis, and
3. after analysis



Before data analysis

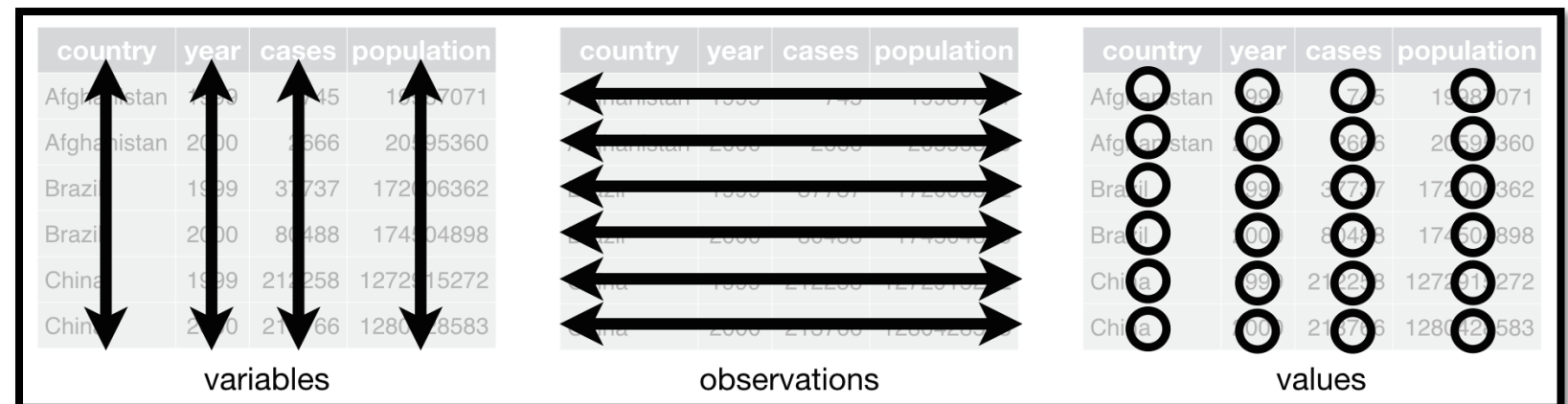
■ Solide Datenverwaltung

- Organisation/Struktur
- Backups

■ „Tidy“ Datenformat ^[1]

- intrinsisch: `install.packages("tidyr"); vignette("tidy-data")`
- Datenformat: csv, txt

[1] H. Wickham, 'Tidy Data', *Journal of Statistical Software*, vol. 59, pp. 1–23, Sep. 2014, doi: [10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10).



During analysis: best coding practices

- Ver-**skript**-en, was geht
- Kommentare im Code
- Sauberer, konsistenter Code-Style
 - camelCase / snake_case
 - demo_data_mat / systol_vec
- Automatisierbare Teile → Funktionen
- Hart-codierte Werte auslagern
- Container



```
7 #make this example reproducible
8 set.seed(0)
9
10 #define data
11 x1 <- rnorm(100)
12 x2 <- rnorm(100)
```

A space between parameters

No space between the function name and parentheses between the parentheses and the parameter

Curly brace { on the same line, after a space

Spaces around operators

Indentation 2 spaces

A space after for/if/while...

A semicolon ; is mandatory

A space between arguments

Lines are not very long

An empty line between logical blocks

} else { without a line break

Spaces around a nested call

```
function pow(x, n) {
  let result = 1;
  for (let i = 0; i < n; i++) {
    result *= x;
  }
  return result;
}

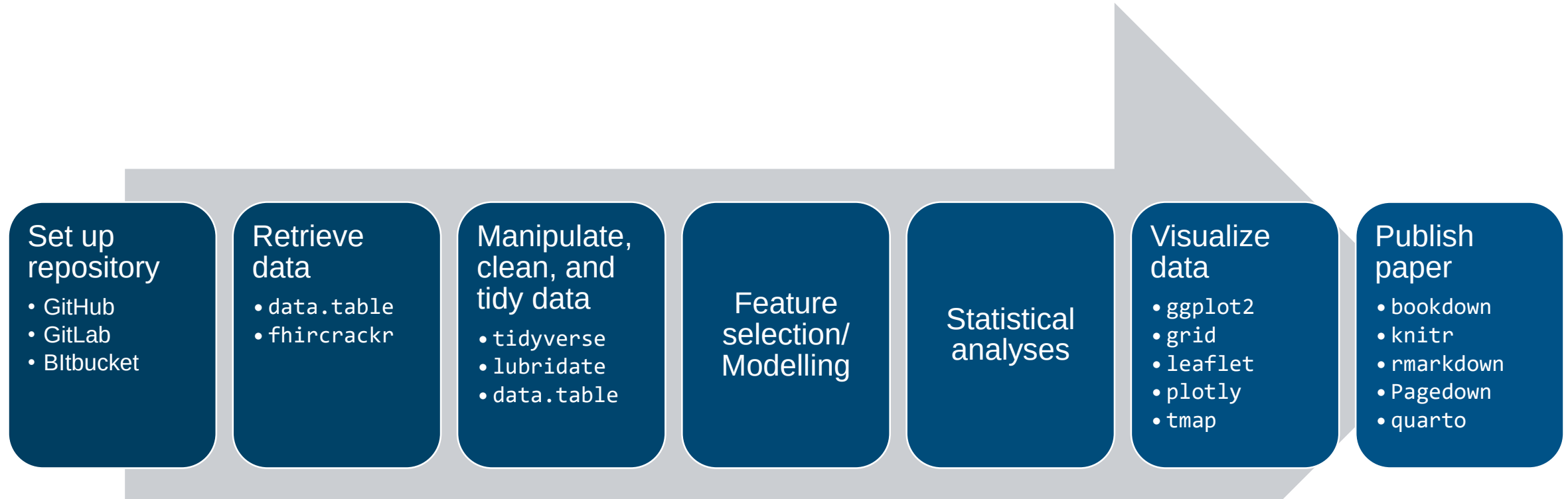
let x = prompt("x?", "");
let n = prompt("n?", "");
if (n < 0) {
  alert(`Power ${n} is not supported,
  please enter a non-negative integer number`);
} else {
  alert( pow(x, n) );
}
```

After analysis

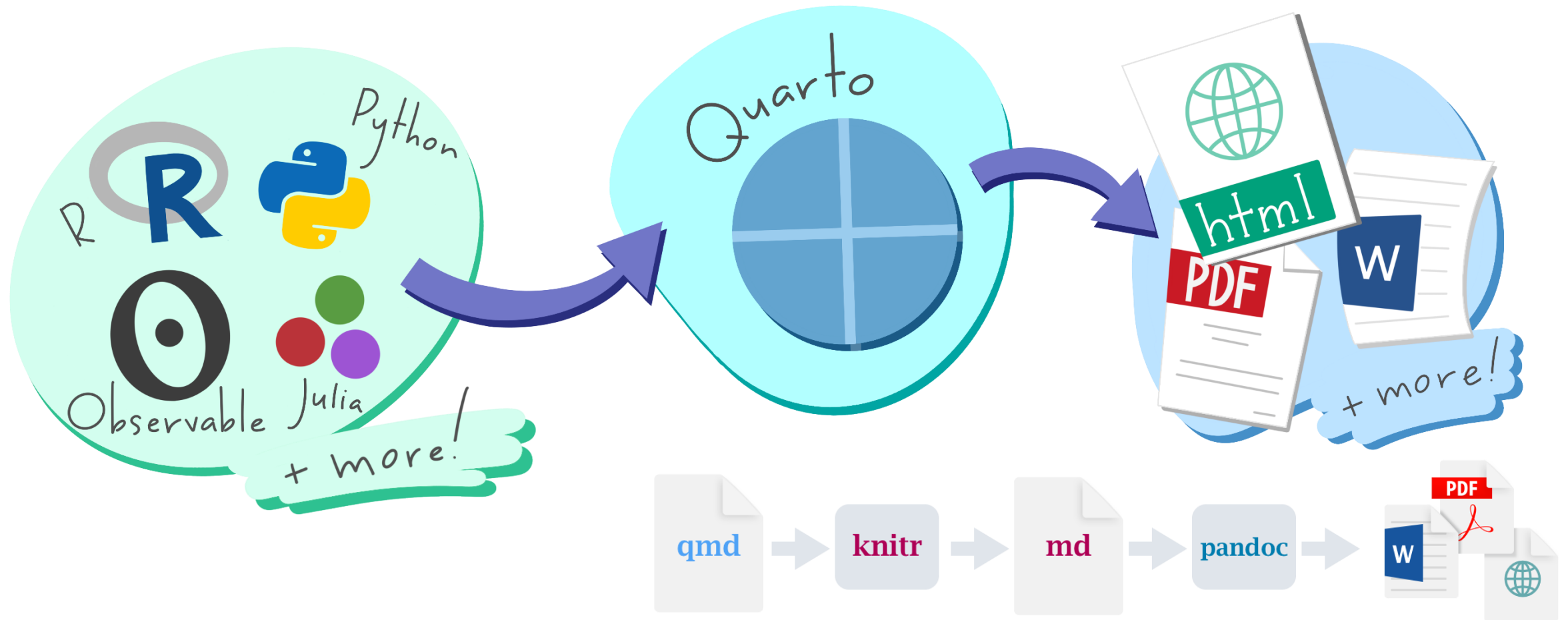
- Details zur Analyse veröffentlichen
 - Code, Versionen, Parameter, Daten
- Abbildungen automatisch erzeugen lassen
 - Kein Illustrator, PPT, Paint, Photoshop
 - Vorteil: Aktuelle Datengrundlage, korrekte Nummerierung
- Daten, Code und Ergebnisse archivieren
 - Forschungsdatenmanagement!



Data pipeline



1x Code → n x Results



Änderungen?

- Neuer Proband?
 - Fehler in den Daten?
 - Datenschema ändern?
-
- → Kein Problem. Ändern, neu „knitr“n/builden, fertig



“Conducting reproducible research is not exceedingly difficult nor does it require encyclopedic knowledge of esoteric research tools and protocols.” ^[1]



“Whether they know it or not, most researchers already perform much of the work required to make research reproducible.” ^[1]

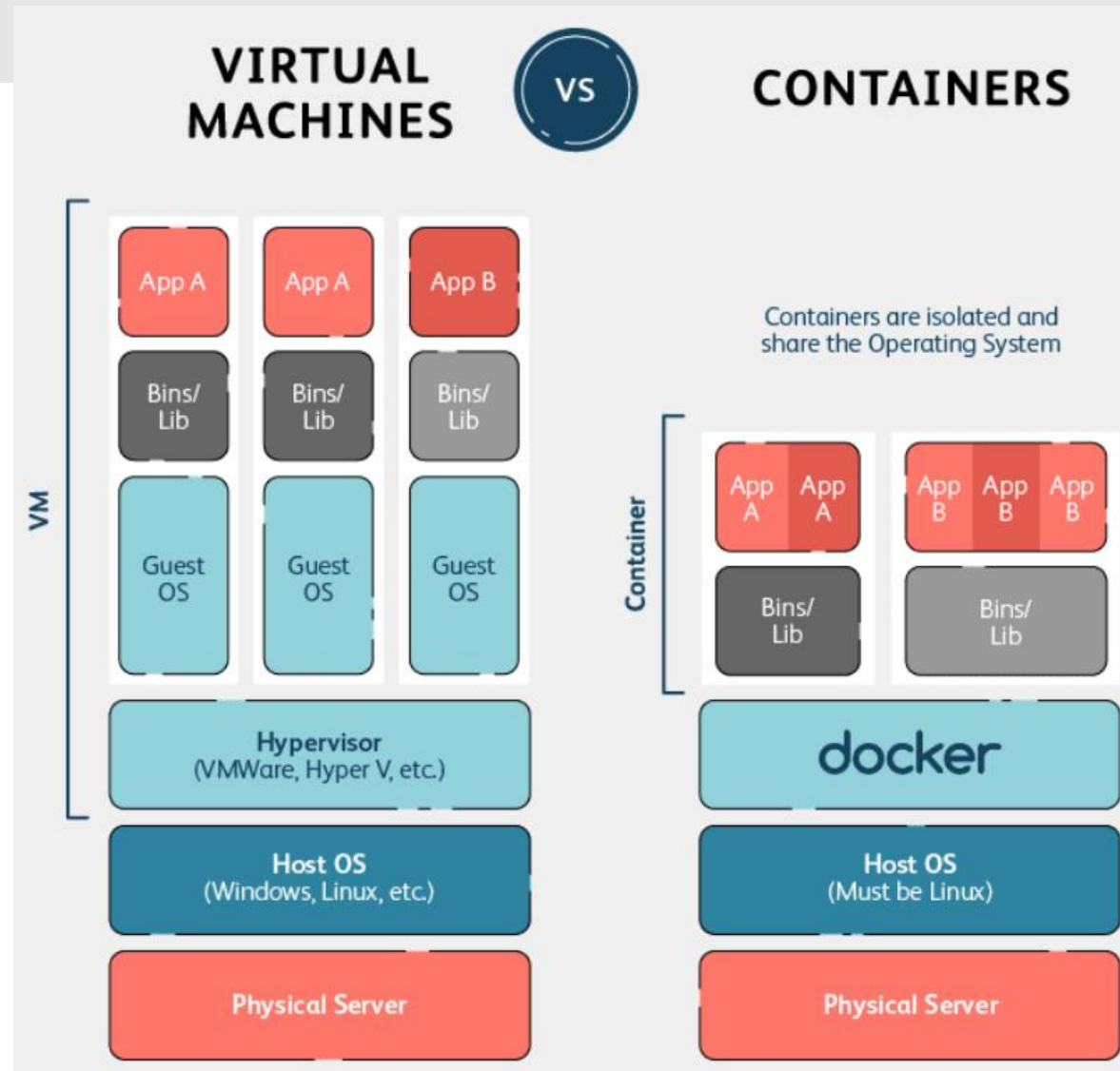


Wie können wir reproduzierbar arbeiten?

Docker is a software platform that allows you to **build, test, and deploy applications** quickly, packaging software into standardized units called **containers**.



Docker Containers vs. Virtual Machines

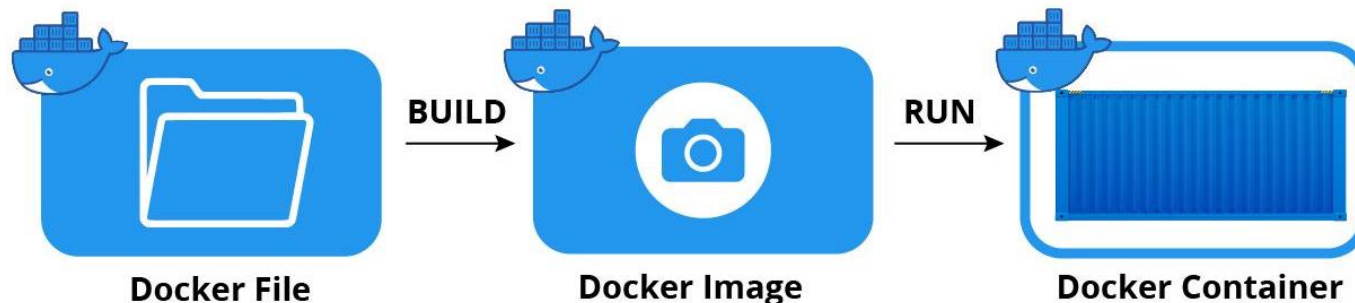


Quelle:
<https://gridscale.io/blog/container-as-a-service-caas/>



Docker Vokabular

- **Docker Image:** The basis of a Docker container. Represents a full application
- **Docker Container:** The standard unit in which the application service resides and executes
- **Docker Engine:** Creates, ships and runs Docker containers deployable on a physical or virtual, host locally, in a datacenter or cloud service provider
- **Registry Service** (Docker Hub or Docker Trusted Registry): Cloud or server based storage and distribution service for your images



Quelle: <https://jfrog.com/knowledge-base/a-beginners-guide-to-understanding-and-building-docker-images/>



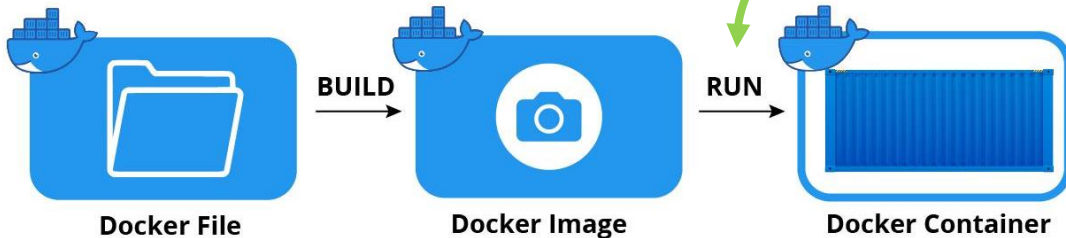
Deployment

Dockerfile

```
FROM centos:7
RUN yum install -y python-devel python-virtualenv
RUN virtualenv /opt/indico/venv
RUN pip install indico
COPY entrypoint.sh /opt/indico/entrypoint.sh
EXPOSE 8000
ENTRYPOINT /opt/indico/entrypoint.sh
```

docker-compose.yml

```
version: "3"
services:
  my-application:
    build: ./
    ports:
      - "8000:8000"
    environment:
      - CONFIG_FILE.env
  db:
    image: postgres:14
  redis:
    image: redis:7-alpine
    command: redis-server --save "" --appendonly no
    ports:
      - "6379"
```



Reproduzierbare Analysen

```
FROM rocker/r-ver:4.2.2
```

```
RUN mkdir -p /Ergebnisse && \  
  mkdir -p /errors && \  
  mkdir -p /Bundles
```

```
RUN apt-get update -qq && \  
  apt-get install -y \  
  --no-install-recommends \  
  curl \  
  ## `libpq-dev` for RPostgres  
  ## (one of the suggests of the packages):  
  libpq-dev \  
  libssl-dev \  
  libxml2-dev
```

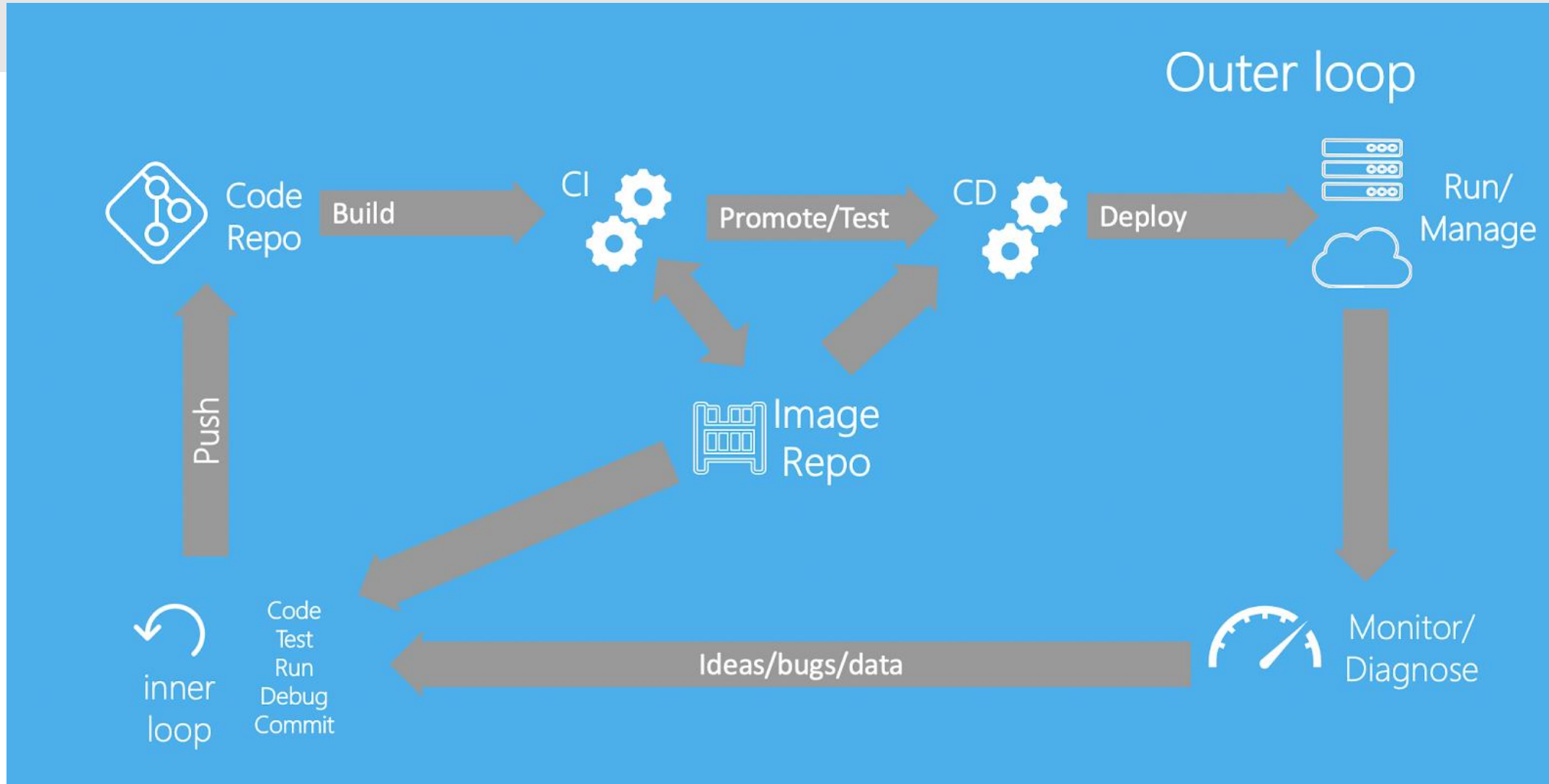
```
RUN install2.r --error \  
  --deps TRUE --skipinstalled \  
  fhircrackr \  
  config \  
  dplyr \  
  zoo \  
  stringr \  
  tidyr \  
  data.table \  
  openxlsx
```

```
COPY script.R script.R  
COPY install_packages.R install_packages.R
```

```
CMD ["Rscript", "script.R"]
```

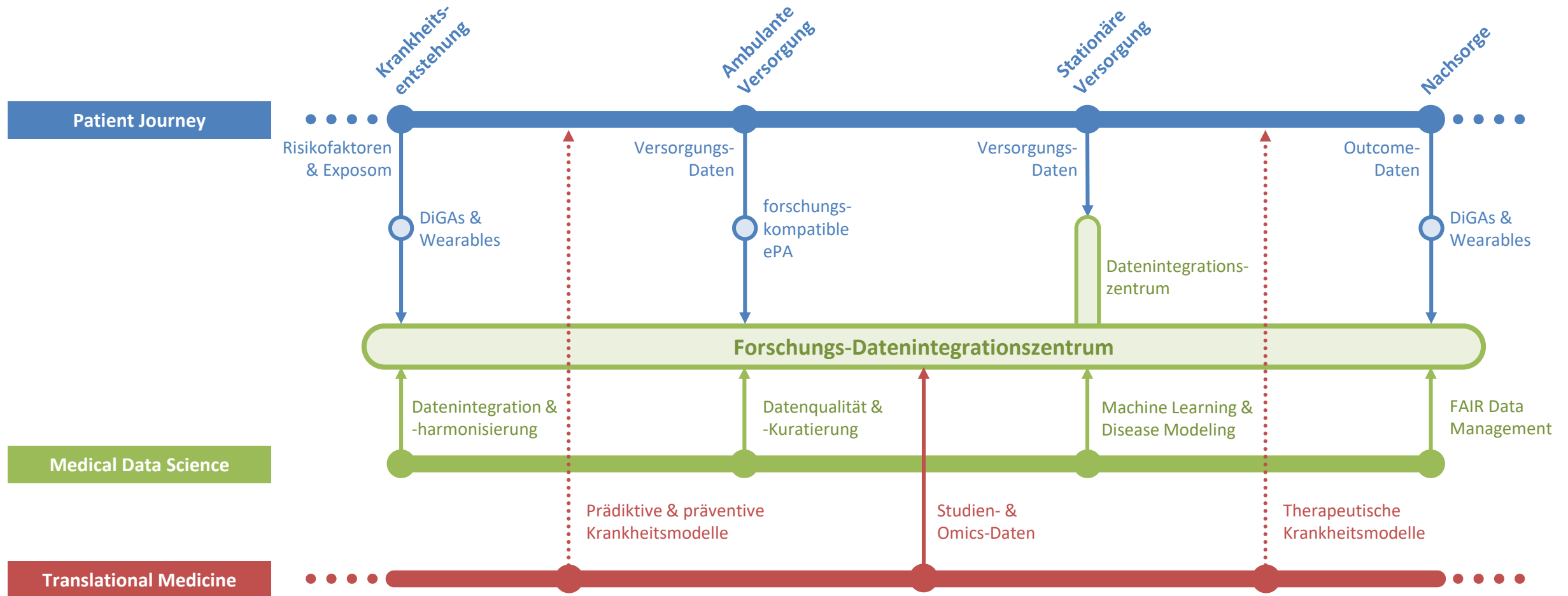


Continuous Integration/Delivery/Deployment (CI/CD)



Digitales Disease & Data Management

entlang der gesamten Patient Journey



Was zusätzlich noch notwendig ist

- Interoperabilität: **Harmonisierung**, Standards (FHIR), Metadaten (MDR)
- **Datenqualität**: Vergleichbare Daten!
- **Rahmenbedingungen**: Datenschutz, Nutzungsrecht, DSGVO, MPG, Zertifizierung, Audits, intern: UAC / Review
- Mechanismen und Tools zur **Erfassung fehlender oder zusätzlicher Daten** (EDC/eCRF)
- Einwilligung, Record-Linkage, Pseudonymisierung != Anonymisierung
- **FAIR**es Datenmanagement (Findable, Accessible, Interoperable, Reusable)
- **Komplexität** (Mathematik, Datenschutz, Bürokratie, Aggregation) **von Multisite-Analysen** „bring the code to the data“



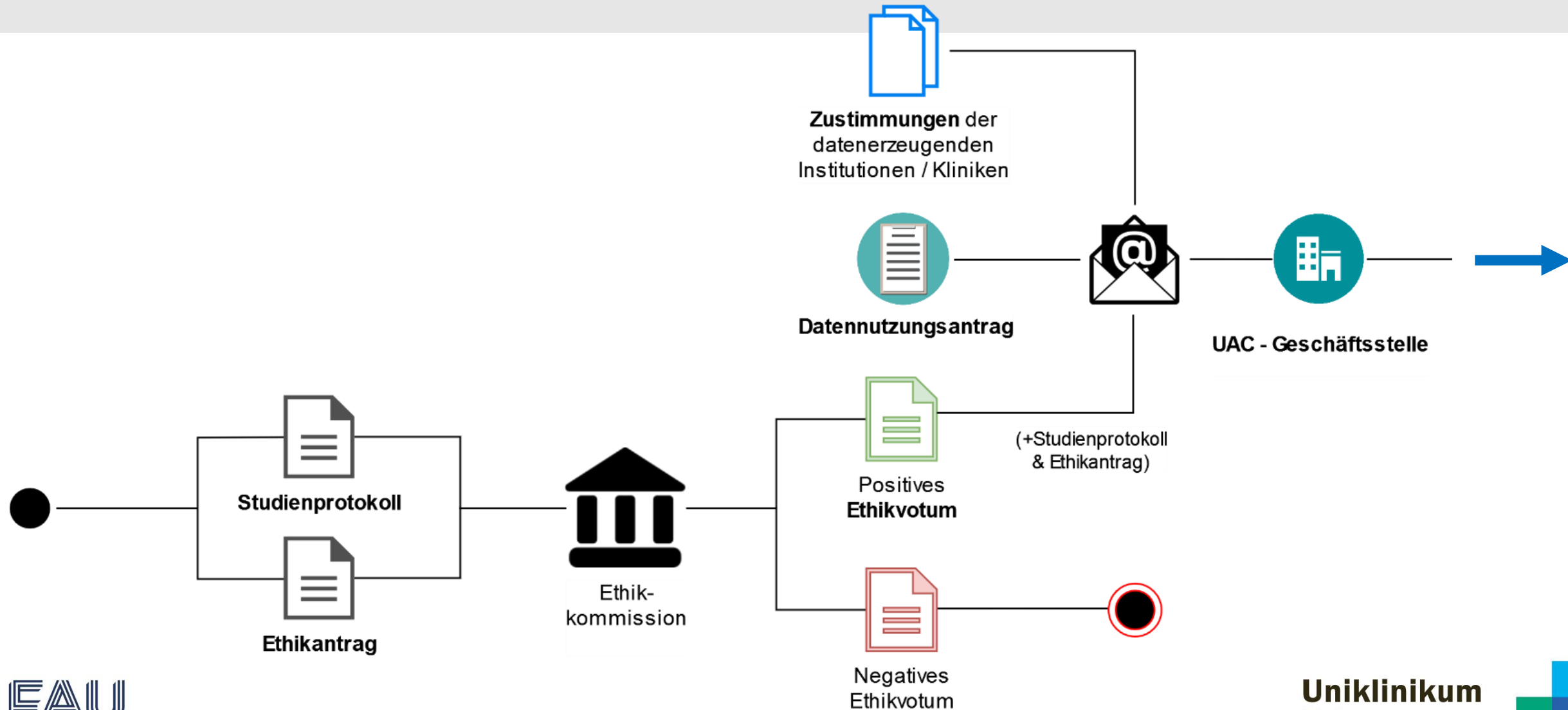
Von der Idee zu den Daten

:D

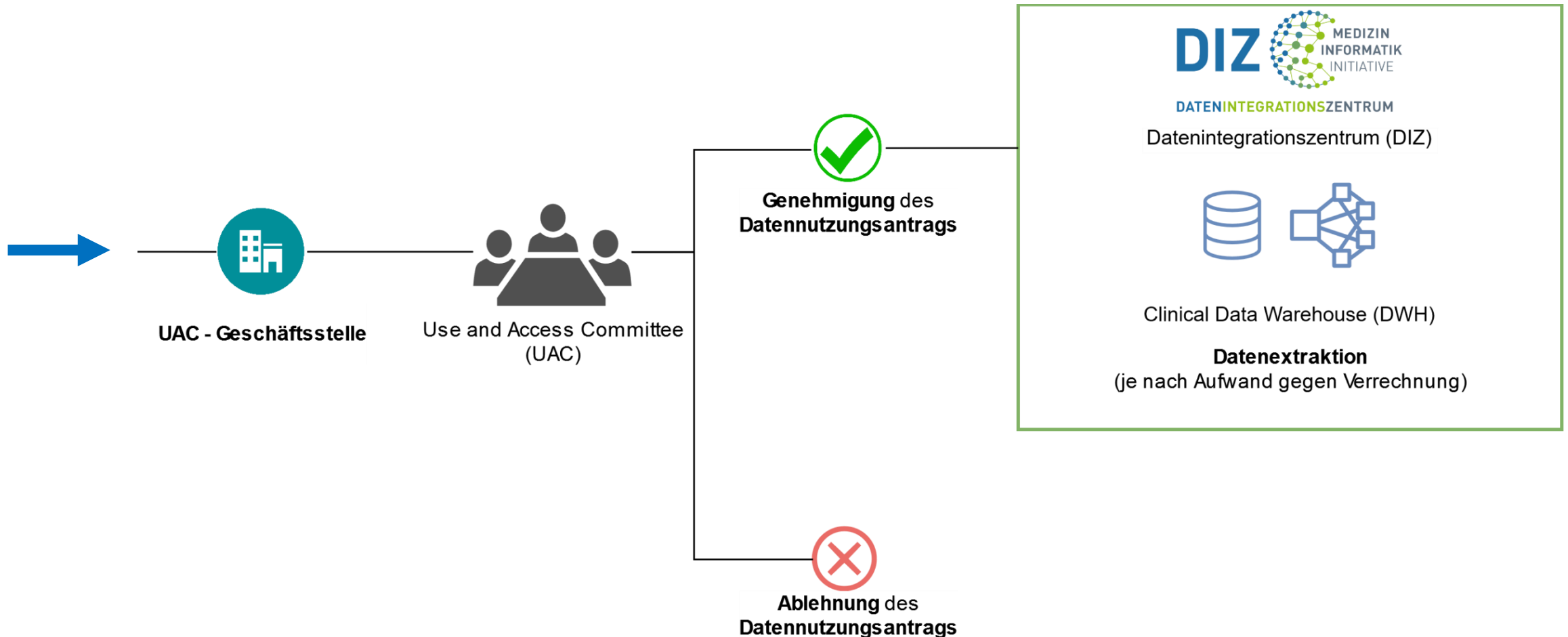
1. Studienprotokoll & Ethikantrag schreiben
2. Auf Antwort der Ethikkommission warten
3. Datennutzungsantrag beim UAC stellen
4. Kostenübernahme zusichern
5. Auf UAC-Votum warten
6. Auf Datenextraktion warten
7. Analyse schreiben
8. (Dasselbe an allen Standorten wiederholen)



Von der Idee zu den Daten



Von der Idee zu den Daten

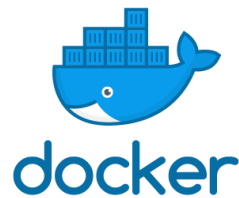


Take Home Messages



kubernetes

- Reproduzierbarkeit: Replikation + Reproduktion
- R kann: Skripte, Reports, Präsentationen, Apps, ...
- Nicht nur die Analyse, auch die IDE, Testumgebung, Infrastruktur, Container, ETL-Jobs, DBs – (fast) alles: Code → Versionierung, Reproduzierbarkeit, Automatisierung
 - Aber: Ordnung! Im Code, Kommentaren, Daten, Methodik
- Es gibt viele Tools zur Unterstützung:



Vielen Dank!



jonathan.mang@fau.de



[+49 \(0\) 9131 / 85 - 47299](tel:+49(0)91318547299)

Fragen?



[joundso](#)



[0000-0003-0518-4710](#)



[jonathan-mang](#)

