Ordnung ist die halbe Miete:
Forschungsdatenmanagement als Voraussetzung für Methoden der Künstlichen Intelligenz

Thüringer Zentrum für Lernende Systeme und Robotik

Oliver Mothes

19.10.2022

Oliver Mothes

- Wissenschaftlicher Mitarbeiter in der **Computer Vision Group** (Prof. Dr. Joachim Denzler) und am **zedif** der FSU Jena
  - Verschiedene Machine Learning / Deep Learning Methoden
  - Multiple Object Tracking
  - Projekte in der Biomedizin und Industrie
- Koordinator des Verbundprojektes **THInKI**
- Mitorganisator des Jenaer KI-Stammtiches **JENA.AI**
- Transferkoordinator Wissenschaft des **TZLR**

Was ist das TZLR?

INITIATOREN

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IOSB
Institutsteil Angewandte Systemtechnik AST

DLR Deutsches Zentrum für Luft- und Raumfahrt

TZLR

FÖRDERUNG

Carl Zeiss Stiftung

Freistaat Thüringen
Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft

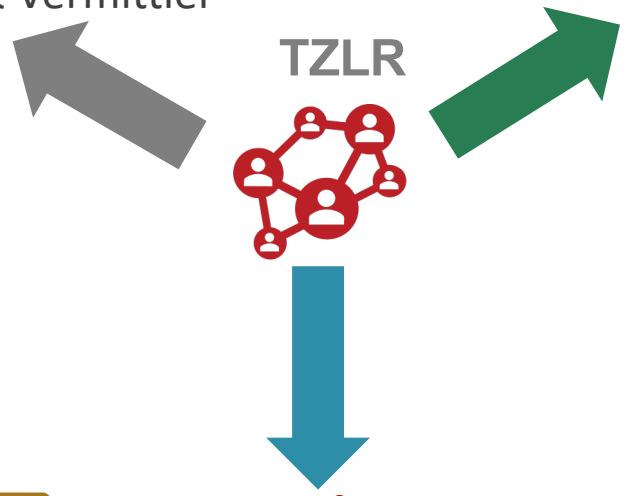
Ziele des TZLR

- Verbesserung der nationalen und internationalen Sichtbarkeit
- Bündelung von Thüringer Kompetenzen
- Verbesserung der Wettbewerbssituation
- Etablierung einer Institution als Ansprechpartner
- Abstimmung und Erfahrungsaustausch

Unsere Handlungsfelder

Wissensaustausch
- Wissenschaftler & Anwender
- Organisation von Workshops
- KI-Forum
- Ansprechpartner & Vermittler

Wissensvermittlung
- Beratung
- Qualifizierungs- & Weiterbildungsmaßnahmen
- Eignung & Einsatz von KI-Methoden



Wissenstransfer
- Auftragsforschung/Dienstleistungen für Unternehmen
- Proof of Concepts
- Machbarkeitsstudien

Was ist Forschungsdatenmanagement (FDM)?

→ Prozess der **Umwandlung**, **Auswahl** und **Speicherung** von Forschungsdaten mit dem Ziel, sie **unabhängig vom Datenautor** über einen **langen Zeitraum hinweg zugänglich, wiederverwendbar** und **reproduzierbar** zu machen.

(Quelle: forschungsdaten.info)

Warum ist FDM notwendig ?

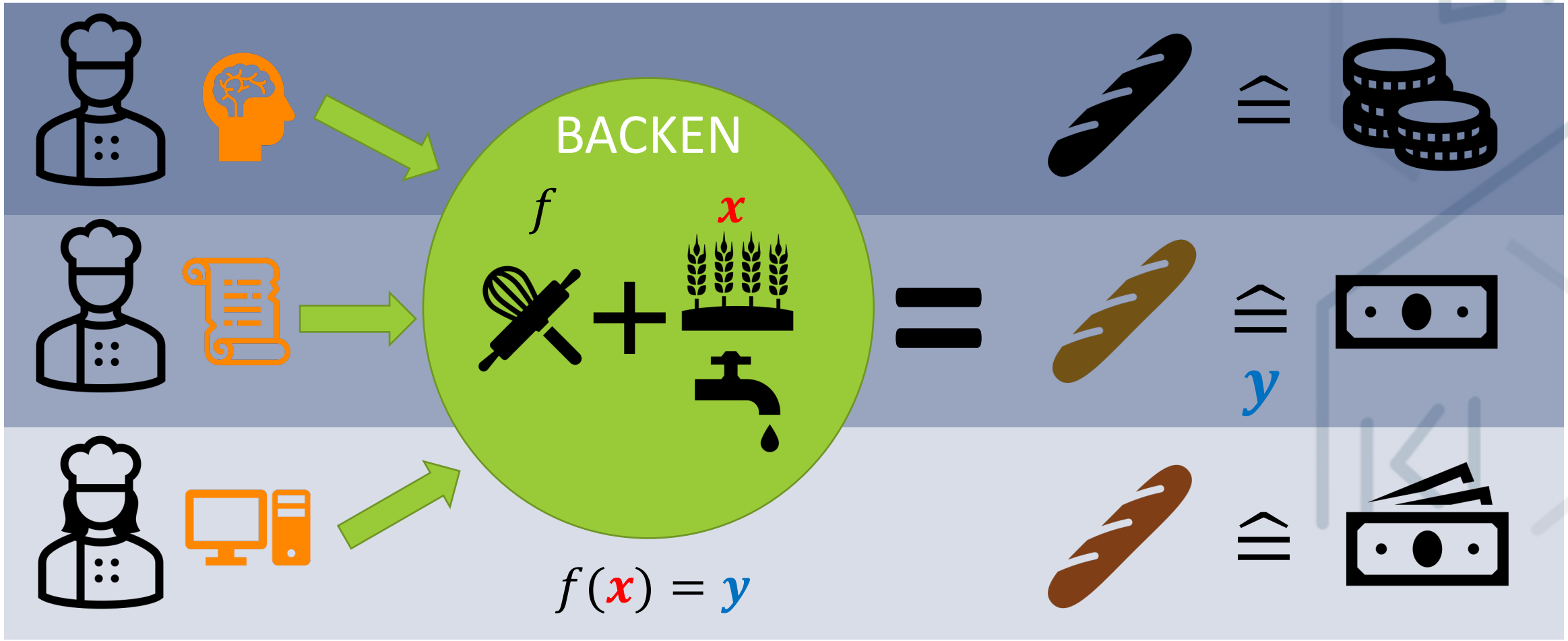
Vorgaben

- Gute wissenschaftliche Praxis
- Leitlinien von Institutionen
- Gesetzliche Regelungen
- Gefordert von Fördergeldgebern

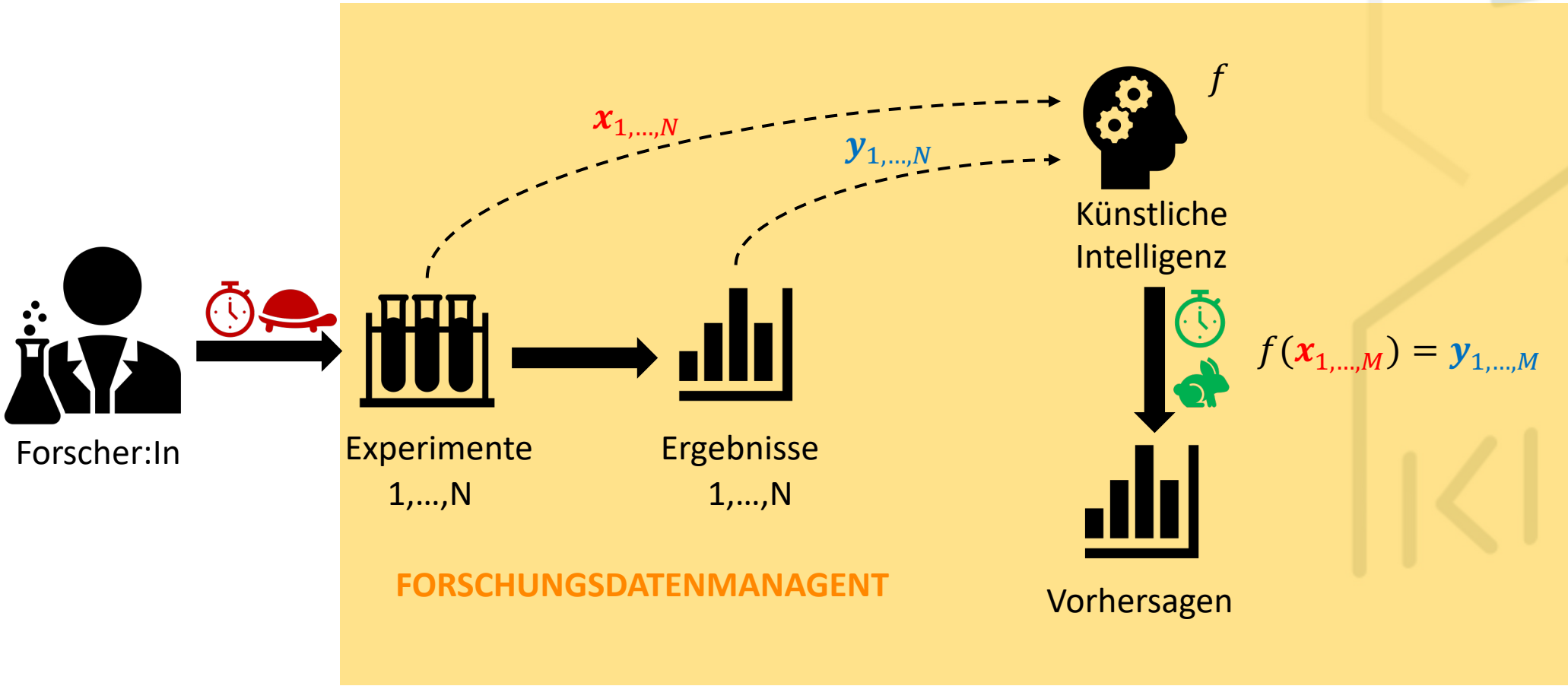
Vorteile

- Ersparnis von Zeit und Ressourcen
- Erhöhte Datensicherheit
- Bessere Zusammenarbeit und Nachnutzung
- Verifizierbarkeit und Reproduzierbarkeit

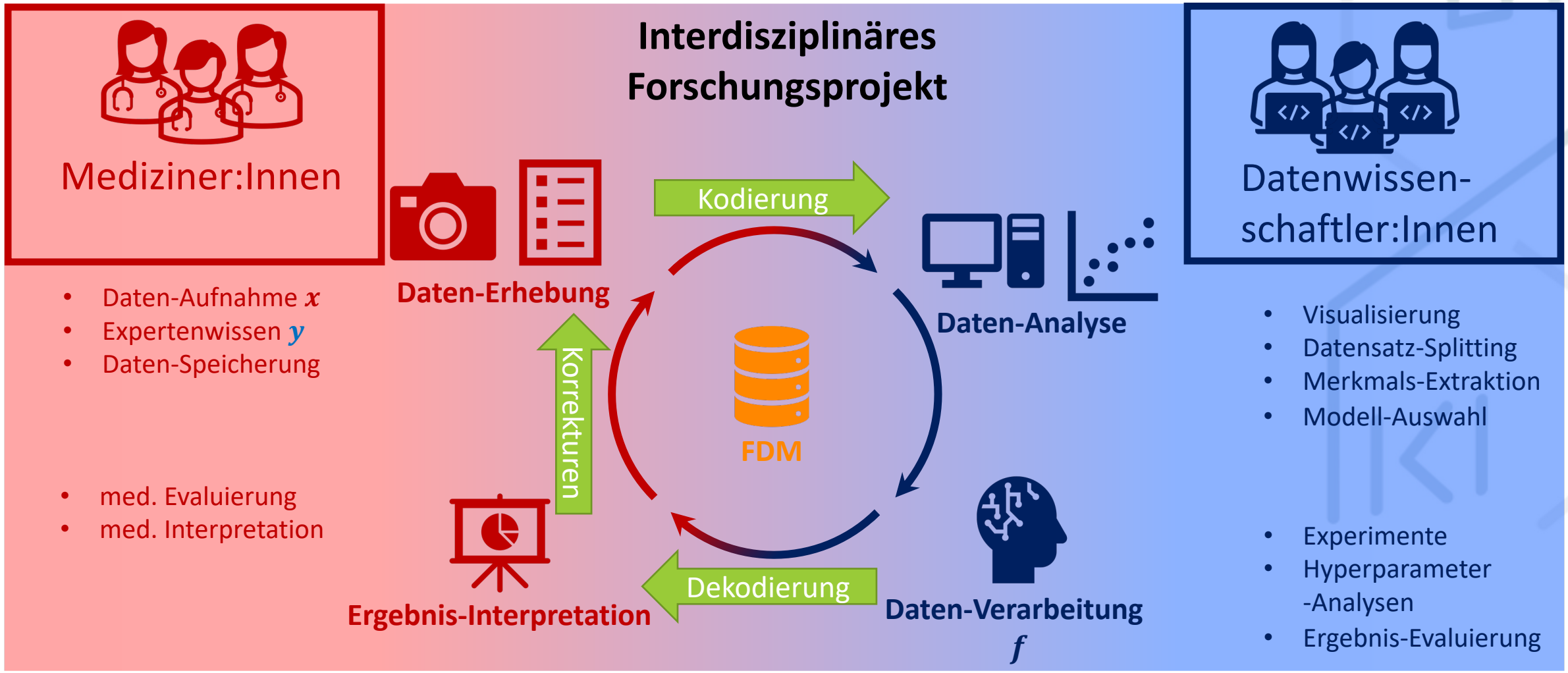
Wer bäckt die besseren Brötchen?



Ein Beispiel aus der Forschung mit KI



Ein Beispiel aus der Praxis



Kontextverluste am Beispiel

- Verteilung von Daten
  - Beispiel 1: Häufigkeiten von Symptomen bei Krankheiten

Krankheit	A	B	C	D
Symptom X	ja	ja	nein	ja
Symptom Y	ja	nein	ja	ja
Häufigkeit (%)	60	17,5	17,5	5

→ Wichtig für Datensatz-Split und Evaluierung

  - Beispiel 2: „gesunde“ und „kranke“ Probanden einer Studie

Zustand	gesund	krank
Häufigkeit (%)	99,99	0,01

→ Wichtig für die Modellwahl

Speicherung & Backup

Speicherung & Backup

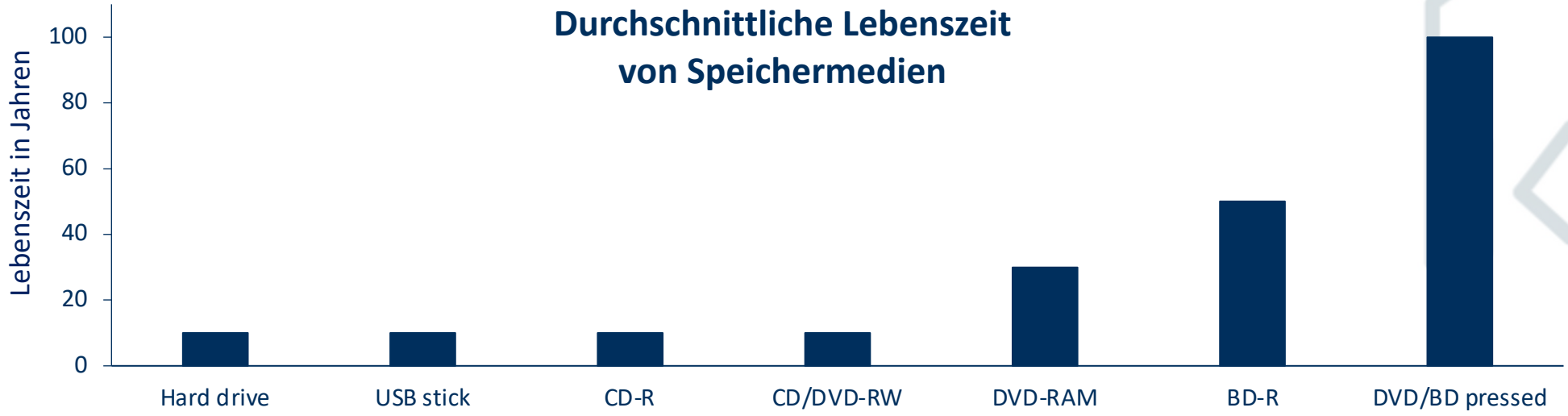
Verwaltung

Dokumentation

Speicher-Medium und Zugriff

Wer hat Zugriff auf Daten?

Privater Laptop / PC **vs.** Arbeits-Laptop / PC **vs.** Externes Medium (z.B. USB-Stick) **vs.** Private / Institutions-Cloud **vs.** Server der Institution



Quelle: J.Rex (2019), DOI: 10.5281/zenodo.2579580

Datenverluste vermeiden

- **Gründe für Datenverluste:**
  - Hardware-, Software- und Benutzerfehler
  - Malware und Hackerangriffe
  - Unfälle und Naturkatastrophen
  - Diebstahl

Backups

3-2-1-Backup-Regel

Daten-Verwaltung

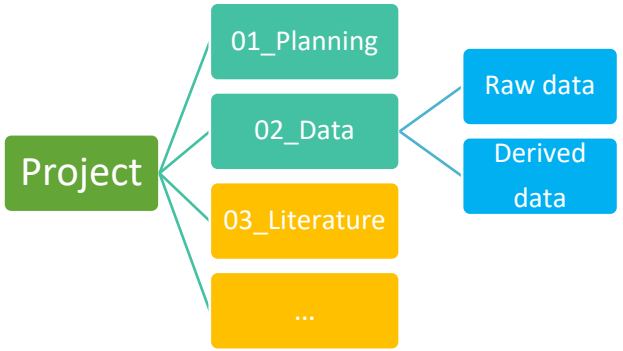
Speicherung & Backup

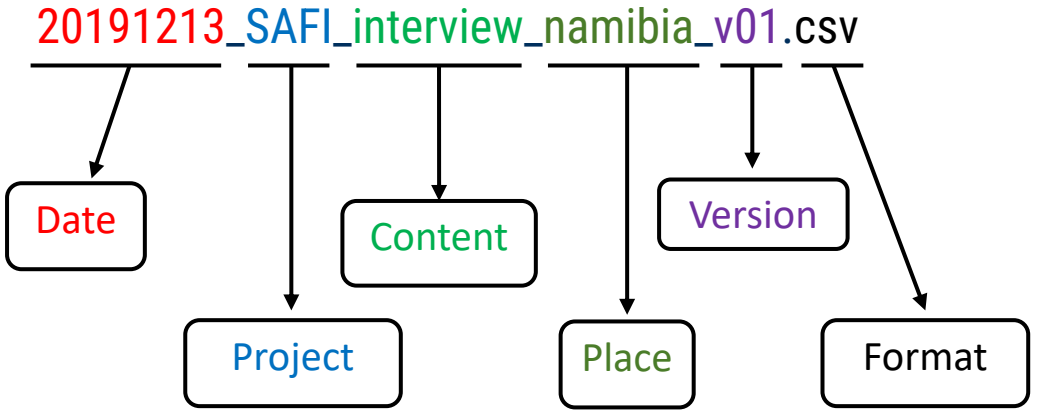
Verwaltung

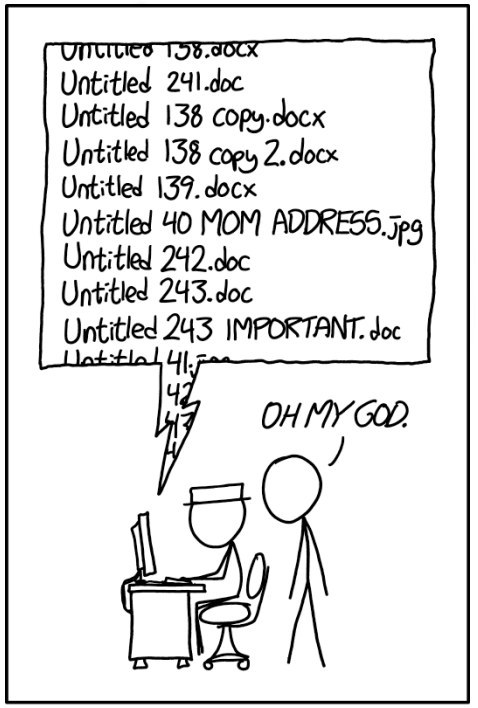
Dokumentation

Verwaltung und Versionierung

- **Datenstrukturen & Namenskonventionen**:







- **Versionsverwaltung**: Git, Subversion, etc.

Dokumentation

Speicherung & Backup

Verwaltung

Dokumentation

Was beinhaltet eine Dokumentation

WER hat die Daten erzeugt?

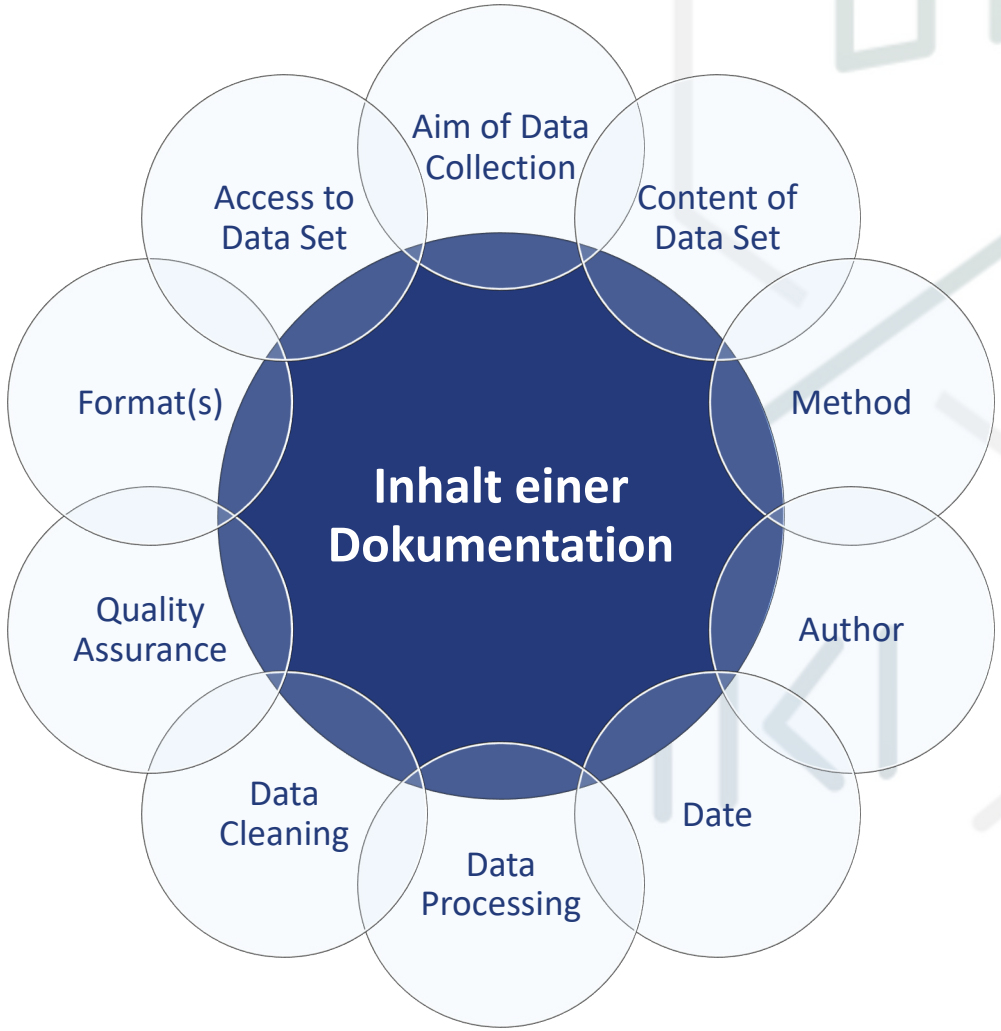
WAS beinhalten die Daten?

WANN wurden Daten erstellt?

WO wurden die Daten erzeugt?

WIE wurden die Daten erzeugt / verarbeitet?

WARUM wurden Daten erzeugt?



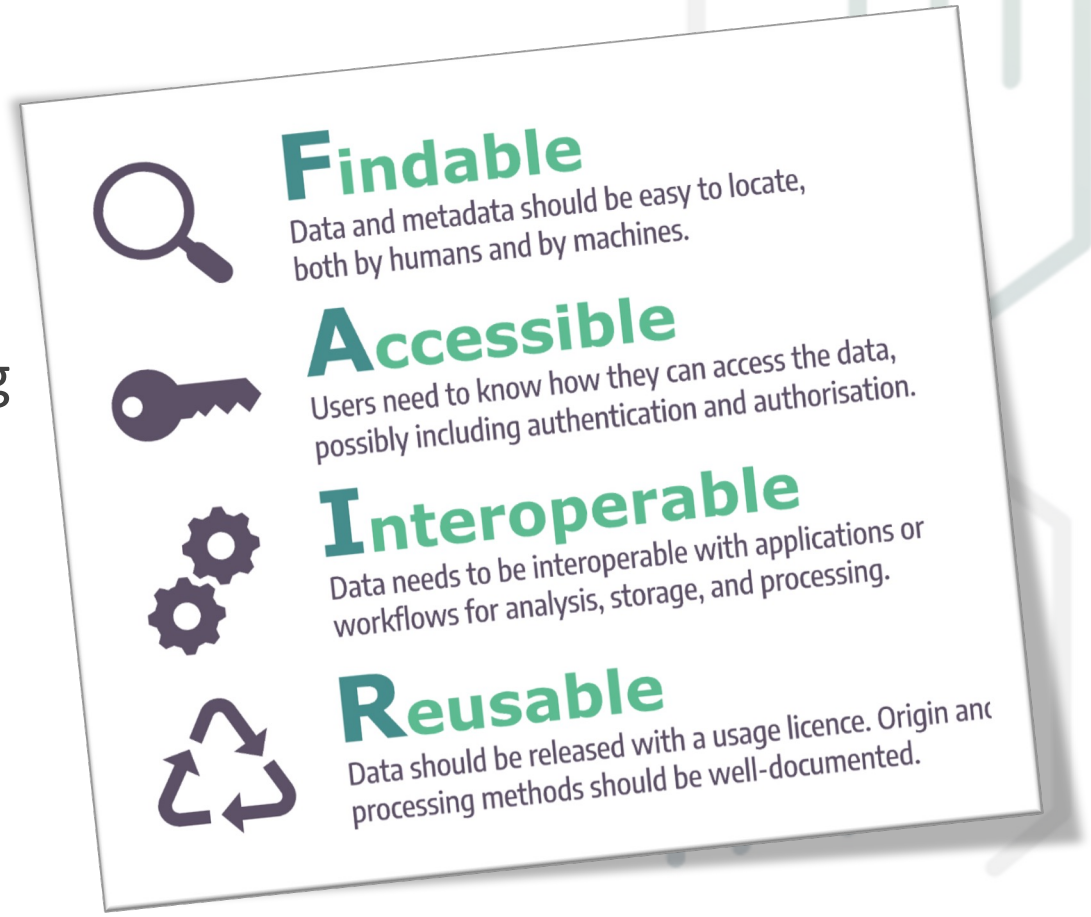
Speicherung & Backup

Verwaltung

Dokumentation

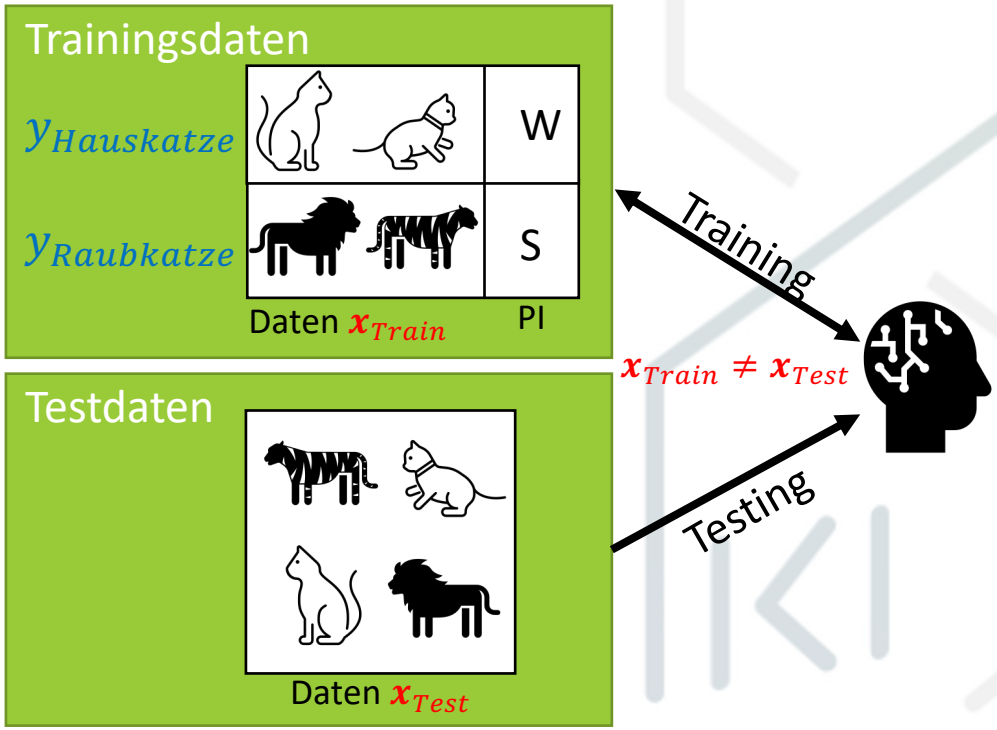
Fazit

- Einhaltung der **FAIR**-Prinzipien
  - **F**indable, **A**ccessible, **I**nteroperable, **R**eusable
  - Konsistenz in der Datenstruktur / -benennung
- Statistiken zu erhobenen Datensätzen schon bei Datenerhebung mit beachten
- Jede Information zu Daten sind hilfreich → Metadaten



Wie kann (Forschungs-) Datenmanagement die KI unterstützen?

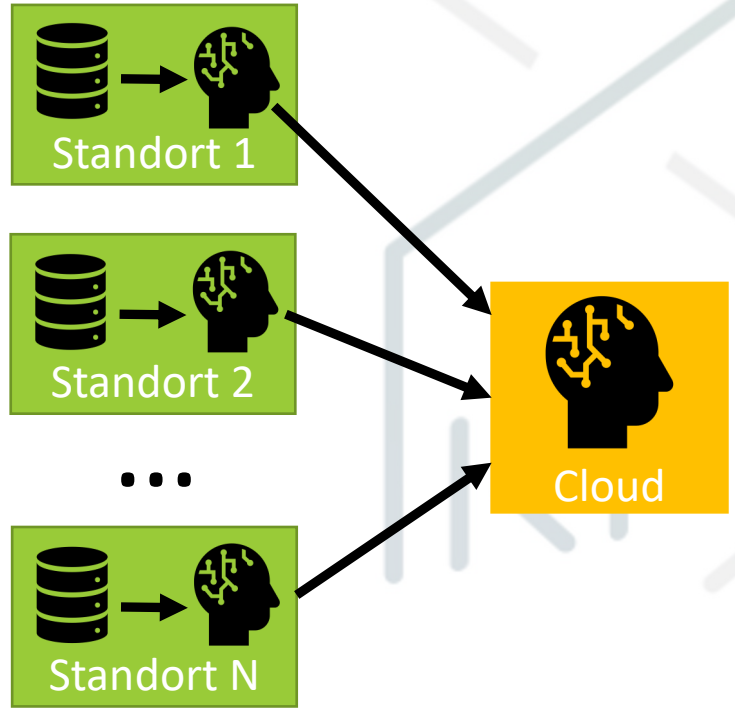
- **Learning using Privileged Information (LUPI)**
  - **Metadaten** zusätzlich als Privileged Information (PI) nutzen
  - PI unterstützt indirekt beim Trainieren des Modells



Vladimir Vapnik, Akshay Vashist, A new learning paradigm: Learning using privileged information, Neural Networks, Volume 22, Issues 5–6, 2009, Pages 544-557,

Wie kann (Forschungs-) Datenmanagement die KI unterstützen?

- **Federated Learning**
  - (Sensible) Daten dezentral kollaborativ nutzen
  - Bsp.: nicht-anonymisierbare Patientendaten (Gesichtsbilder)



Konečný, Jakub; McMahan, Brendan; Ramage, Daniel (2015). "Federated Optimization: Distributed Optimization Beyond the Datacenter". arXiv:1511.03575

Wie kann (Forschungs-) Datenmanagement die KI unterstützen?

- **FDM unterstützt außerdem:**
  - **Reproduzierbarkeit** von KI-Methoden und deren Ergebnisse
  - **Verstehen** von KI-Methoden
  - **Teilen** von KI-Modellen und deren Methoden
  - uvm.

oliver.mothes@tzlr.de

Vielen Dank für Ihre Aufmerksamkeit!