Datenintegration in der Medizin

Aktuelle Herausforderungen aus der Sicht eines Wanderer's zwischen den Welten

Toralf Kirsten

Win-Meeting, Leipzig, 10.09.2016

Motivation

Aktuelle Untersuchungsdaten
- Symptome,
- „Messdaten“

kontinuierliche Behandlungsdaten:
- Messdaten
- Labordaten
- „Leistungen“
- Dokumentation

Diagnosen

Behandlungsstrategien (Leitlinien)

Rekonstruktion des Tatherganges, ~ der Krankheitsgeschichte

Lokalität und Umgebung

Literatur
- Dokumentation ähnlicher Fälle
- Forschungsergebnisse

Erfassung von Untersuchungsdaten

- Arten
  - Anamnese (Interview, Fragebögen)
  - Messdaten
- Erfassung von Anamnesedaten unterschiedlich
  - Versorgung
    - bei Hausärzten / Ambulanzen oftmals unstrukturiert: Text
    - in Kliniken mit extra Modul im hauseigenen SAP-System
  - Forschung
    - vermehrt strukturiert :-)
    - vorangig verwendete Software: Excel :-(
    - komplexere Software-Systeme: RedCap, LimeSurvey, ...

Erfassung von Untersuchungsdaten

- Genauigkeit von Messungen
  - Einsatz von genormten Instrumenten
  - Trainingsaufwand bei manuellen Messungen zur Reduzierung untersucherabhängiger Effekte
- Messsysteme als Appliance-Lösung (Preis!)
  - Getrennte Systeme: Patientenverwaltung vs. Messsysteme
  - Häufige Lösung: Aufnahme des Ergebnisses in deskriptiver Form

Dokumentation

- „Umfangreiche“ Dokumentation in Form verschiedener Dokumente
  - Arztbriefe, Pflegedokumente, Laborberichte, ...
- Probleme:
  - Deidentifikation
  - unstrukturierte Daten zu verschiedenen Entitäten
  - Paragrammatik: Spezialsprache der Mediziner
  - Text Mining Algorithmen vor allem für Englisch
  - Keine Verfügbarkeit eines deutschen Text-Corpus
  - Kein kostenloser Zugriff auf deutschsprachige Ressourcen (Nachschlagewerke, Terminologien, Ontologien, …)
- Vorarbeiten: Text Mining zum Zwecke der Abrechnung

Beispiel: Medikamentenanamnese

1. 1x Dekristol 400IE
2. 1x1 Tbl. Dekristol 400 IE
3. anfangs Dekristol
4. D-Christol
5. Decristol
6. Deekristol
7. Dekrestol
8. Dekristol
9. dekristol
10. Dekristol 400 IE
11. Dekristol 400µg
12. Dekristol (Vit. D)
13. Dekristol 400
14. Dekristol 400 Zµg
15. Dekristol 400 Žµg
16. Dekristol 400 I.E.
17. Dekristol 400 I.E. tgl
18. Dekristol 400 IE 1x
19. Dekristol 400 µg
20. dekristol 400µg
21. Dekristol 800 OE
22. Dekriston
23. Dekriston 400
24. Dekristrol
25. Dektristol
26. Dektritol
27. Destristol 400 IE
28. Dkristol 400µg
29. Vitamin D (Dekristol)
30. Dekristol (Vitamin D)
31. Vit. D3, Dekristol
32. Dekritsol
33. Dwkristol

<u>Aufnahme von Medikamentengebrauch</u>
- Medikamentenname (oder Pharmazentralnr.)
- Dosis, Darreichungsform, Einnahmeregel

Orangendrink Brausetablette

1. damit der zyklus regelmäßig ist
2. Cef...
3. Cholesterol
4. Cortisonpräparat
5. Nasentropfen (Osana)
6. Natriumhyluronat
7. ?
8. 999
9. Fiebersaft
10. Fieberzäpfen
11. Formel 90 Protein
12. Pantomine
13. Proteinpulver Whey
14. Proteinshakes
15. Taurin
16. Tolotricin
17. Tropfen gegen Birkenpollen, Sedo...
18. Xylit Lutschtablette
19. emser
20. gegen Magenbeschwerden
21. salbe mit Citronensäure und 85% Glycerol
22. unbekannte Beruhigugnstabletten
23. unbekanntes Antibiotikum

- Falschschreibungen
- Schwankende Genauigkeit der Angaben (Produkt- vs. Wirkstoffnamen)
- Angabe inkl. Verabreichungsform und Einnahmeregel

- Was ist ein Medikament?
  - Apothekenpflichtige M.
  - Drogerieartikel
- Abschätzung der Auswirkung von Nicht-Medikamentangaben
  - Systemische vs. lokale Wirkung

➡ **Aufnahme abhängig vom Ziel**

Beispiel: Medikamenetenanamnese

- Adaptiver Thesaurus-basierter Ansatz zur Abbildung aller Schreibweisen (Falsch~ + Richtig~)
- **Kernidee**:
  - Thesaurus enthält Korrektur jeglicher Falschschreibung
  - Iterative Verbesserung (Verifizierung von Neueinträgen)
  - Korrekturaufwand mit wachsendem Thesaurus stark sinkend (Hypothese)
- Lernphase
  - Paarweiser Vergleich mit bekannten Falschschreibungen und *Namen bei noch unbekannten Schreibweisen
  - Kontrolle bei Einfügungen in Thesaurus

T. Kirsten et al.: On Integrating and Preparing Drug Data in an Epidemiological Study. In preparation.

Thesaurusbasierter Ansatz

Medikamentname aus Eingabe: **Dwkristol**

Auszug aus **Thesaurus**

DRUG_INPUT_NAME	DRUG_ORIG_NAME
1x Dekristol 400IE	Dekristol
1x1 Tbl. Dekristol 400 IE	Dekristol
anfangs Dekristol	Dekristol
D-Christol	Dekristol
Decristol	Dekristol
Deekristol	Dekristol
Dekrestol	Dekristol
Dekristol	Dekristol
dekristol	Dekristol
Dekristol 400 IE	Dekristol
Dekristol 400µg	Dekristol
Dekristol (Vit. D)	Dekristol
Dekristol 400	Dekristol
Dekristol 400 Zµg	Dekristol
Dekristol 400 Žµg	Dekristol
Dekristol 400 I.E.	Dekristol
Dekristol 400 I.E. tgl	Dekristol
Dekristol 400 IE 1x	Dekristol
Dekristol 400 µg	Dekristol
dekristol 400µg	Dekristol
Dekristol 800 OE	Dekristol
Dekriston	Dekristol

nGram-basierter Vergleich von S_1 und S_2

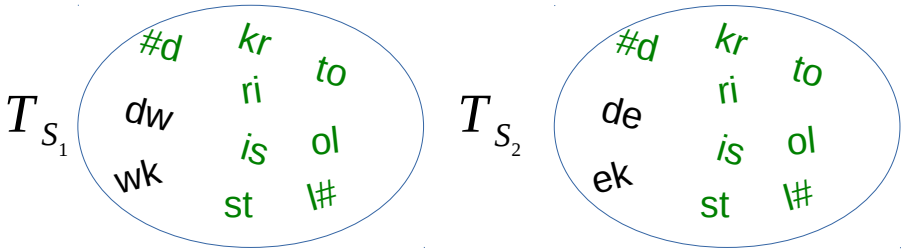
mit allen Eingabenamen

- BiGram
- Dice-Ähnlichkeit

$$s_{Dice}(S_1, S_2) = \frac{2 * |(T_{S_1} \cap T_{S_2})|}{|T_{S_1}| + |T_{S_2}|} \in [0,1] \subset \mathbb{R}$$

Beispiel:

S_1 = #dwkristol# S_2 = #dekristol#



$$s_{Dice}(S_1, S_2) = \frac{2 * 8}{10 + 10} = \frac{4}{5}$$

$$s_{Dice}(S_1, S_2) \geq t$$

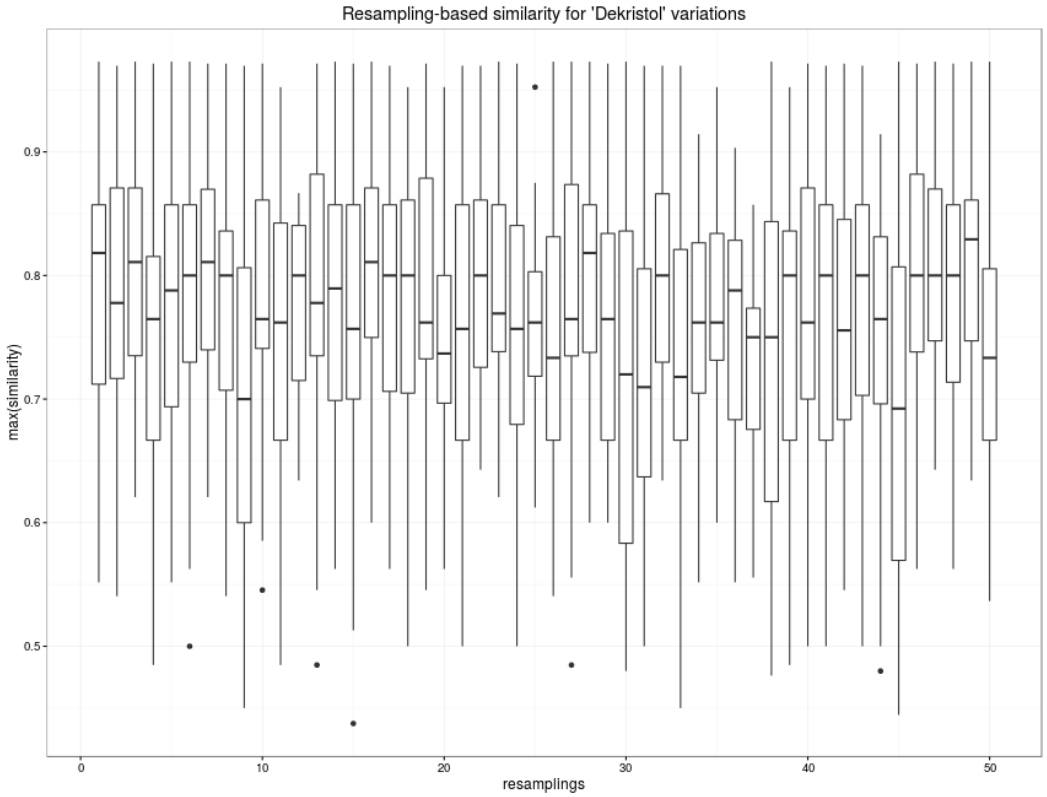
gefunden: s_{Dice}!=1?, dann Einfügen (Eingabename → *Name)

nicht gefunden: einfügen (Eingabename → *Name)

Kontrolle und mgl. Korrektur notwendig

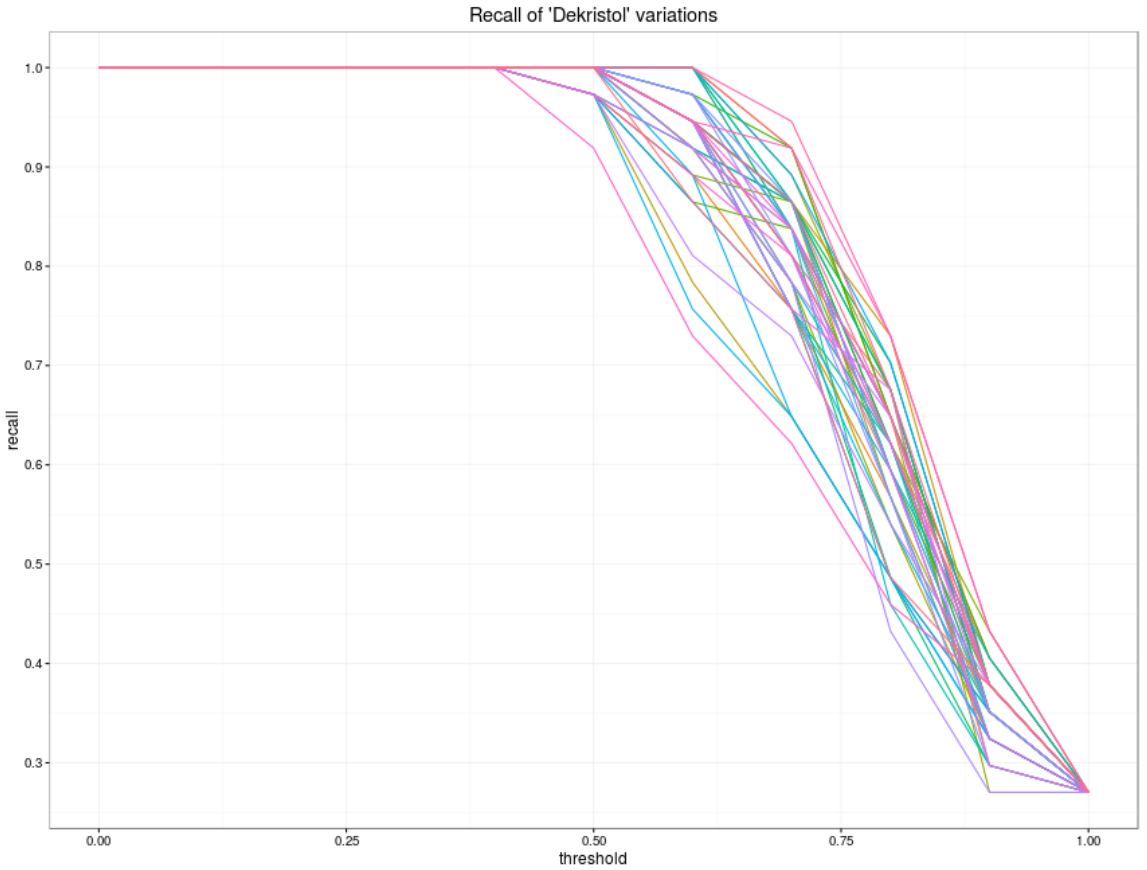
Evaluierung

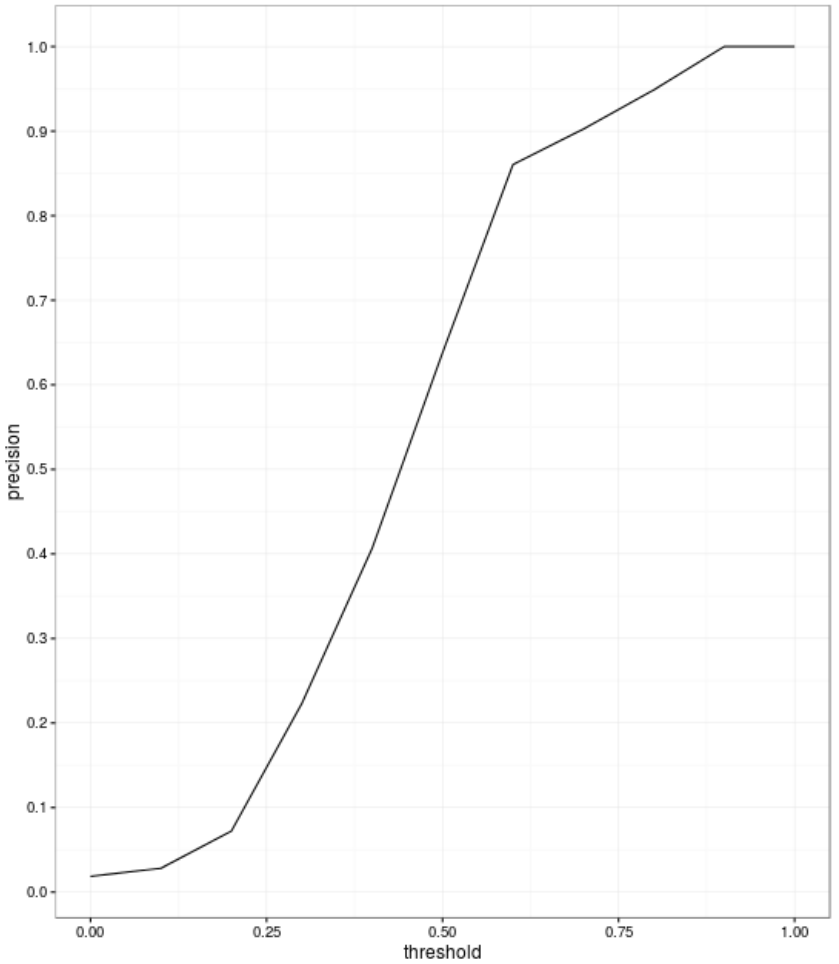
- Triviale Tests nicht zielführend
  - Neustart (leerer Thes.), automatisch: $Thes_{V_0} \cap Thes_{V_1} \approx 0$
  - Test Med. gegen bestehenden Thesaurus: 100%
- Sampling
  - |'Dekristol'-Einträge|=37
  - |Stichprobe|=10 aus allen 'Dekristol' zugeordneten Einträgen als Thesaurus
  - Test 27 Namen → Thes.
  - 50 Wiederholungen (Resampling)



Evaluierung

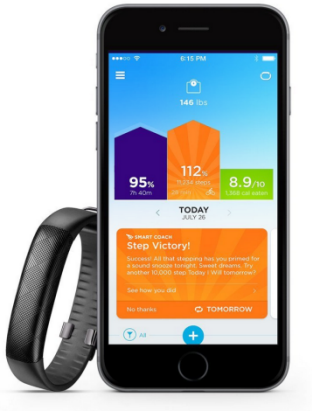
- Recall-Analyse: #Treffer?
  - Threshold 0.9 → 0.27-0.4 Recall
  - Threshold 0.5 → 0.95 – 1.0 Recall
- Precision-Analyse: # FP?
  - Threshold 0.9 → Precision 1.0
  - Threshold 0.5 → Precision 0.63



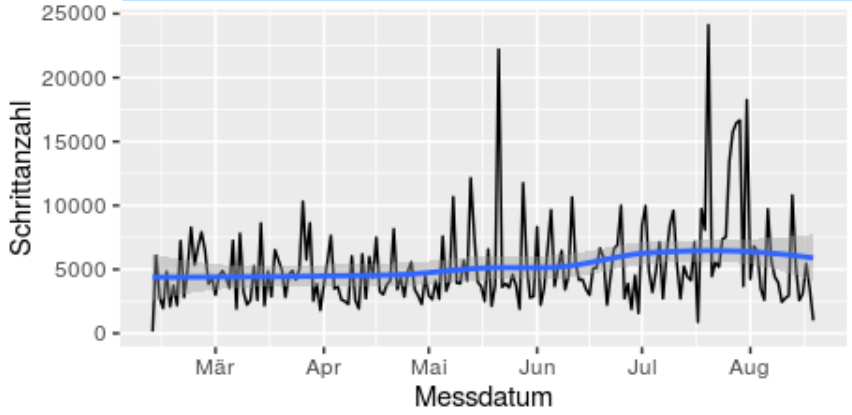


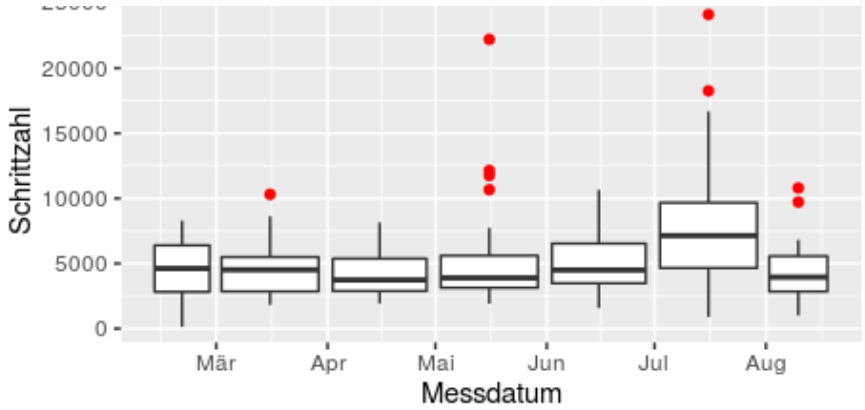
Sensoren zum Monitoring

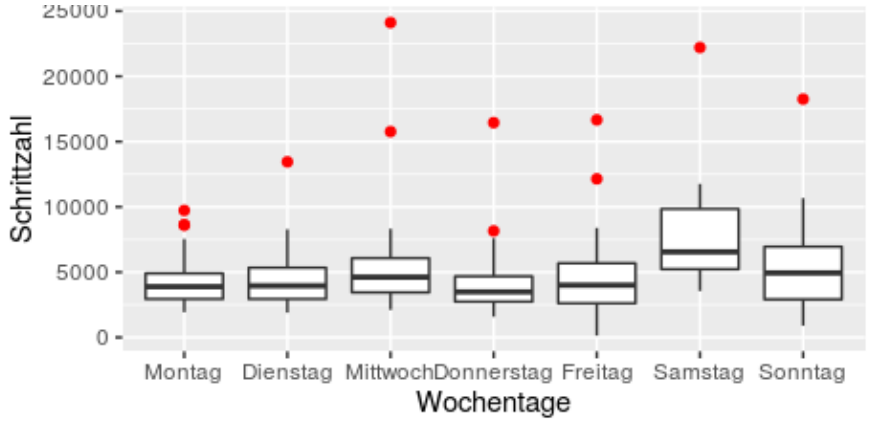
- Vorteil Zeitraumbetrachtung
- Verschiedene Hersteller
- Verschiedene Produkte
- Unterschiedliche Sensorik
  - Herzfrequenz
  - Fitness Tracker (Schritte, Kalorien, …)
  - Schlaf-Tracking
  - Temperatur
- Genauigkeit abhängig von Qualität des Sensors, Kalibirierung und Trageeigenschaften



Personengebundene Analysen

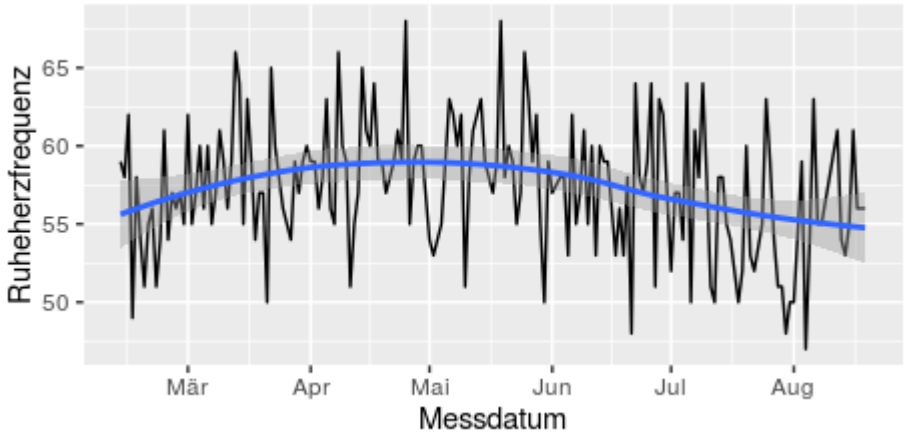


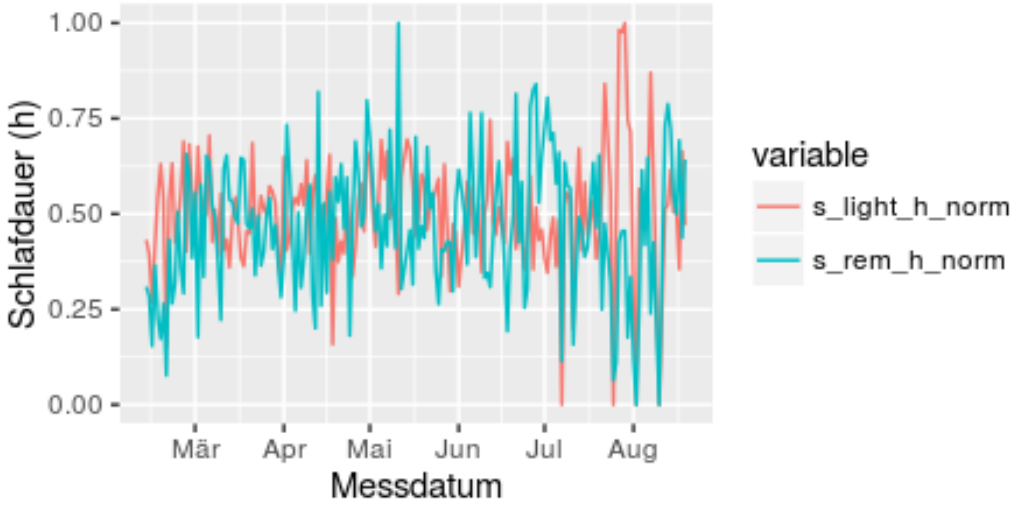




<u>Dimensionen</u>: Personen, Zeit, Variable

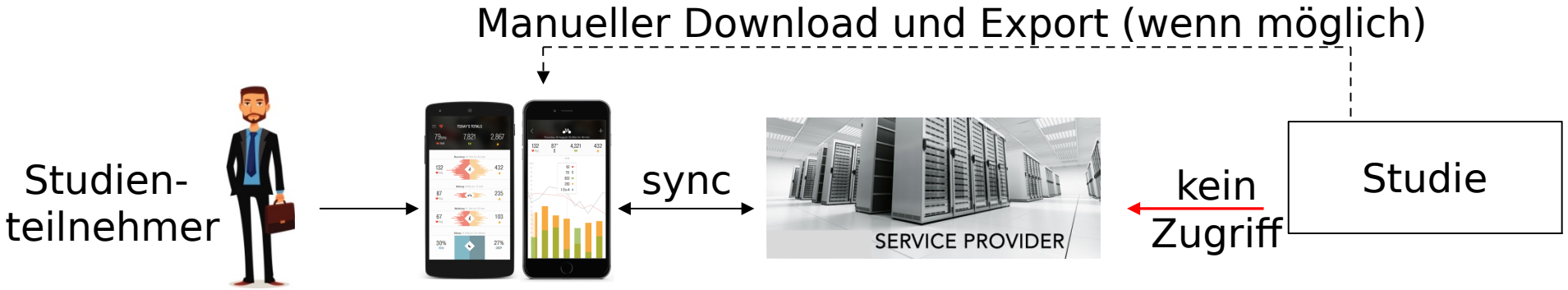
alle Grafiken für n(Person) = 1,
n(Zeit) = 192





Monitoring: Datenzugriff

- Unterschiedliche Daten
  - Allgemeine: Temperatur, Luftfeuchte, …
  - Personenspez.: Schrittzahl, Herzfrequenz, …
- Smartphones: Apps synchronisieren Daten mit Service Provider oder Hersteller
  - Privater Zugang (account)
  - Begrenzter Zugang und Datenaustausch



Perspektive:Citizen Science

- Idee: Einbezug der interessierten Öffentlichkeit bei der Aufnahme und Analyse von Daten zu helfen
  - Viele Vorgängerprojekte ...
  - Kartierung von Pflanzenvorkommen (z.B. Ambrosia), „Food shops“, …
  - Mehrwert für Bürger bislang schwer transportierbar

Herausforderung Datenschutz

- Kernidee: Schutz der Privatsphäre 
- Aber oftmals Behinderung vieler Forschungsaktivitäten trotz Möglichkeiten der
  - Anonymisierung (bis hin zur k-Anonymisierung)
  - (multiplen) Pseudonymisierung
- Folgen
  - Erschwerung des Datenaustausches (Datenzugang!)
  - Kaum Freigabe von Forschungsdaten
  - Doppelfinanzierung von Forschungsvorhaben (mit denselben / leicht unterschiedlichen Themen / Schwerpunkten

Herausforderung Metadaten

- Beschreibung der Daten (Items)
  - Name, Beschreibung / Definition, Wertebereiche / Kategorien von Codelisten, Skalen, ...
- Zusammenfassung in einem separaten MDR (Metadaten Repository)
  - Zugriff und Nutzung von verschiedenen Tools
- Herausforderungen
  - Sammlung und Codierung
  - Harmonisierung (interner Abgleich verschiedener Formulare, Versionen, Varianten, Eingabesysteme, ...)
  - Linking (Abstraktion) zu verfügbaren Terminologien / Ontologien

Metadaten Repository in LIFE

- Grundlage für
  - automatisch gesteuerte ETL Prozesse in die Forschungsdatenbank
  - Generierung von Datenbeschreibungsblättern (aCRF)
  - Erstellung automatisch generierter Reports
  - Generierung von SQL-Statements für Datenabfragen und Datenexporte
- Harmonisierung (Matching & Merging): Automatische Ableitung des DBS-Schema der Forschungsdatenbank
- Folgen
  - Vereinfachung und Automatisierung von Aufgaben und Prozessen

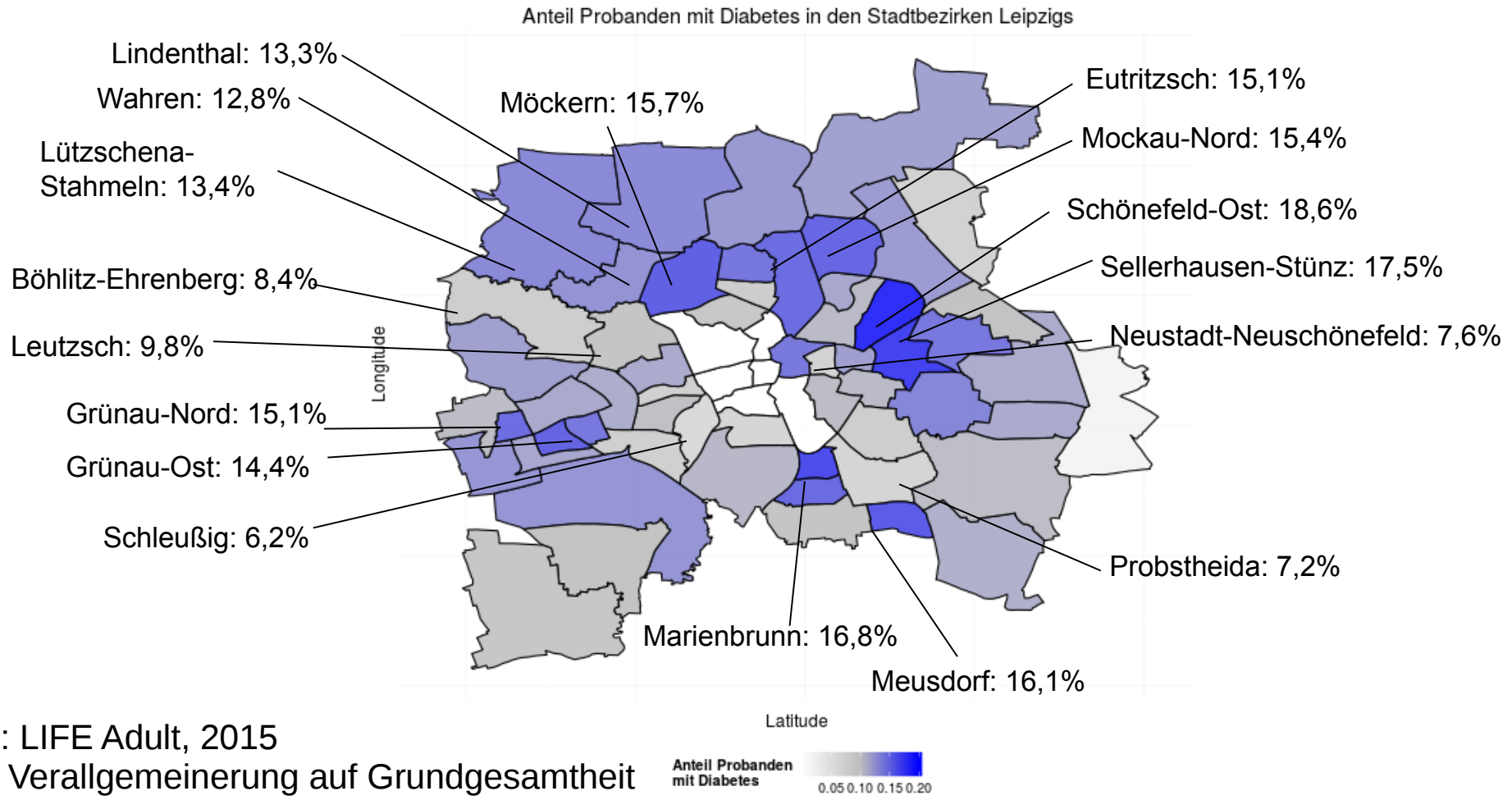
T. Kirsten et al.: On Integrating a large Corpus of Input Forms – Experiences from the LIFE Study. In preparation.

Leipzig Health Atlas (LHA)

- Projekt Uni Leipzig (Loeffler, Binder, Kirsten)
- Beginn: 03.2016
- Ziel: Aufbau einer Plattform, mit der Daten, Metadaten und Modelle (Ergebnisse) aus vergangenen klinischen und epidemiologischen Studien zur Verfügung gestellt werden
  - Novum in D (im Gegensatz zu USA)
  - Datensätze je Publikation
  - Vielzahl von Applikationen (Modelle)
- Metadatenmanagement + ausgewählte Apps (TK)

Geo-bezogene Auswertungen

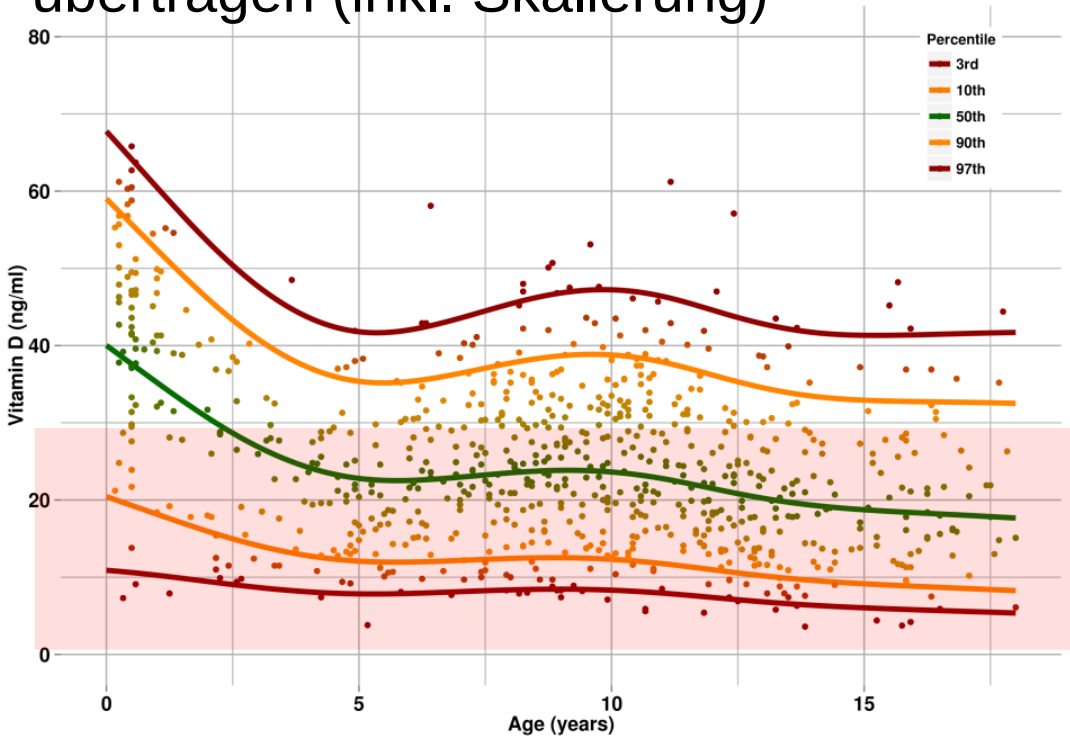
Diabetes nach Stadtbezirken



Basis: LIFE Adult, 2015
keine Verallgemeinerung auf Grundgesamtheit

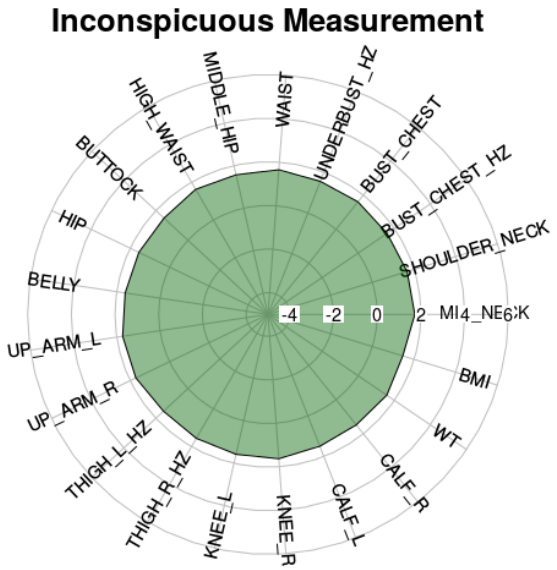
Referenzbereiche

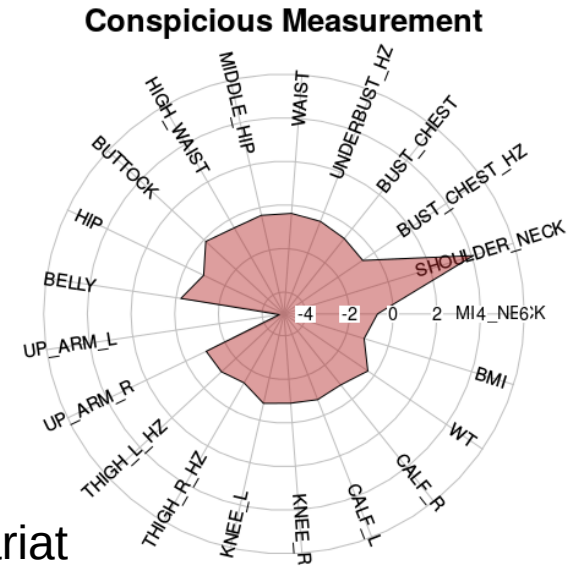
- Notwendig zur Einschätzung von Werten
- Häufigster Anwendungsfall: Labordaten
- Bisher kaum für andere Phänotypdaten verfügbar
- Oftmals auf kindliche Altersbereiche übertragen (inkl. Skalierung)



Grenzwert Vitamin D Mangel

univariat





multivariat

Zusammenfassung

- Datenintegration in der Medizin ist ein MUSS
  - Data sharing
  - Erkenntnisgewinn in der Medizin befördern
  - Teure Doppelförderung vermeiden
- Viele unterschiedliche Herausforderungen
  - Text Mining (entity recognition) deutscher Texte
  - Kein kostenloser Zugriff auf „DE“ Terminologien / Ontologien
  - Verwendung von Wearables → Monitoring
  - Citizen Science
  - Metadaten Management (Harmonisierung & Linking)
  - Datenschutz

Sommermeeting, 15./16.9.2016

http://www.gender.medizin.uni-leipzig.de