Programmierkonzepte in der Physikalischen Chemie

28. Datenverarbeitung und -Analyse in der PC

UNI FREIBURG

Albert-Ludwigs-Universität Freiburg

Dr. Till Biskup

Institut für Physikalische Chemie
Albert-Ludwigs-Universität Freiburg
Wintersemester 2017/18

Zentrale Aspekte

- Daten sind die Grundlage der empirischen Wissenschaften. Sie sollten Jahrzehnte überdauern.
- Datenverarbeitung sollte der Wissenschaftlichkeit, insbesondere der Nachvollziehbarkeit, genügen.
- Datenverarbeitung sollte systematisch erfolgen und jeder einzelne Schritt dokumentiert werden.
- Ein System zur Datenverarbeitung muss einfach nutzbar sein und klare Vorteile bieten, um genutzt zu werden.
- Ein System zur Datenverarbeitung ist viel umfassender als einzelne Programme zur Datenanalyse.

Standortbestimmung

UNI FREIBURG

Grobgliederung der Vorlesung „Programmierkonzepte“

1. Motivation
2. Infrastruktur
3. Code
4. Architektur
5. **Datenauswertung in der PC**

- Bislang ging es um Programmieraspekte und damit hauptsächlich um technische Voraussetzungen.
- Jetzt steht die Anwendung im Rahmen der wissenschaftlichen Datenauswertung im Fokus.

Übersicht

UNI FREIBURG

Datenverarbeitung, die wissenschaftlichen Kriterien genügt

Gründe für ein System zur Datenverarbeitung

Aspekte eines Systems zur Datenverarbeitung

Ansprüche an ein System zur Datenverarbeitung

Was ist Wissenschaft?

Eine Wiederholung – und (immer noch) keine formale Antwort

UNI FREIBURG

> *If I have seen further*
> *it is by standing on y^e shoulders of giants.*
>
> – Sir Isaac Newton

Was ist der Kern von Wissenschaft?

- Erkenntnisgewinn
- Unabhängigkeit vom Betrachter/Experimentator
- gegründet auf den Erkenntnissen früherer Generationen
- überprüfbar, nachvollziehbar, ggf. reproduzierbar

☛ Wissenschaftler tragen Verantwortung gegenüber denen, die auf den gewonnenen Erkenntnissen aufbauen.

Sir Isaac Newton: Brief an Robert Hooke, 5. Februar 1676

Die Verantwortung der Wissenschaftler

Saubere, nachvollziehbare Arbeit nach bestem Wissen und Gewissen

UNI FREIBURG

Empirische Wissenschaften

- Interpretationen ändern sich, Daten sollten Bestand haben.
- Voraussetzung: Daten nach bestem Wissen und Gewissen akkurat aufgenommen (und dokumentiert)

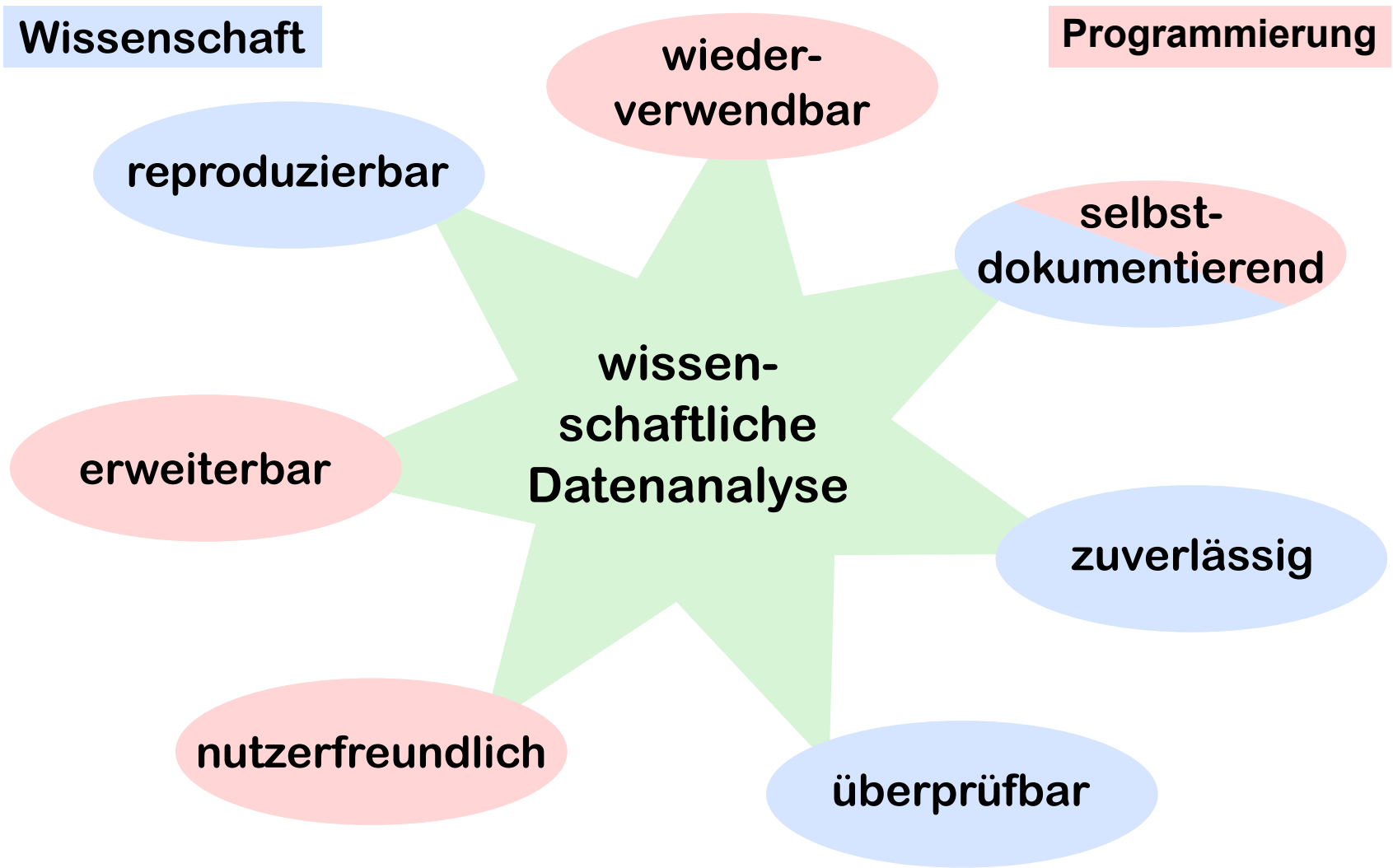
☛ Nachvollziehbarkeit und saubere Dokumentation sind wesentliche Aspekte von Wissenschaftlichkeit.

These

Datenauswertung, die nicht vollständig dokumentiert und nachvollziehbar ist, ist letztlich unwissenschaftlich.

Wissenschaftliche Datenanalyse

Anforderungen aus wissenschaftlicher und Programmierersicht



Anforderungen aus wissenschaftlicher Sicht

- selbstdokumentierend
  - Alle Parameter sollten automatisch mitgeschrieben werden.
  - Voraussetzung für vollständige Dokumentation
- zuverlässig
  - Auswertungen sind die Grundlage des Erkenntnisgewinns.
  - Korrektheit lässt sich u.a. durch (Unit-)Tests sicherstellen.
- überprüfbar
  - Wissenschaft lebt vom Vielaugenprinzip.
  - Unabhängige Überprüfbarkeit ist essentiell.
- reproduzierbar
  - Auswertungen sollten (unabhängig) wiederholbar sein.
  - setzt Versionsverwaltung und Versionsnummern voraus

Anforderungen aus Programmierersicht

- wiederverwendbar
  - Ökonomie: nicht das Rad immer wieder neu erfinden
  - setzt saubere Dokumentation und lesbaren Code voraus
- selbstdokumentierend
  - Code sollte so ausdrucksstark wie möglich sein.
  - setzt treffende und gut gewählte Abstraktionen voraus
- nutzerfreundlich
  - Voraussetzung für die Nutzung eines Systems
  - Robustheit, Nähe zum Nutzer, Flexibilität
- erweiterbar
  - Anforderungen entwickeln sich mit dem Verständnis.
  - Neue Funktionalität sollte einfach implementierbar sein.

Domain Driven Design

Ein gutes Modell der komplexen Realität und Fragestellung bildet die Grundlage für den Kern einer Anwendung.

- Modell der wissenschaftlichen Datenverarbeitung
  - Abstraktion der Erfahrung aus dem Forscheralltag
- Anwendung um dieses Modell herum entwickeln
  - Wechselspiel von Implementierung und Modellentwicklung
- saubere Architektur
  - sorgt für die Entkopplung der einzelnen Teile
  - ermöglichst Fokussierung und Beherrschbarkeit
  - Fokus von Anfang an auf guten Schnittstellen

Gründe für ein System zur Datenverarbeitung

Der Wert von Daten in den empirischen Wissenschaften

UNI FREIBURG

These

Wissenschaftler haben Verantwortung sowohl gegenüber der Wissenschaft als auch gegenüber dem Steuerzahler.

- Daten sind die Grundlage empirischer Wissenschaften.
  - Ausgangspunkt jeglichen Erkenntnisgewinns
  - unabhängig ihrer Herkunft (Messungen, Rechnungen)
  - Datenaufnahme immer nach bestem Wissen und Gewissen
  - Daten ohne zugehörige Informationen sind wertlos.
- Gewinnung von Daten ist in der Regel teuer.
  - unabhängig ihrer Herkunft (Messungen, Rechnungen)
  - Probenherstellung, Aufbauten, Messungen oft kostspielig
  - Rechnungen oft mit erheblichem Zeitbedarf verbunden
  - Personalaufwand häufig auch nicht vernachlässigbar

Voraussetzung für Wissenschaftlichkeit

System zur vollständigen Dokumentation und Nachvollziehbarkeit

UNI FREIBURG

These

Ein System zur Datenauswertung ist viel umfassender als einzelne Programme zur Datenanalyse.

- Kriterien für Wissenschaftlichkeit
  - vollständige Dokumentation
  - Zuverlässigkeit
  - Überprüfbarkeit
  - Nachvollziehbarkeit
- lässt sich nur durch systematisches Vorgehen erreichen
  - Frage der persönlichen Einstellung und Arbeitsweise
  - nicht (zwingend) in einer Software-Anwendung abbildbar
  - beginnt mit der Datenaufnahme und endet bei der Reproduzierbarkeit von Abbildungen aus den Rohdaten

Ansprüche an ein System zur Datenverarbeitung

Antworten auf fünf grundlegende Fragen

UNI FREIBURG

Fünf grundlegende Fragen

Wer hat **was** mit **wem** **wann** und **wie** gemacht?

- Modularität auf allen Ebenen
  - selten eine in sich geschlossene Software-Plattform
- Nachvollziehbarkeit und Selbstdokumentation
  - Rückgriff auf die Rohdaten und Verarbeitungsschritte von einer Abbildung/Datenrepräsentation aus
- Wartbarkeit und Erweiterbarkeit
  - sollte durch andere Personen später weiter nutzbar sein

☛ Die konkrete Ausgestaltung des Systems ist flexibel – solange die gestellten Anforderungen erfüllt werden.

Aspekte eines Systems zur Datenverarbeitung

Von der Datenaufnahme zur Publikation

UNI FREIBURG

- Datenverarbeitung beginnt mit der Datenaufnahme.
  - Daten ohne Zusatzinformationen (Metadaten) sind wertlos.
  - möglichst vollständige Dokumentation der Bedingungen
- Rohdaten sollten sauber und dauerhaft archiviert werden.
  - (*von anderen*) nachvollziehbares Ablagesystem
  - standardisierte, dokumentierte Formate
  - Zielstellung: Archivierung über Jahrzehnte
- Datenverarbeitung erfordert ein systematisches Vorgehen.
  - dokumentiert, nachvollziehbar, überprüfbar, reproduzierbar
  - Reproduzierbarkeit erfordert hinreichende Dokumentation und Archivierung von Daten und verwendeten Routinen.
- Repräsentationen sollten automatisch reproduzierbar sein.
  - Eineindeutigkeit in beide Richtungen gewährleisten.

Bausteine eines Systems zur Datenverarbeitung

Tipps aus der Praxis und der eigenen Erfahrung

UNI FREIBURG

- eindeutige Kennzeichnung aller Proben
  - einfachste Variante: durchnummerieren
  - in der Synthese guter Standard (meist mit Namenskürzel)
- vollständiger Satz von Metadaten zu einer Messung
  - vollständige Beschreibung der Messapparatur
  - *formalisierte* Dokumentation aller wichtigen Messparameter
- vollständiger Parametersatz jedes Verarbeitungsschritts
  - muss (automatische) Reproduzierbarkeit gewährleisten
  - *vollständige* Historie zu jedem Datensatz
- systematische Ablage von Daten und Ergebnissen
  - Rohdaten, abgeleitete Daten, Metadaten, Repräsentationen
  - modulare, erweiterbare Verzeichnisstruktur
  - Konventionen für Dateinamen mit konsistenter Verwendung

Vorschau: nachfolgend behandelte Aspekte

Die Summe von über einem Jahrzehnt eigener Erfahrungen

UNI FREIBURG

- Datenformate: Beständigkeit und Plattformunabhängigkeit
  - Ziel: Archivierung und Lesbarkeit über Jahrzehnte
  - Beispiele für offene, (selbst-)dokumentierte Formate
- Datenaufnahme: Metadaten
  - Das verfügbare Wissen ist während der Messung maximal.
  - Ziel: formalisiertes, möglichst vollständiges Laborbuch
- Datenverarbeitung und -Analyse: selbstdokumentierend
  - vollständiger Parametersatz zu jedem Verarbeitungsschritt
  - System zur automatischen Generierung einer Historie
- Datenpräsentation: Abbildungs- und Berichterstellung
  - automatisierte Abbildungserstellung aus Datensätzen
  - Bericht: Dokumentation der Verarbeitung und Auswertung und Präsentation der Charakteristika eines Datensatzes

Ansprüche an ein System zur Datenverarbeitung

Nutzerfreundlichkeit und offensichtliche Vorteile der Nutzung

UNI FREIBURG

- Ansprüche stellen den Anwender in den Fokus.
  - Kriterien der Wissenschaftlichkeit wurden schon behandelt.

zwei Ansprüche aus Sicht der Anwender

- hinreichend einfach nutzbar
  - möglichst niedrige Eintrittsschwelle
  - gute Dokumentation inkl. praxisnaher Beispiele
  - intuitive Bedienung dank guter Abstraktionen
  - flexibel erweiterbar und auf Bedürfnisse anpassbar
- offensichtliche Vorteile bei Verwendung des Systems
  - anderweitig nur mit erheblichem Aufwand erreichbar
  - korreliert mit (externen, durchgesetzten) Anforderungen: je höher die Anforderungen, desto größer der Nutzen

Nutzerfreundlichkeit

So einfach wie möglich, so komplex und flexibel wie nötig

UNI FREIBURG

- zwei Arten von Komplexität (nach Fred Brooks)
  - unvermeidliche Komplexität (*essential complexity*)
  - vermeidbare Komplexität (*accidental complexity*)
- Wissenschaftliche Datenverarbeitung ist inhärent komplex.
  - Entsprechendes Hintergrundwissen ist eine Voraussetzung.
  - Viele Anforderungen sind unvorhersehbar.
  - Flexibilität ist wesentlich für die Nutzbarkeit.
- Das System ist für den Nutzer da – *nicht* umgekehrt.
  - Nutzerbedürfnisse möglichst genau analysieren
  - wichtig: treffende Abstraktionen (*domain-driven design*)
  - Prinzip der „geringsten Überraschung“

☛ Nur ein möglichst einfach nutzbares System, das offensichtliche Vorteile bietet, wird auch genutzt werden.

Frederick P. Brooks: The Mythical Man Month. Addison Wesley Longman, Boston 1995

Offensichtliche Vorteile der Nutzung

Kompensation für den (initialen) Mehraufwand

UNI FREIBURG

- Ein System, das alle Kriterien erfüllt, ist komplex.
  - lässt sich nicht „schnell mal nebenher“ entwickeln
  - Viele Einzelaspekte sind einfach umsetzbar.
  - Entscheidend ist das Bewusstsein für einzelne Aspekte und die Existenz einfacher, bewährter Lösungen.
- „Köder“: Automatisierung und Vereinfachung von Abläufen
  - appelliert an die menschliche Faulheit
  - Automatisierung sorgt für Konsistenz.
- Entscheidend sind die (extern) gestellten Anforderungen.
  - Nur wenn Wissenschaftlichkeit real eingefordert wird, ergibt sich ein Vorteil aus der Verwendung des Systems.

☛ Nur eine konsequente Nutzung liefert reale Vorteile. Entsprechend prominent sollten sie beworben werden.

Zentrale Aspekte

- Daten sind die Grundlage der empirischen Wissenschaften. Sie sollten Jahrzehnte überdauern.
- Datenverarbeitung sollte der Wissenschaftlichkeit, insbesondere der Nachvollziehbarkeit, genügen.
- Datenverarbeitung sollte systematisch erfolgen und jeder einzelne Schritt dokumentiert werden.
- Ein System zur Datenverarbeitung muss einfach nutzbar sein und klare Vorteile bieten, um genutzt zu werden.
- Ein System zur Datenverarbeitung ist viel umfassender als einzelne Programme zur Datenanalyse.