Praktikum Datamining und Sequenzanalyse

Einführung

Markus Fleischhauer – markus.fleischauer@uni-jena.de

Organisatorisches

Vorträge

Zu bestimmten Terminen im SR 3423

freies Arbeiten

Montag und Freitag im Linuxpool

Projekte

1. Exakte Suche

naive Suche, KMP, Boyer-Moore

2. Alignments

globales/lokales Alignment, Alignmentscore mit linearem Speicher

3. Clustering, phylogenetische Bäume

UPGMA, WPGMA , Neighbor-Joining

Projekte bauen aufeinander auf!

Gruppen und Projekte

- 3-5 Personen pro Gruppe
- 3-5 Gruppen
- jede Gruppe bekommt einen gemeinsamen
  - Projektbereich („Group“) im Gitlab.
  - Projekte IMMER in eurer „Group“ anlegen
  - Neues Projekt: `Settings → Repository → Protected Branches → Unprotect Master`
- Jedes Projekt bekommt ein eigenes git-Repository

`https://git.bio.informatik.uni-jena.de`

Gruppen und Projekte

In der heutigen Veranstaltung:

- Gitlab Account erstellen
- Unsername beliebig, Echten Namen angeben,
  Uni E-Mail-Adresse Verwenden
- Gruppen bilden
- Pro Gruppe eine E-Mail mit den Gitlab
  Benutzernamen der Gruppenmitglieder an:

  `markus.fleischauer@uni-jena.de`

`https://git.bio.informatik.uni-jena.de`

Ziele

- Teamarbeit
- Implementierung bekannter Algorithmen
- Auswirkung der Implementierung auf die Performance
- Arbeiten mit:
  - Versionskontrolle
  - Build-Management
  - IDE (integrated development environment)

Aufgaben

Ein Aufgabenblatt pro Projekt (online)

1. Programmieren
   - Wartbarer Code – Objektorientiert (OOP)
   - Effizienter Code
   - Command Line Interface (**CLI**)
   - Dokumentation
2. Evaluation (Protokoll)
3. Präsentation (Vorträge)
   - Siehe Leitfaden Projektpräsentation (online)

Tools für die Aufgaben

- Objektorientierte Programmierung mit **`Java`**
- Dokumentation mit **`Javadoc`**
- Arbeiten in einer **`Linux`** Umgebung
- Versionskontrolle mit **`git`**
- Projektmanagement mit **`Gradle`**
- Als IDE verwenden wir **`IntelliJ IDEA`**
- Benutzerinterface (CLI) mit **`picocli`**

Auswertung und Vortrag

Auswertung

- Aufgabenzettel bearbeiten (**online**)
- Laufzeitmessungen
- auf mögliche Fehler/Probleme eingehen
- Protokoll **vor** Vortrag abgeben! (pushen)

Vortrag (Leitfaden online)

- Jede Gruppe trägt einmal vor (Dauer 30min)
  - Vorstellung und Diskussion der Ergebnisse
  - Vorführen der Benutzerschnittstelle
  - Details zur Implementierung

Linux

Linux bringt Vielfalt

Distributionen

Mint Debian Ubuntu openSUSE

Fedora Mageia Manjaro CentOS

LXLE Arch elementary OS

...

Desktopumgebungen

KDE Gnome LXDE Xfce

MATE Unity Cinnamon Panteon

Verzeichnisstruktur

UNIX, Linux, ZETA

- / [Wurzelverzeichnis]
  - bin/
  - lib/
  - etc/
  - dev/
    - sda1 [Gerätedatei]
    - sda2 [Gerätedatei]
  - home/
    - benutzername/
    - anwendername/
      - Office Dokument.odt
      - Foto.png
      - Verzeichnis/
  - media/
    - disk1/ [Einhängpunkt für Datenträger]
    - usbdisk1/ [Einhängpunkt für Datenträger]
      - tabelle.ods
      - Bild.jpg
      - Text.txt
      - Unterordner/
  - usr/

Mac OSX

- / [Wurzelverzeichnis]
  - bin/ [versteckt]
  - System/
  - etc/ [versteckt]
  - dev/ [versteckt]
    - disk [Gerätedatei]
    - disk2 [Gerätedatei]
  - Users/
    - benutzername/
    - anwendername/
      - Office Dokument.odt
      - Foto.png
      - Verzeichnis/
  - Volumes/
    - macosx/ [Einhängpunkt für Datenträger]
    - usbmedium/ [Einhängpunkt für Datenträger]
      - tabelle.ods
      - Bild.jpg
      - Text.txt
      - Unterordner/
  - Applications/
  - usr/ [versteckt]

Windows NT/2000/XP

- C:\ [Laufwerk]
  - Programme\
  - Dokumente und Einstellungen\
    - Adminstrator\
    - Benutzername\
      - Office Dokument.odt
      - Foto.png
      - Verzeichnis\
  - Windows\
    - System32
- E:\ [Laufwerk]
- F:\ [Laufwerk, USB-Medium]
  - tabelle.ods
  - Bild.jpg
  - Text.txt
  - Unterordner\

Besonderheiten der Shell

Hilfe zu Befehlen

`info, man <Befehl>`	Hilfe zu einem Befehl
`<Befehl> --help`	kurze Hilfe

Completion

Name-Completion mit Tab-Taste

Befehle

`ls`	Anzeige des Verzeichnisinhalts
`cd`	Wechseln des Verzeichnisses
`mkdir`	Erzeugen eines Verzeichnisses
`rmdir`	Löschen eines leeren Verzeichnisses
`rm`	Löschen einer Datei
`rm -r`	Löscht rekursiv einen nicht leeren Ordners
`mv`	Verschieben
`cp`	Kopieren
`cat / less`	Anzeigen von Dateien
`grep`	Suchen in einer Datei
`head`	Anfang einer Datei anzeigen
`df -h`	Anzeigen der Festplattenbelegung

Benutzerrechte

- Jede Datei und jedes Verzeichnis ist einem Eigentümer und einer Gruppe zugeordnet.
- verschiedene Rechte für Eigentümer, Gruppe und andere
- Anzeigen der Rechte mit z.B. `ls -la`

`chmod`	Setzen der Dateirechte
`chown`	Ändern des Eigentümers
`chgrp`	Ändern der Gruppe
umask	Setzen der Standardrechte für neue Dateien

Java und OOP

Grundlagen

Klassen definieren Eigenschaften v. Objekten (Blaupausen)

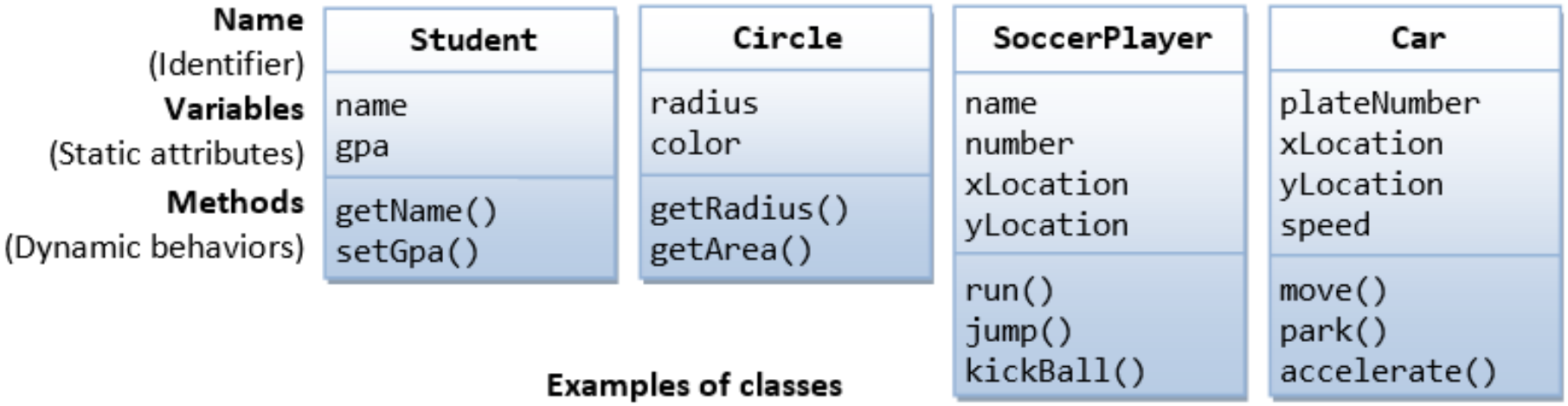
1. Name/Identifier
2. Attribute (Felder, Variablen)
3. Behaviour (Methoden)

Objekte sind Instanzen von Klassen

1. Felder – Belegung
2. Methoden – Abhängig v. Belegung

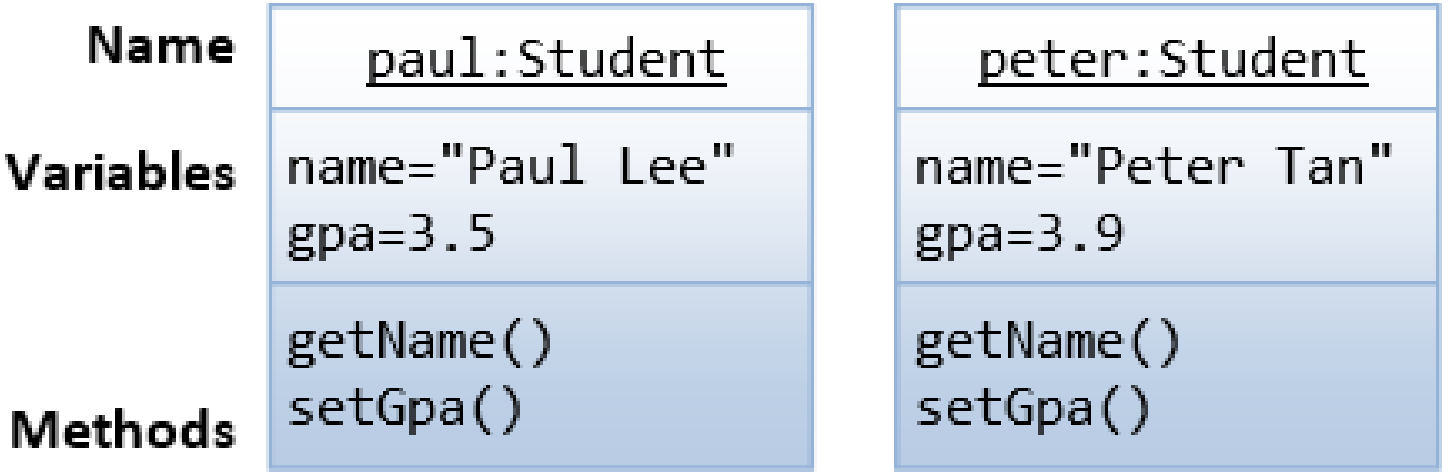
Repräsentation von physikalischen oder logischen Einheiten der Echtwelt

Klassen



Examples of classes

Objekte - Instanzen



Two instances - paul and peter - of the class Student

Sichtbarkeit

Modifier	Class	Package	Subclass	World
public	✓	✓	✓	✓
protected	✓	✓	✓	✗
default	✓	✓	✗	✗
private	✓	✗	✗	✗

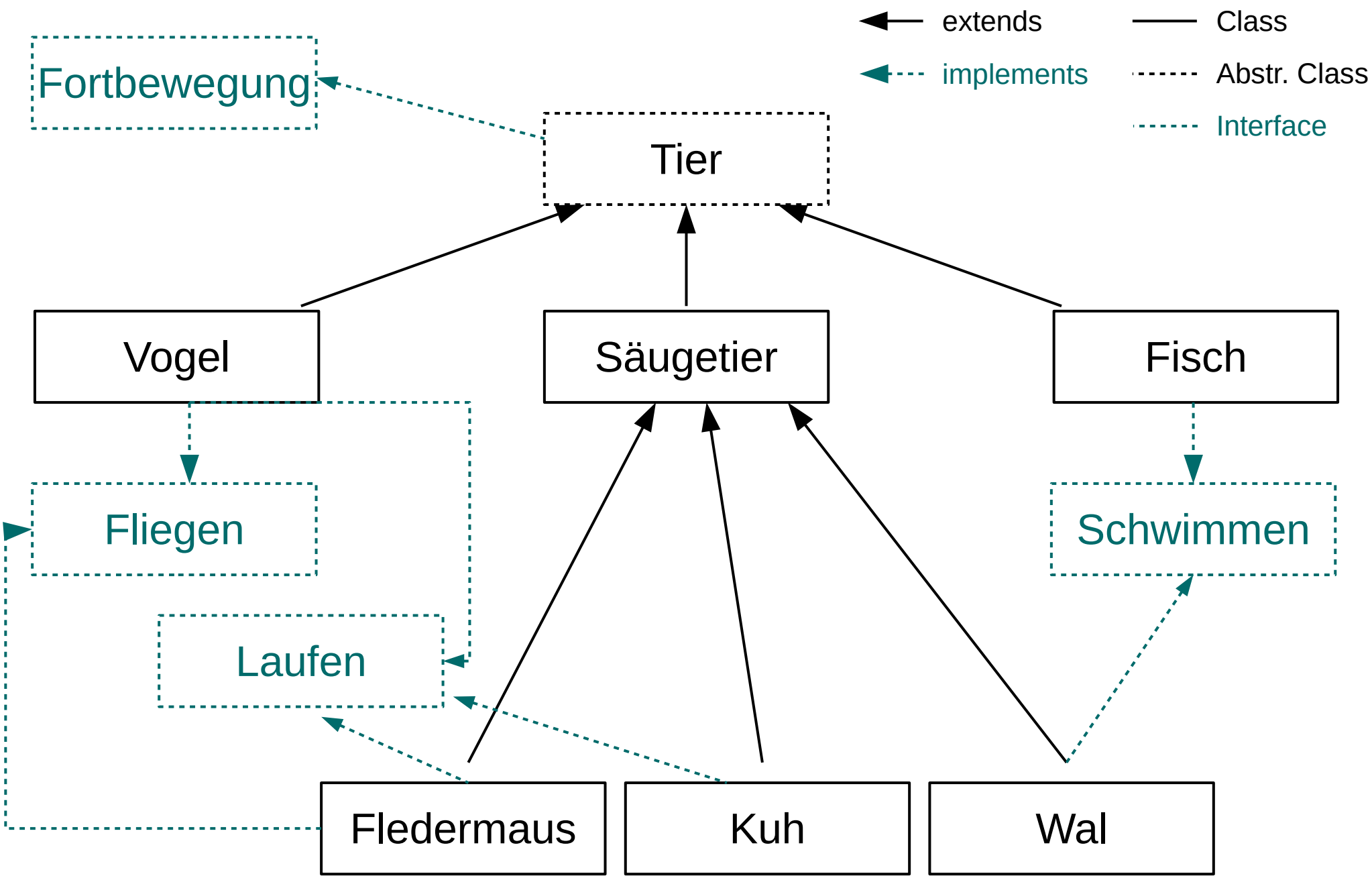
Klassentypen in Java

Klasse	• Instanzierbar • Nur aus-implementiert Methoden • Beliebige Felder und Sichtbarkeiten • Einfachvererbung
Abstrakte Klasse	• NICHT Instanzierbar • Abstrakte und implementierte Methoden • Beliebige Felder und Sichtbarkeiten • Einfachvererbung
Interface	• NICHT Instanzierbar • Abstrakte Methoden mit default implementierung • keine Felder, nur public Methoden • Mehrfachvererbung

Wann welche Klassentypen

Klasse	• Beliebige Objekte
Abstrakte Klasse	• Vermeidung von doppeltem Code • Basis Implementierung bei der essentielle Funktionalität fehlt
Interface	• Definition von Schnittstellen • Gewährleistung von Funktionalitäten ohne die konkrete Implementierung vorzugeben

Vererbung


extends
implements
Class
Abstr. Class
Interface
Fortbewegung
Tier
Vogel
Säugetier
Fisch
Fliegen
Schwimmen
Laufen
Fledermaus
Kuh
Wal

Interfaces?

```
public interface Distance{
    public double distance(Point p1, Point p2);
}

public class ManhattanMetrik() implements Distance{
    public double distance(Point p1, Point p2) {…};
}

public class EuclideanMetrik() implements Distance{
    public double distance(Point p1, Point p2) {…};
}


public class NearestNeighbor{
    public NNResult cluster(List<Point> points, Distance distance){
        …
    }
}
```

Mehrfachvererbung?

```
public abstract class Tier{
    public abstract void move();
}

public interface CanFly{
    public void fly();
}

public class Bird extends Tier implements CanFly{
    public void move(){
        //do something
    }
    public void fly(){
        //do something
    }
}
```

Polymorphismus

```
public class Clock{
   public static getFormat(){…}
   public void setTime(long ns){…}
   public void setTime(int h, int m, int s, int ms){...}
}

public class MoreSpecificClock extends Clock{
   public static getFormat(){…}
  @Override
   public void setTime(long ns){…}
}
```

Ueberladen

Ueberdecken

Ueberschreiben

Zugriff auf Attribute durch Getter/Setter

```
public class WithoutEncapsulation{

    public String value;

    public void printLength(){
        System.out.println("Länge: " + value.length());
    }

    public static void main(String[] args){
        WithoutEncapsulation we = new WithoutEncapsulation();
        we.value = null;
        we.printLength();
    }

}
```

NullPointerException

Zugriff auf Attribute durch Getter/Setter

```
public class WithEncapsulation{

    private String value;

    public void setValue(String value){
        if(value == null) this.value = “”;
        else this.value = value;

    public void printLength(){
        System.out.println(“Länge: “ + value.length());
    }

    public static void main(String[] args){
        WithoutEncapsulation we = new WithoutEncapsulation();
        we.setValue(null);
        we.printLength();          →    Länge: 0
    }

}
```

Hinweise

Zeit messen

via command line time

```
#$ time sleep 10
real 0m10.116s
user 0m0.001s
sys 0m0.007s
```

via Java

```
System.currentTimeMillis()
System.nanoTime()
long time = System.nanoTime();
//do something
long duration = System.nanoTime() - time;
```

Zeit messen

Vorgehen

- mehrfach messen
- Minimum aller Messungen verwenden

Testumgebung beschreiben

- Betriebssystem
- Hauptspeicher
- CPU
- Java VM version
- Java Heapspace (`-Xmx`)

IO

```
try(BufferedReader reader =
          new BufferedReader(new FileReader(path))){
   String temp = null;
   while((temp = reader.readLine()) != null){
      //werte aus
   }
}catch(IOException e){
   // ...
}
```

Fragen?

Jetzt bzw. Freitag gemeinsam in der Gruppe:

- Projekt „exactSearch“ in euer Group erstellen
  - Readme anlegen, Master unprotecten
- JDK, Gradle und IDE zum laufen bekommen
- Projekt in ein Verzeichnis auf eurem System clonen:

```
git clone
https://git.bio.informatik.uni-jena.de/datamining/ws2019/groupXX/exactSearch.git
```

- .gitignore file in Verzeichnis packen (online)

`https://git.bio.informatik.uni-jena.de`

Jetzt bzw. Freitag gemeinsam in der Gruppe:

- Mit IntelliJ ein neues Gradle project anlegen
  - Files → new → Project → Gradle → Select Java

```
ArtifactId 'exactSearch'
GroupId 'de.unijena.bioinf.dm.20[year].grp[number]'
Version '1.0-SNAPSHOT'
```

- Als Speicherort geclonetes git repo auswaehlen
- .gitignore, build.gradle, settings.gradle zu git hinzufuegen
- Aenderungen commiten und pushen

`https://git.bio.informatik.uni-jena.de`