Anwendung von Data Mining auf Daten der Durchflusszytometrie

- von Jan-Christoph Meier

- Hamburg, 09.01.2013

Ablauf

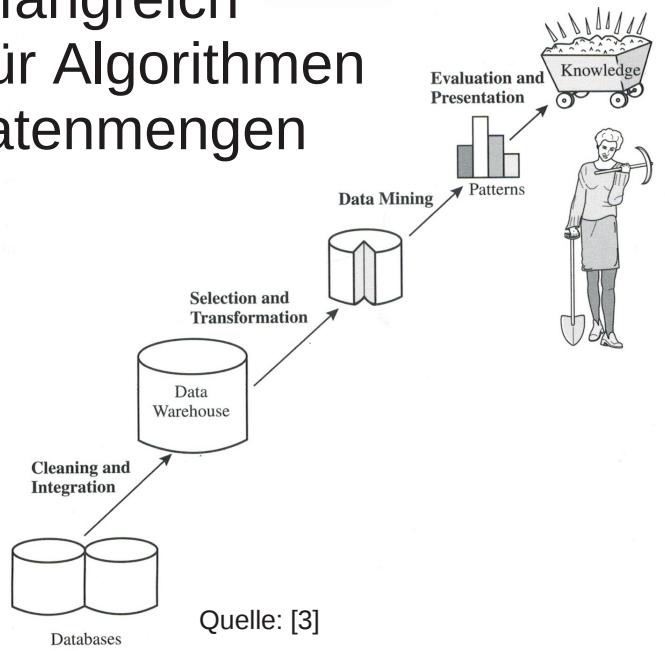
- Einführung
- Rückblick – Projekt 1
- Überblick Masterarbeit
- Aktuell – Projekt 2
- Chancen / Risiken
- Fazit

Ablauf

- Einführung 
- Rückblick – Projekt 1
- Überblick Masterarbeit
- Aktuell – Projekt 2
- Chancen / Risiken
- Fazit

Einführung

- In AW1, AW2 und Projekt 1 wurde sich mit verschiedenen Data Mining-Algorithmen für Clustering und „Frequent pattern mining“ beschäftigt.
- Die Thematik soll im Rahmen der Masterarbeit vertieft und die Algorithmen auf Daten aus der Multiple Sklerose-Forschung angewendet werden.
- Die hierbei zu analysierenden Daten sind sehr umfangreich (Größenordnung 1 Terabyte), daher müssen hierfür Algorithmen eingesetzt werden, die die Verarbeitung großer Datenmengen ermöglichen.



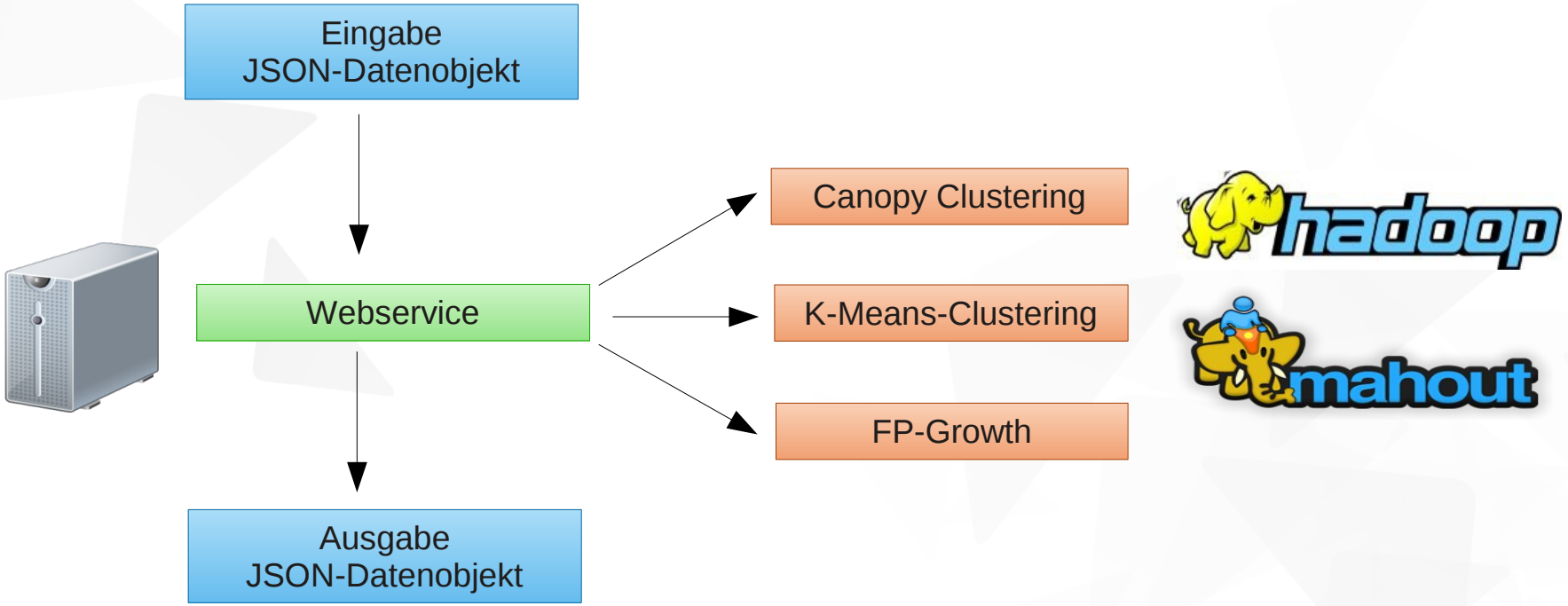
Quelle: [3]

Ablauf

- Einführung
- Rückblick – Projekt 1 
- Überblick Masterarbeit
- Aktuell – Projekt 2
- Chancen / Risiken
- Fazit

Rückblick Projekt 1

- In Projekt 1 wurde eine Webservice-Plattform entwickelt, die es ermöglicht, Daten mit verschiedenen Data Mining-Algorithmen zu analysieren.
- Als Basis hierfür dienten die Frameworks Hadoop und Mahout.

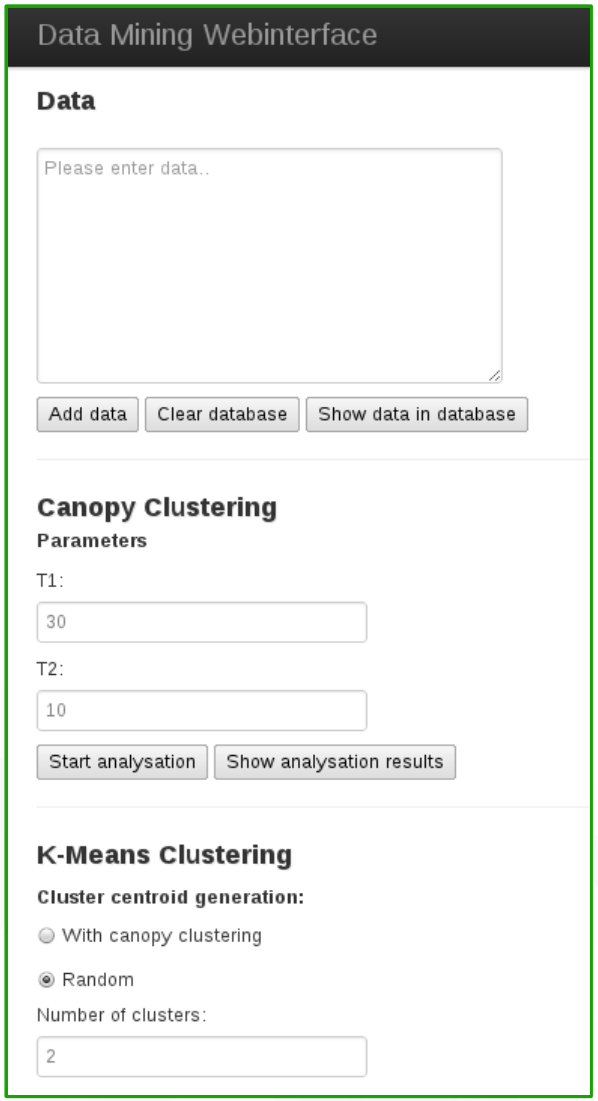


Rückblick Projekt 1

- Die Plattform wurde in eine Anwendung zur Analyse von Proteinsequenzen integriert.
- Hierdurch konnten Proteinsequenzen mit Data Mining untersucht werden, z.B. mit dem FP-Growth Algorithmus.

```
[A with L, V, ]: 17897
[I with L, ]:    17760
[A with L, S, ]: 17271
[L with S, V, ]: 16911
[A with G, L, ]: 16118
[...]
```

Abbildung 1: Aminosäuren, die häufig gemeinsam auftreten

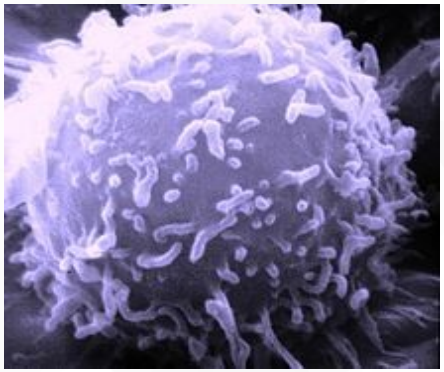


Ablauf

- Einführung
- Rückblick – Projekt 1
- Überblick Masterarbeit 
- Aktuell – Projekt 2
- Chancen / Risiken
- Fazit

Motivation

- Die Masterarbeit ist im Bereich der Multiple Sklerose Forschung angesiedelt. Ziel dieser ist es, die Ursache einer Multiple Sklerose Erkrankung zu erforschen und neue Therapien zu entwickeln.
- Bei einer Multiple Sklerose Erkrankung wird der eigene Körper durch das Immunsystem angegriffen.
- Auslöser hierfür sind die weißen Blutkörperchen (Leukozyten), die für die Immunabwehr zuständig sind.
- Die Leukozyten werden unterschieden in
  - Lymphozyten
  - Granulozyten
  - Monozyten



Warum werden die Zellen untersucht?

- Die Anzahl der Lymphozyten kann Auskunft über den aktuellen Zustand der Körperabwehr geben.
- Bei einer Entzündung oder Infektion steigt die Anzahl der Lymphozyten an.
- Bei folgenden Erkrankungen kommt es zu einer verringerten Anzahl an Lymphozyten:
  - AIDS
  - Verschiedene Krebsarten
  - Autoimmunerkrankungen (z.B. Multiple Sklerose)

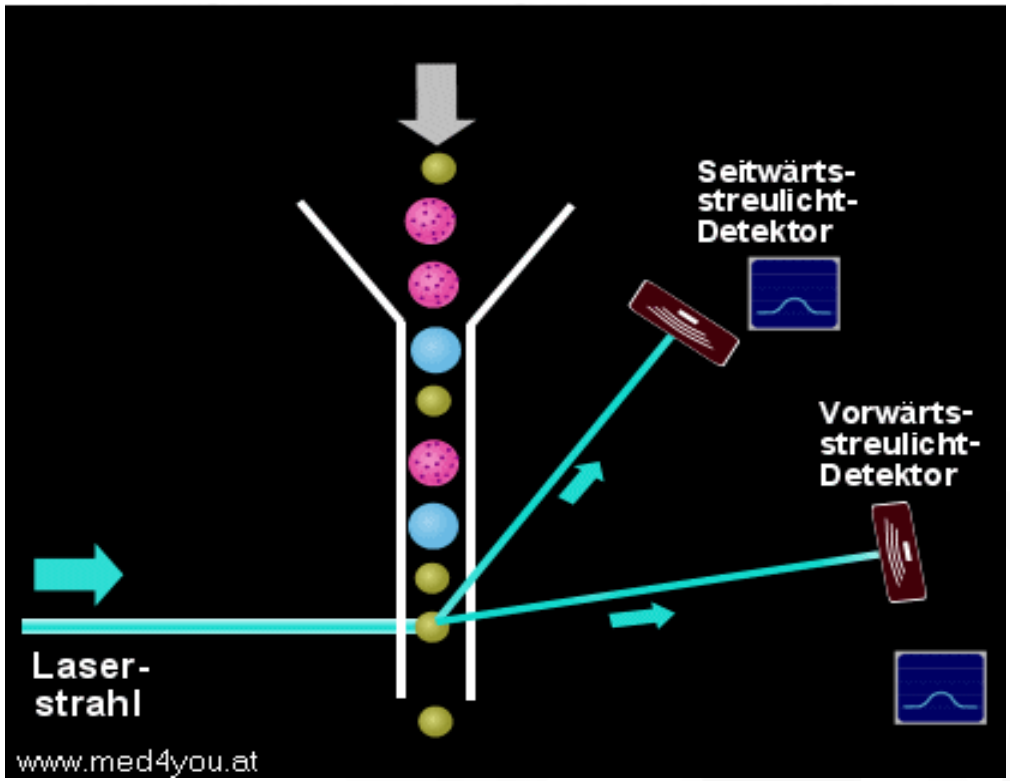
Was ist die Durchflusszytometrie?

- Bei der Durchflusszytometrie werden Zellpopulationen im Blut gemessen, indem dieses durch eine dünne Messkammer fließt und mit einem Laser beschossen wird.
- Die hierbei verwendeten Geräte werden als Durchflusszytometer oder auch „Fluorescence Activated Cell Sorting“-Gerät (FACS-Gerät) bezeichnet.
- Die Messergebnisse werden mit dem Computer erfasst und mit einer Software analysiert, z.B. FACSDiva oder Flowjo.

Quelle: [1]

Durchflusszytometrie im Detail

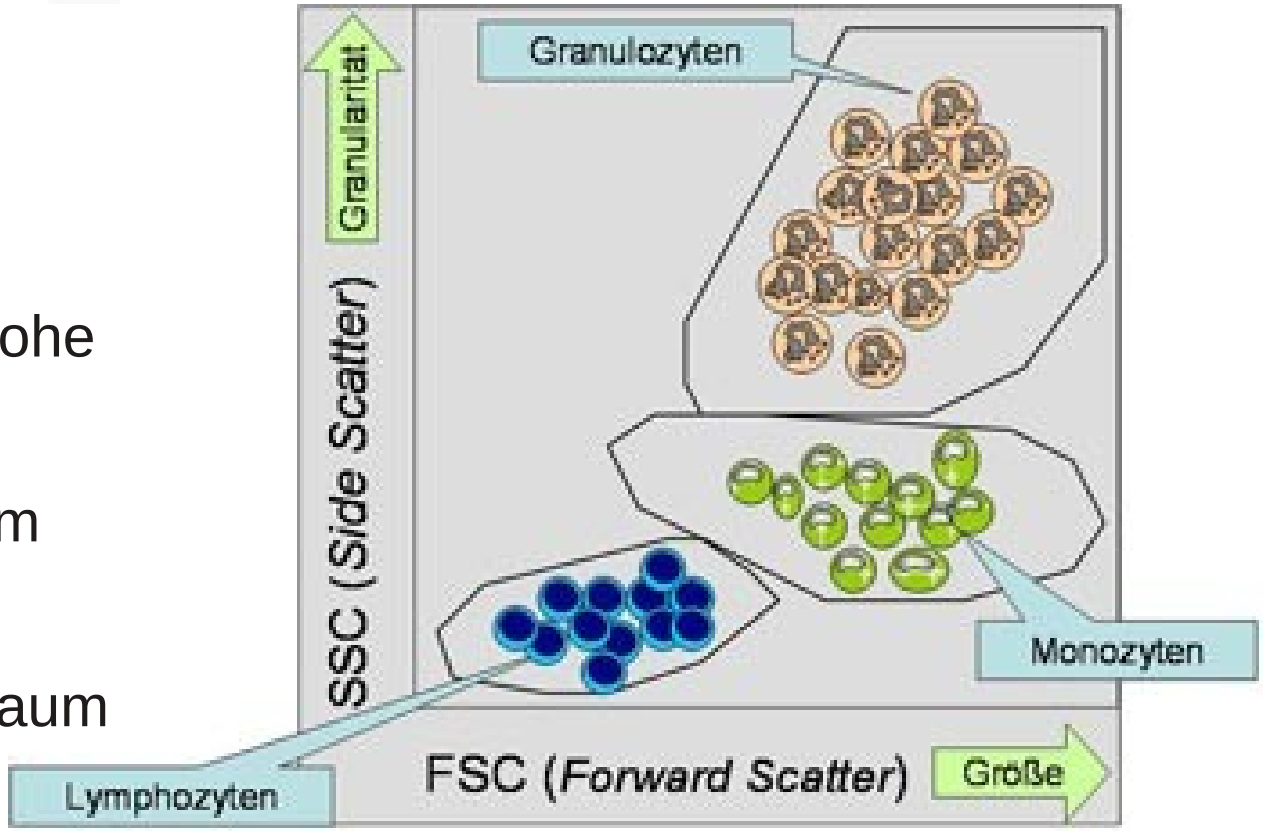
- Der Laser beschießt die Zellen und das Seitwärtsstreulicht sowie das Vorwärtsstreulicht werden eingefangen.
- Das Vorwärtsstreulicht gibt Auskunft über die Größe der Zelle.
- Das Seitwärtsstreulicht gibt Auskunft über die Körnigkeit der Zelle (Granularität).



Quelle: [1]

Messergebnis

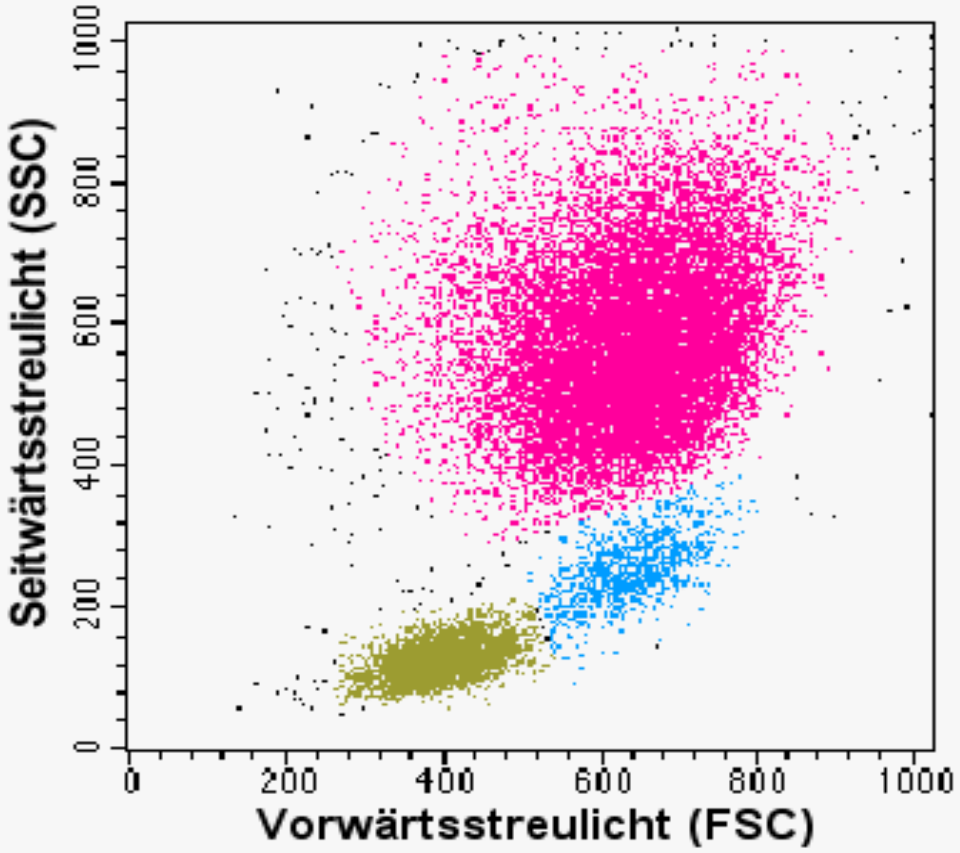
- Die einzelnen Ereignisse (Signale) des Lasers werden in einem X-Y-Diagramm eingetragen.
- X-Achse ist das Signal des Vorwärtsstreulicht.
- Y-Achse ist das Signal des Seitwärtsstreulicht.
- **Granulozyten**: Groß und hohe Granularität.
- **Monozyten**: Groß und kaum Granularität.
- **Lymphozyten**: Klein und kaum Granularität.



Quelle: [2]

Messergebnis im Detail

- Im Messergebnis sind teilweise aufgrund von Ungenauigkeiten der Sensoren Störungen vorhanden.



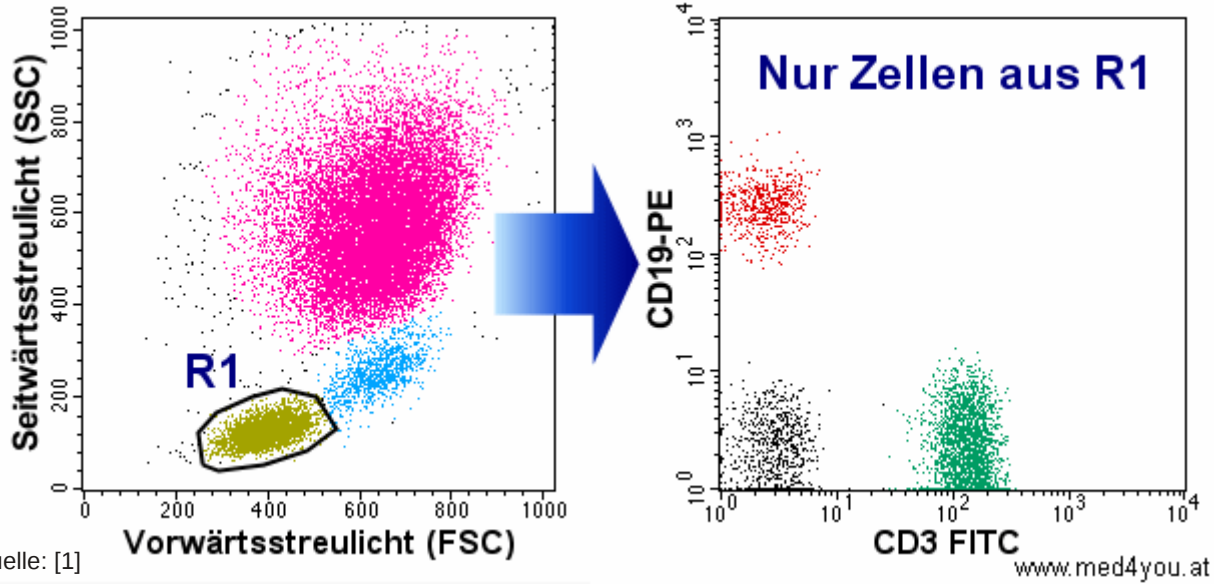
Quelle: [1]

Messen weiterer Zelltypen

- Die Lymphozyten können in B-Zellen, T-Zellen und NK-Zellen unterschieden werden.
- Damit diese mit dem Laser gemessen werden können, wird ein Antikörper zur Blutprobe hinzugefügt.
- Der Antikörper haftet an der Oberfläche der Zelle und erzeugt farbiges Laserlicht, sobald die Zelle mit dem Laser beschossen wird.
- Anhand der Farbe des Lasers kann dann zwischen den unterschiedlichen Zelltypen unterschieden werden.

Gaten

- Beim Gaten werden bestimmte Zellen selektiert, für die Antigene zur Blutprobe hinzugefügt wurden.
- Nach der Selektion werden die durch die Antigene hervorgerufenen Signale in einem weiteren Dot-Plot angezeigt.



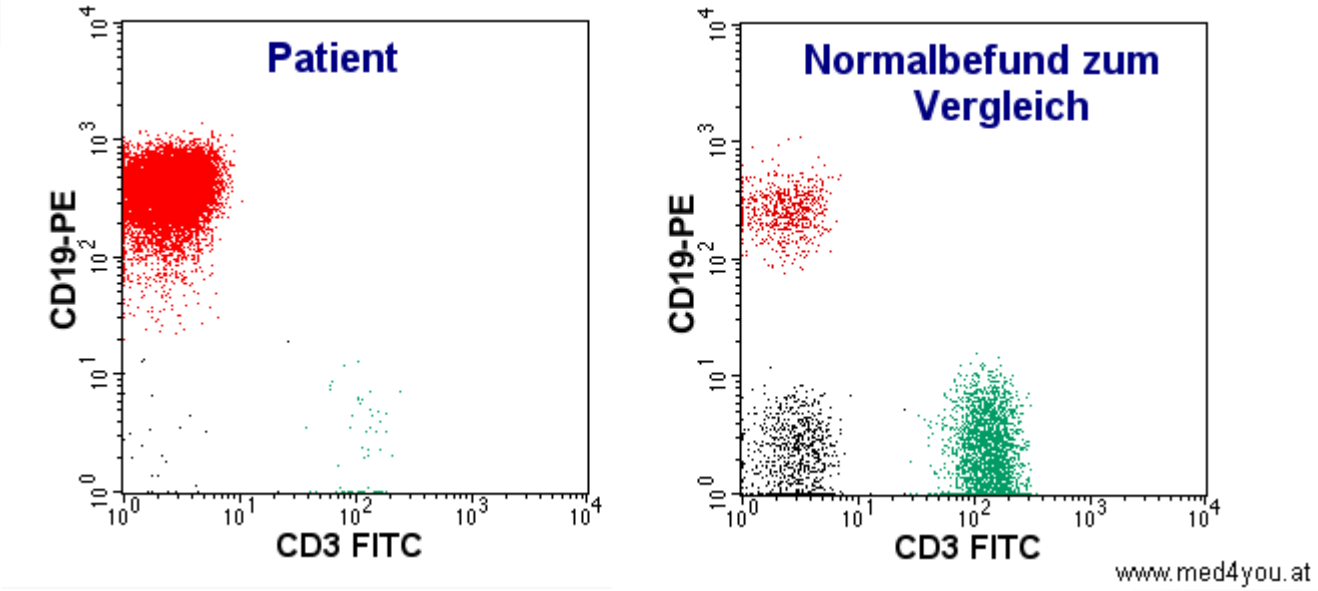
Quelle: [1]

Über die Selektion R1 wurden Lymphozyten ausgewählt.

CD19-PE = B-Zellen
CD3-FITC = T-Zellen

Ergebnis / Diagnose

- Aus den Populationen der verschiedenen Zelltypen kann eine Diagnose abgeleitet werden.
- Der kranke Patient hat ausschließlich B-Lymphozyten (rot dargestellt), im Gegensatz zum gesunden Patienten der überwiegend T-Lymphozyten hat (grün dargestellt).



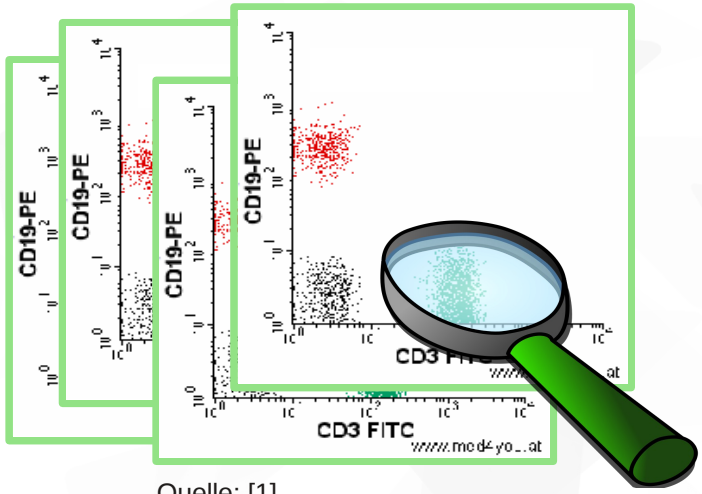
Quelle: [1]

Ist-Zustand

- FACS-Daten in der Größenordnung von insgesamt ca. 1 Terabyte vorhanden.
- Analysen werden von Hand mit der Software FacsDiva durchgeführt.
- Organisation der Daten eher willkürlich, jede Messung wird individuell im Dateisystem organisiert.
- Finden von Messergebnissen anhand bestimmter Kriterien kaum möglich.
- Im aktuellen Workflow ist wenig automatisiert.

Ziele für die Masterarbeit

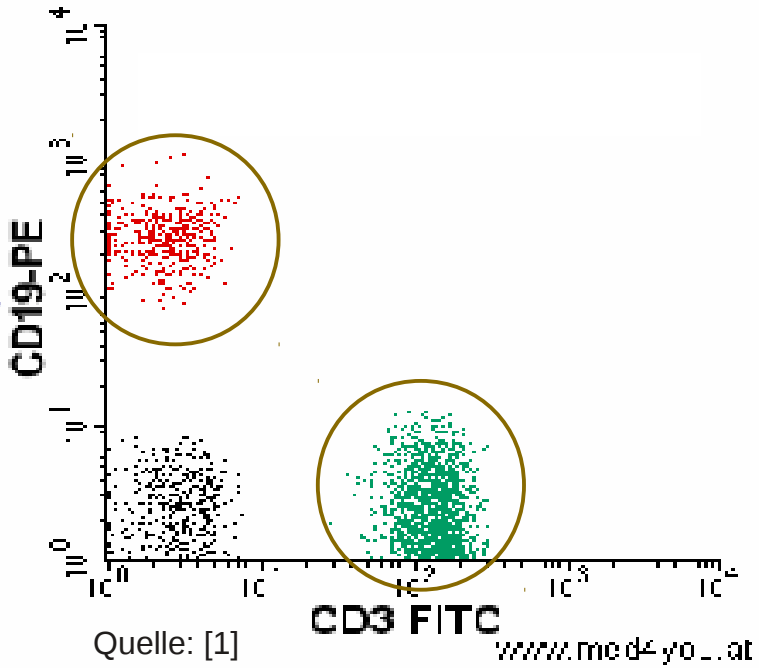
- Verbesserung der Organisation der FACS-Daten durch Entwicklung eines „Data management systems“.
- Messergebnisse sollen anhand verschiedener Kriterien im Datenbestand gefunden werden können.
- Automatisieren des Gating und Analyse der Zellpopulationen mithilfe von Data Mining-Algorithmen.



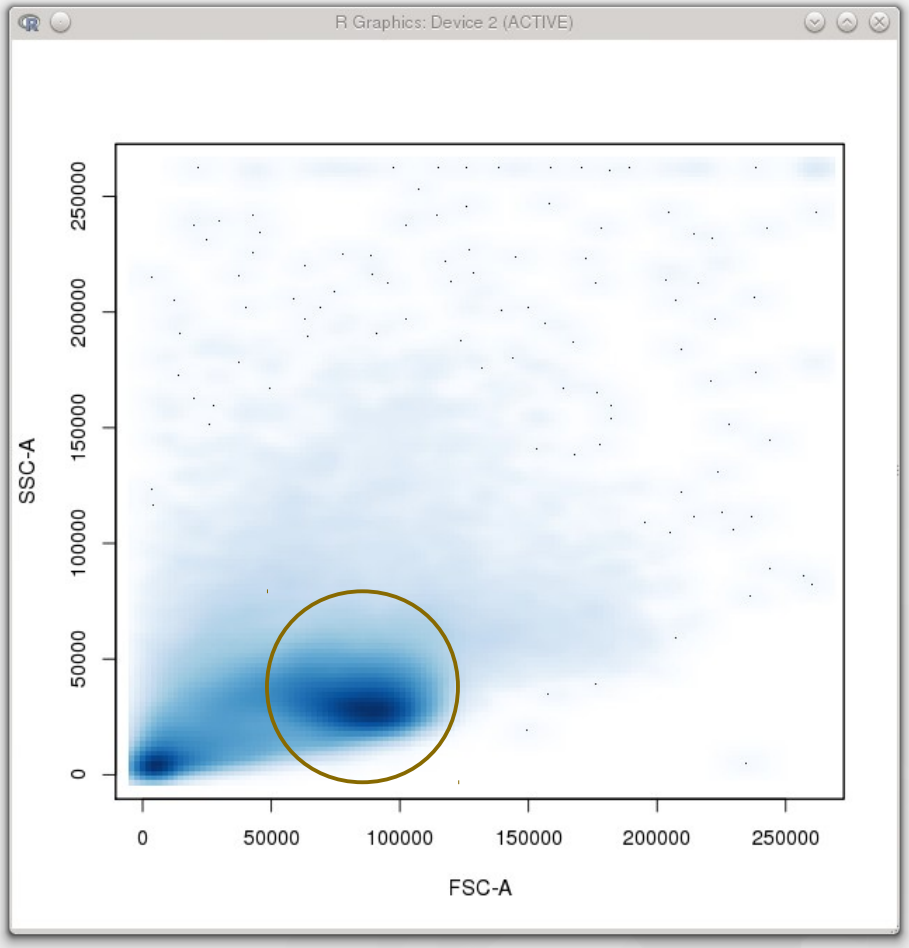
Quelle: [1]

Zellpopluationen finden

- Mit Clustering Algorithmen, wie z.B. K-Means werden die Zellpopulationen bestimmt.

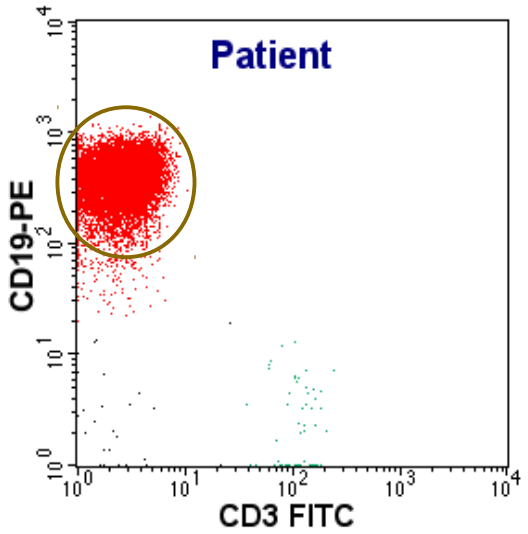


Quelle: [1]



Einbindung klinischer Daten

- Neben den FACS-Daten stehen noch weitere klinische Daten zum Krankheitsverlauf zur Verfügung, die mit in die Analyse eingebunden werden können.
- Es sollen Zusammenhänge zwischen den klinischen Daten und den Zellpopluationen im Blut gefunden werden.



Quelle: [1]

Korrelation

- Eingenommene Medikamente
- Aktueller Krankheitsstatus
- Symptome

Analyse mit Frequent Pattern Mining

- Mithilfe von „Frequent Pattern Mining“ kann ein Zusammenhang zwischen den Zellpopulationen, Krankheitsverläufen und eingenommenen Medikamenten gefunden werden.

Zellpopulation	Medikament	Krankheitszustand
Wenig Lymphozyten	Keine Medikamente eingenommen	Schlecht
Viele Lymphozyten	M1	Gut
Wenig Lymphozyten	M2	Schlecht
...	...	...

- Auf den ersten Blick könnte aus „Wenig Lymphozyten“ in Zusammenhang mit der Einnahme von Medikament „M2“ und einem schlechten Krankheitszustand abgeleitet werden, dass das Medikament nicht wirkt.
- Diese Aussage ist nicht repräsentativ, da nur wenige Datensätze betrachtet wurden.

Ablauf

- Einführung
- Rückblick – Projekt 1
- Überblick Masterarbeit
- Aktuell – Projekt 2 
- Chancen / Risiken
- Fazit

Projekt 2

- Aufbau eines „Data Management Systems“ (DMS) zur Verwaltung der Messdaten.
- Importieren der FACS-Rohdaten in das DMS.
- Automatische Durchführung des Gatings.
- Senden der Daten an den in Projekt 1 entwickelten Webservice für Data Mining.

Projekt 2 – Schematik der Software

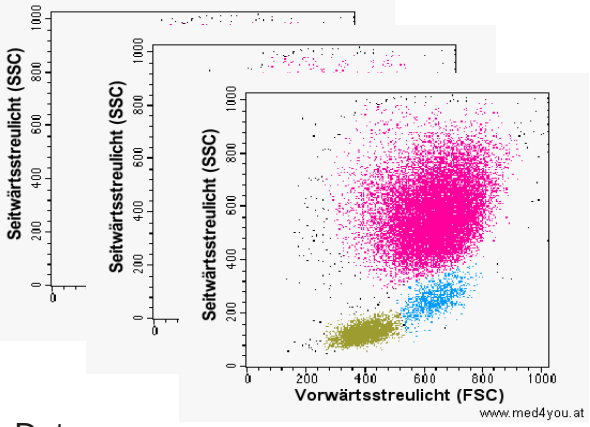
- 80122.fcs
- 53211.fcs
- 75532.fcs

FACS-Daten

Importieren der Daten

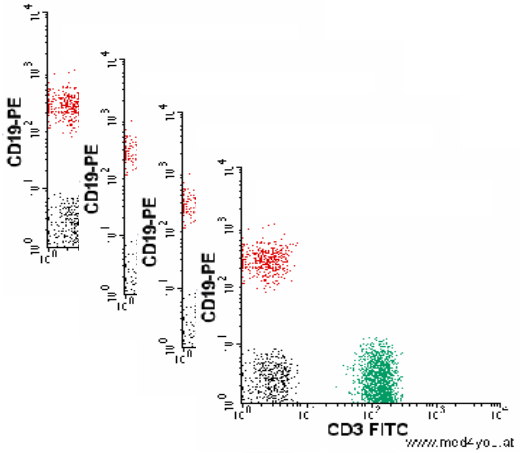
Data Management System

Selektion von Daten durch den Benutzer



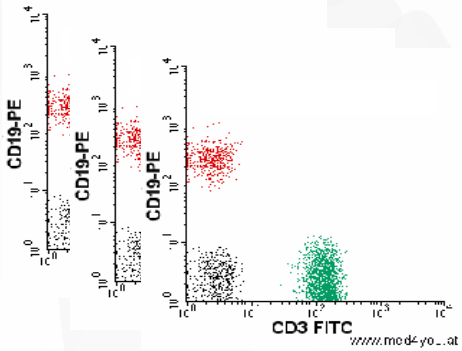
FACS-Daten

Gating durchführen



Zellpopulationen

Projekt 2 – Koppelung an Webservice



Zellpopulationen

Klinische Daten

- Eingenommene Medikamente
- Aktueller Krankheitsstatus
- Symptome

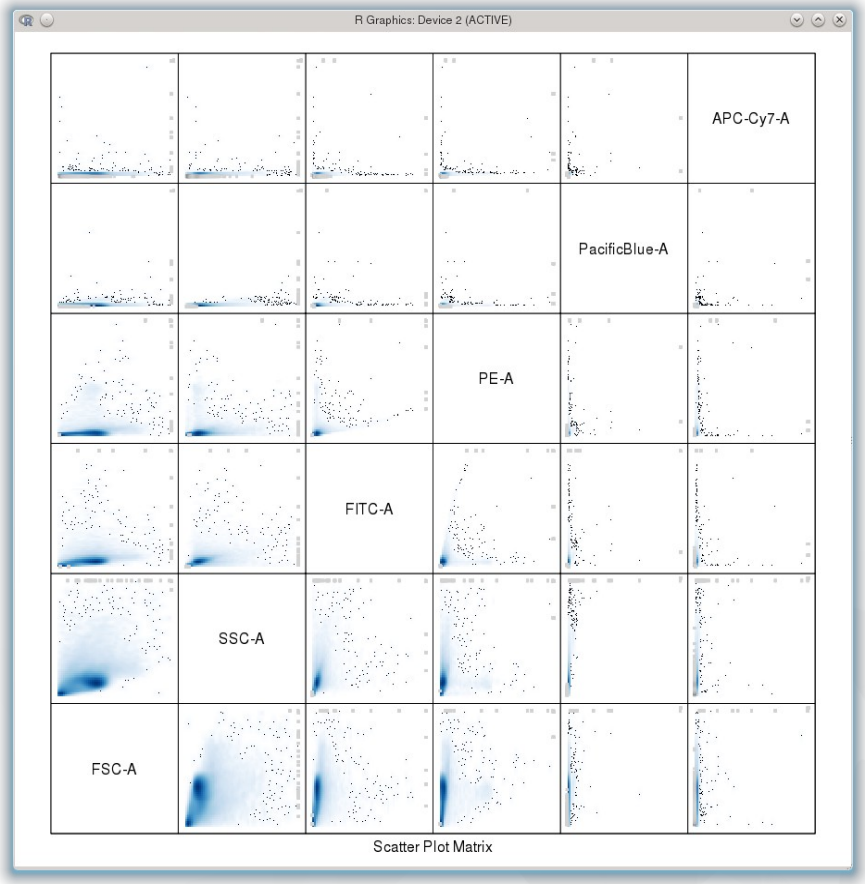
Webservice

FP-Growth

K-Means-Clustering

Analyse der Daten

- Es stehen verschiedene Programmiersprachen und Bibliotheken zum Auslesen der Facsdaten zur Verfügung.
- Die Programmiersprache R bietet mit der Bibliothek „flowCore“ umfangreiche Funktionen zum Auslesen von Facsdaten.
- Mit R kann direkt auf die einzelnen Ereignisse des Lasers zugegriffen werden oder die Daten in einem Dot-Plot ausgegeben werden.



Ablauf

- Einführung
- Rückblick – Projekt 1
- Überblick Masterarbeit
- Aktuell – Projekt 2
- Chancen / Risiken
- Fazit

Chancen

- Aufgrund der großen Menge an Daten konnten diese bisher nur schwer analysiert werden.
- Durch Anwendung von Data Mining können neue Erkentnisse aus den Daten gewonnen werden.
- Eine Analyse des gesamten Datenbestandes wurde noch nicht durchgeführt.
- Praktische Anwendung von Data Mining auf großen Datenmengen.

Risiken

- Als Ergebnis wird nur „weißes Rauschen“ geliefert, wodurch keine neuen Erkenntnisse gewonnen werden können.
- Die Daten sind sehr inkonsistent und können nur mit extrem viel Aufwand in einen konsistenten Stand gebracht werden.
- Die Selektion beim Gaten erfolgt häufig durch „scharfes Hinsehen“ und Erfahrung, was unter Umständen nur schwer automatisiert werden kann.

Ablauf

- Einführung
- Rückblick – Projekt 1
- Überblick Masterarbeit
- Aktuell – Projekt 2
- Chancen / Risiken
- Fazit 

Fazit

- Es sind sehr große Datenmengen vorhanden, die sowohl schlecht organisiert als auch nur schwer manuell analysiert werden können.
- Mit dem in der Masterarbeit zu entwickelndem System soll sowohl die Organisation sowie die Analyse der Messdaten wesentlich verbessert werden.
- Mithilfe von Data Mining soll neues Wissen aus den Daten und so neue Erkenntnisse für die Multiple Sklerose Forschung gewonnen werden.

Quellen

[1] http://www.med4you.at/laborbefunde/techniken/durchflusszytometrie/lbef_durchflusszytometrie.htm

[2] http://www.antikoerper-online.de/resources/17/607/Durchflusszytometrie+FACS+Messprinzip++Aufbau/

[3] **Data Mining, Concepts and Techniques**
Jiawei Han, Micheline Kamber, Jian Pei
Morgan Kaufmann 2011

[4] **Zellulare Diagnostik. Grundlagen, Methoden und klinische Anwendungen der Durchflusszytometrie**
U. Sack, A. Tarnok, G. Roth
Basel, Karger, 2007, pp 27–70

[5] **Automated high-dimensional flow cytometric data analysis**
Saumyadipta Pyne, Xinli Hu, Kui Wang, Elizabeth Rossin, Tsung-I Lin, Lisa Maier, Clare Baecher-Allan, Geoffrey McLachlan, Pablo Tamayo, David Hafler, Philip De Jager, and Jill Mesirov
Proceedings of the 14th Annual international conference on Research in Computational Molecular Biology

[6] **Flow: Statistics, visualization and informatics for flow cytometry**
Frelinger, Jacob and Kepler, Thomas and Chan, Cliburn
Source Code for Biology and Medicine 2008, 3:10

[7] **Scalable Analysis of Flow Cytometry Data using R/Bioconductor3**
David J. Klinke, Kathleen M. Brundage
Cytometry A. 2009 August; 75(8): 699–706.

Ende

Vielen Dank für die Aufmerksamkeit!

Fragen?