Tutorium: Reproduzierbare Forschung

15.03.2015, 12:00 – ca. 17:30 Uhr

Dozenten

Benjamin Hofner ist PostDoc am Institut für Medizininformatik, Biometrie und Epidemiologie der FAU Erlangen-Nürnberg. Er ist derzeit Reproducible Research Editor des Biometrical Journal und Autor/Co-Autor diverser R-Pakete (u.a. mboost, gamboostLSS, stabs, opm und papeR). Er beschäftigt sich auch in seiner täglichen Arbeit mit reproduzierbarer Forschung. Weitere Forschungsschwerpunkte sind statistische Modellbildung und Modellwahl, Modelle für biologische High-Throughput Experimente und die Entwicklung und Erweiterung moderner, modelbasierter Boostingansätze für biomedizinische Daten.

Lutz Edler leitete bis 2010 die Abteilung Biostatistik des Deutschen Krebsforschungszentrums (DKFZ) und ist derzeit einer der beiden Herausgeber des Biometrical Journal. Forschungsbereiche sind mathematische und statistische Modellbildung und Datenanalyse, Design von experimentellen und klinischen Studien und Methoden der Risikoanalyse zur Krebsentstehung und ihrer Anwendung zur Sicherheit von Nahrungs- und Futtermitteln in der EU.

Reproduzierbare Forschung…

… bezeichnet die Idee, dass das letztendliche Resultat der Wissenschaft die Publikation zusammen mit der kompletten Computerumgebung (d.h. Code, die Daten, Computerprogramme etc.) ist, welche benötigt wird um die Ergebnisse zu reproduzieren. Dies soll die Nachvollziehbarkeit und Überprüfbarkeit der Ergebnisse garantieren und zu neuen wissenschaftlichen Erkenntnissen führen.

Inhalt

Ziel des Tutoriums ist die Vermittlung der Einsicht in die Notwendigkeit reproduzierbarer Forschung in der biometrischen Forschung und ihren Anwendungen, sowie die Befähigung zur Nutzung reproduzierbarer Forschung als einen übergreifenden Ansatz moderner Datenanalyse, um somit die Reproduzierbarkeit der Ergebnisse zu einem ein essentieller Bestandteil der täglichen Arbeit zu machen.

Wir beginnen mit einer Einführung in das Konzept der reproduzierbaren Forschung und geben Einblicke in die Problematik reproduzierbarer Forschung. Dazu wird die Praxis der Biometrischen Zeitschrift (Biometrical Journal) bei der Umsetzung von reproduzierbarer Forschung herangezogen und es werden mögliche Hürden und Fallstricke aus der Sicht von Zeitschriften der angewandten Statistik und der Biometrie besprochen. Alltägliche Herausforderungen reproduzierbarer Forschung werden aufgezeigt und Lösungsansätze vermittelt.

Literate Programming, d.h. das „verweben“ von Text und Code ist eine hilfreiche Methode. Hierzu wird Sweave vorgestellt, welches die Verknüpfung von R-Code und LaTeX erlaubt. Hiermit können einfach und schnell Berichte oder Publikationen generiert werden. Ändern sich die Daten, so ändert sich auch das fertige Dokument. Ein kurzer Einblick in knitr, eine Weiterentwicklung von Sweave wird gegeben.

Versionskontrolle ist ein weiteres wichtiges Werkzeug zur Unterstützung reproduzierbarer Forschung und im *Projektmanagement* welche es erlaubt Änderungen an Dateien zu protokollieren und zu speichern (Backup). Somit werden Änderungen nachvollziehbar und können jederzeit Rückgängig gemacht werden. Auch das Arbeiten in Teams, sogar mit Unterschiedlichen Standorten, wird dabei

erleichtert. Hierzu werden Subversion (SVN) und Git/GitHub vorgestellt. Diese Systeme finden häufig in der Softwareentwicklung Verwendung, erleichtern jedoch auch die Erstellung und Verwaltung von Dissertationen, Publikationen und sonstigen Dokumenten.

Ablauf (vorläufige Planung)

12:00 - 13:00	Notwendigkeit und Ziele reproduzierbarer Forschung und ihre praktische Umsetzung in Publikationen
13:00 - 13:30	Reproduzierbare Forschung im Arbeitsalltag (z.B. bei der Erstellung von Berichten, Dissertation, Veröffentlichungen)
13:30 - 14:00	Kaffeepause
14:00 - 15:30	Einführung in Sweave und knitr mit Hilfe von RStudio
15:30 - 16:00	Kaffeepause
16:00 - 17:30	Einführung in "Projektmanagement"-Software: Subversion (SVN) und Git/GitHub für eigene Projekte und in Kollaborationen

Erforderliche Vorkenntnisse

Die Teilnehmer sollten grundlegende Computerkenntnisse haben. (Grund-)kenntnisse in LaTeX und der Programmiersprache R sind von Vorteil.

Teile des Kurses erfolgen interaktiv. Hierzu ist es von Vorteil wenn jeder Teilnehmer einen eignen Laptop zur Verfügung hat auf dem RStudio (1), R (2) und LaTeX (3) installiert sind.

Softwarequellen:

(1) http://www.rstudio.com/products/rstudio/download/
(2) http://cran.rstudio.com
(3) z.B. MiKTeX: http://miktex.org