BIG DATA
Ozapft is!

Spätestens seit Edward Snowden können sich selbst eingefleischte Computermuffel unter dem Begriff *Big Data* etwas vorstellen. Es geht da nicht einfach um „große Datenmengen“, sondern um ungeheuerliche Datenmassen im Peta- und Exabytebereich (10^{15} bis 10^{18} Byte), die mehr oder weniger wahllos aus reichlich sprudelnden Quellen abgezapft werden – nicht nur aus den Glasfaserkabeln der internationalen Telekommunikation oder den Logdateien von Google und Facebook, sondern auch aus wissenschaftlichen Beobachtungen und Experimenten, beispielsweise der Klimaforschung oder der Kernphysik.

Ein paar Petabyte kommen schnell zusammen

Zu den größten Datenlieferanten des 21. Jahrhunderts gehören die Biowissenschaften, die sogenannten „omics“, also *Genomics, Proteomics und Metabolomics* mit ihren Untervarianten (*Metagenomics, Transcriptomics, Lipidomics* etc.). Bei einer Genomanalyse in der molekularen Onkologie kommen schnell ein paar Petabyte (Millionen Gigabyte) zusammen.

Pro Fall muss man drei Milliarden Basenpaare mittels *Deep Sequencing* mehrfach absichern, um am Ende durchschnittlich 30.000 signifikante Abweichungen vom „Normalgenom“ zu finden. Will man dann eine nur halbwegs aussagekräftige Stichprobe von zum Beispiel je hundert Tumoren und Normalgeweben analysieren und die gefundenen drei Millionen Mutationen jedes einzelnen Falles mit den in der Literatur beschriebenen Funktionen von rund 5.000 „Tumorgenen“ verknüpfen, so kann man sich leicht ausrechnen, welch extrem große Speicherkapazitäten und Rechenleistungen man benötigt, um aus 100 x 30.000 x 5.000 wissenschaftlich fundierten Verknüpfungen ein medizinisch aussagekräftiges Ergebnis zu extrahieren.

Vor allem die USA und China konkurrieren derzeit auf dem Gebiet der Supercomputer und sind momentan bei schier unglaublichen 10^{16} Rechenoperationen pro Sekunde (*Flops*) angelangt. Von größerer praktischer Bedeutung ist allerdings das *v*erteilte Rechnen auf Clustern handelsüblicher Computer, wie es beispielsweise Google einsetzt, um Billionen von Webseiten zu durchforsten.

Attraktives Betätigungsfeld auch für Ärzte

Wichtiger als die Hardware sind natürlich die Menschen, die die Maschinen bedienen und die passenden Algorithmen entwickeln. Diese *Data Scientists* müssen nicht nur in Mathematik und Informationstechnik, sondern auch im jeweiligen Fachgebiet fit sein. In der Wirtschaft gehören solche Multitalente bereits zu den meistgesuchten IT-Spezialisten, zum Beispiel für Kundenanalysen und Wettbewerbsbeobachtungen im Onlinegeschäft, zur Energieverbrauchssteuerung komplexer Industrieanlagen oder auch zur Aufdeckung von betrügerischen Finanztransaktionen.

In den Biowissenschaften findet man sie vorerst fast ausschließlich in Informatikinstituten; einer von ihnen ist Prof. Michael Schröder, der Autor des Beitrags auf der nächsten Seite. Es lässt sich aber vorhersagen, dass auch der Bedarf an versierten Ärzten durch den immer leichteren Zugang zu großen Rechen- und Speicherkapazitäten wachsen wird. Für junge Mediziner – zum Beispiel Humangenetiker, Molekularpathologen, Laborärzte – ist es also durchaus lohnend, sich aktiv mit der Bioinformatik und dem *Big Data Management* auseinanderzusetzen. Der wissenschaftliche Reiz liegt vor allem in der Unvorhersehbarkeit der Ergebnisse, die sich durch das (primär ungezielte) Anzapfen reichlich sprudelnder „*omics*-Quellen“ ergeben. Den Neugierigen gehört die Zukunft. Prosit!

gh