GWDG NACHRICHTEN 05|18

Phishing

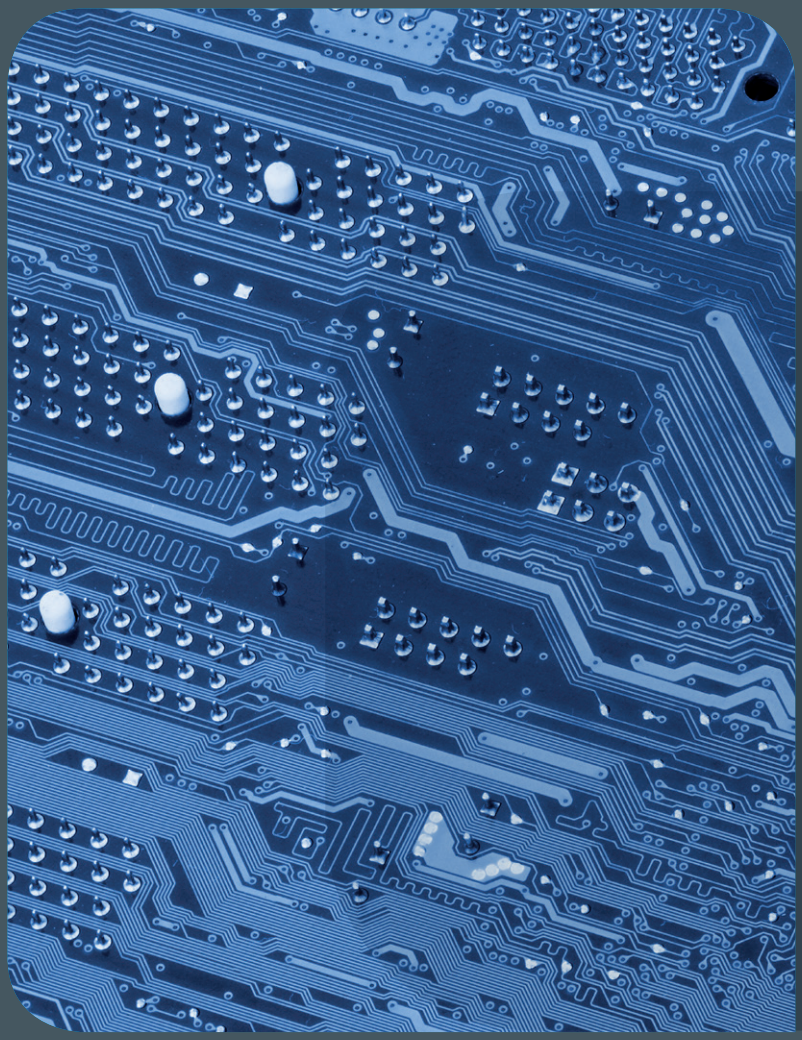
Bioinformatik-Software

Speicherstrategien

Data Mining

ZEITSCHRIFT FÜR DIE KUNDEN DER GWDG

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen



GWDG
NACHRICHTEN

05|18

Inhalt

Impressum

Zeitschrift für die Kunden der GWDG

ISSN 0940-4686
41. Jahrgang
Ausgabe 5/2018

Erscheinungsweise:
monatlich

www.gwdg.de/gwdg-nr

Auflage:
550

Fotos:
© Rogatnev - Fotolia.com (1)
© chagin - Fotolia.com (9)
© vege - Fotolia.com (19)
© Contrastwerkstatt - Fotolia.com (20)
© momius - Fotolia.com (23)
© MPIbpc-Medienservice (3)
© GWDG (2, 21)

Herausgeber:
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
Am Faßberg 11
37077 Göttingen
Tel.: 0551 201-1510
Fax: 0551 201-2150

Redaktion:
Dr. Thomas Otto
E-Mail: thomas.otto@gwdg.de

Herstellung:
Franziska Schimek
E-Mail: franziska.schimek@gwdg.de

Druck:
Kreationszeit GmbH, Rosdorf

Prof. Dr. Ramin Yahyapour
ramin.yahyapour@gwdg.de
0551 201-1545

Liebe Kunden und Freunde der GWDG,

es ist immer wieder ärgerlich, wenn trotz diverser Filter Spam-Nachrichten im Postfach landen. Ca. 50 % aller E-Mails, die im Internet transportiert werden, lassen sich Spam-Nachrichten zuordnen. Dank diverser Filter erreichen erfreulicherweise nur sehr wenige den Nutzer. Manche Spam-Nachrichten haben dabei durchaus solide Verkaufsabsicht, sind jedoch unerwünscht und störend. Viele Nachrichten haben jedoch leider einen kriminellen und betrügerischen Hintergrund.

Fast jeder kennt Phishing-Versuche, bei denen man angeblich dringend bei DHL, Amazon oder PayPal sein Konto überprüfen muss. Diese Nachrichten sehen teilweise täuschend echt aus, so dass man sehr genau hinschauen muss, um den Betrug zu erkennen. In ähnlicher Form gibt es auch E-Mails, die auf Nutzer der GWDG, der Universität Göttingen oder der Max-Planck-Gesellschaft abzielen. Teilweise sind diese mit Verweis auf irgendwelche Postfach-Änderungen immer noch recht gut zu durchschauen; teilweise gib es jedoch auch sehr gezielte Angriffe, die durchaus glaubwürdig erscheinen. Trotz aller technischen Filtermaßnahmen sind diese E-Mails leider nicht zuverlässig automatisiert zu unterbinden.

Die letzte Verteidigungslinie verbleibt daher beim Nutzer. In dieser Ausgabe haben wir deswegen einen Artikel, der ein wenig helfen soll, die Sensibilisierung zu schärfen, um nicht auf solche Nachrichten hereinzufallen. Es wäre auch bedauerlich, wenn jemand anderes im Internet mit Ihren Namen und Daten agiert.

Ich wünsche Ihnen viel Freude beim Lesen dieser Ausgabe.

Ramin Yahyapour

GWDG – IT in der Wissenschaft

Vorsicht Phishing!

Text und Kontakt:
Dr. Holger Beck
holger.beck@gwdg.de
0551 201-1554
Jan-Nikolas Fahrenholz
jan-nikolas.fahrenholz@gwdg.de
0551 201-1536

Täglich stöhnen wir über unsere überfüllten Postfächer. Noch schlimmer: Da sind dann noch wichtige E-Mails von unserer Bank, vom Online-Shop oder gar von unserem Arbeitgeber oder dem Rechenzentrum dabei, die unbedingt sofort beantwortet werden müssen, damit das Konto sicher ist, das tolle Geschenk nicht verpasst wird oder das E-Mail-Konto nicht gesperrt wird. Also wenigstens das schnell erledigen, nur mal kurz anmelden, ein Häkchen an der richtigen Stelle setzen … Oder war das etwa gar nicht echt?! Wollte da ein Online-Betrüger etwas von mir? Ja, in den vergangenen Wochen waren solche Betrüger auch im Umfeld der GWDG unterwegs und leider gelegentlich mit ihrer Betrugsmasche erfolgreich. Davor wollen wir hier warnen, über Folgen berichten und Tipps geben, wie man betrügerische von echten Anfragen unterscheiden kann.

IHRE E-MAIL KONNTE NICHT ZUGESTELLT WERDEN!

Eine solche Nachricht haben in den vergangenen Wochen gelegentlich Nutzerinnen und Nutzer des E-Mail-Service der GWDG erhalten (siehe Abb. 1). Dabei hatten sie doch nur eine ganz normale E-Mail verschickt. Nun steht da aber etwas von Spam!

Von: Mail Delivery System <Mailer-Daemon@gwdg.de>
Datum: 21. März 2018 um 23:23:30 MEZ
An: < @ mpg.de>
Betreff: Mail delivery failed: returning message to sender

This message was created automatically by mail delivery software.

A message that you sent could not be delivered to one or more of its recipients. This is a permanent error. The following address(es) failed:

d .d @ .de
host mx0.webpack.hosteurope.de [80.237.138.5]
SMTP error from remote mail server after HELO emailer.gwdg.de:
550-REJECT: 134.76.10.24 is in ix.dnsbl.manitu.net :
550-Your e-mail service was detected by mail.ixlab.de (NiX Spam) as spamming at
550-Wed, 21 Mar 2018 22:39:46 +0100. Your admin should visit
550 http://www.dnsbl.manitu.net/lookup.php?value=134.76.10.24 (ID:550:3:0)

1_Spam-Fehlermeldung

Ja, die Betroffenen waren gänzlich unschuldig. Nicht sie selbst hatten Spam verschickt, aber eben doch einzelne andere Nutzer des E-Mail-Service der GWDG. Und das kann weitreichende Folgen haben.

Wer Spam empfängt, kann das bei verschiedenen Dienstleistern melden, die solche Meldungen sammeln und auswerten. Wenn solche Meldungen sich häufen, trägt der Dienstleister den E-Mail-Server (genauer: dessen IP-Adresse), über den Spam verschickt wurde, in eine „Schwarze Liste" oder „Blacklist" ein. Dort, wo Betreiber von E-Mail-Diensten solche Blacklists nutzen, um ihren Kunden den Empfang von Spam zu ersparen, werden dann alle E-Mails abgelehnt, die über einen E-Mail-Server verschickt wurden, der gerade auf der Blacklist steht. So kam es durch Spam-Versand über einzelne E-Mail-Konten des E-Mail-Service der GWDG zu entsprechenden Unzustellbarkeitsmeldungen an unschuldige andere Absender. Zudem können die Blacklist-Einträge nicht nur zu einer Ablehnung der E-Mail-Zustellung, sondern auch zu – lediglich – einer Verlangsamung der Zustellung führen. Das kann sogar noch ärgerlicher sein, weil bei einer Ablehnung i. d. R. sofort eine Information an den Absender geschickt wird. Über eine Verzögerung wird jedoch meistens nicht informiert. Absender wundern sich dann vielleicht, warum die Kommunikationspartner so zögerlich reagieren, ohne den Grund auf Anhieb zu erkennen.

Um umfangreiche Störungen zu vermeiden, waren immer wieder schnelle Reaktionen von Seiten der GWDG erforderlich. Die Konten, über die Spam verschickt wurden, wurden gesperrt und die Betreiber der Blacklists kontaktiert, um den Eintrag in der Blacklist wieder löschen zu lassen. Die administrativen Aufwände, den durch solche Vorfälle entstandenen Schaden gering zu halten, sind enorm.

Der Schaden hielt sich auch deshalb in Grenzen, weil der E-Mail-Service der GWDG mit mehreren Servern leistungsfähig genug ausgelegt ist und die Konfiguration des Dienstes die Möglichkeiten des Spam-Versands einschränkt: Für jedes Konto (mit wenigen Ausnahmen) ist der E-Mail-Versand auf 500 E-Mails in 24 Stunden begrenzt. Ohne diese Begrenzung hätten die Spammer auch versuchen können, den E-Mail-Service durch Überlastung in die Knie zu zwingen.

SPAMMER BEI DER GWDG?

Aber wie konnte es überhaupt so weit kommen? Hat die GWDG in ihrem Nutzerkreis kriminelle Personen, die sich mit Spam-Versand ein Zubrot verdienen?

Nein, – obwohl seit Jahrzehnten immer wieder vor den berüchtigten Innentätern gewarnt wird – die Spam-Versender waren selbst nur Opfer. Sie hatten selbst betrügerische E-Mails erhalten und waren im guten Glauben den Anweisungen in diesen

E-Mails gefolgt, in denen behauptet wurde, dass diese von der GWDG, der Universität Göttingen oder der MPG kämen. Wer den Anweisungen folgte, um eine wichtige Information zu erhalten oder eine Überprüfung des eigenen Kontos zuzulassen, meldete sich dann nicht bei einem Dienst der GWDG, Universität Göttingen oder MPG, sondern auf einer gefälschten Webseite mit seinen Zugangsdaten an. Das war das Ziel der Betrüger: Über die gefälschten Webseiten haben sie Benutzernamen und Passwörter der Opfer erhalten. Vom englischen „fishing" abgeleitet, wird diese Vorgehen „Phishing" genannt. (Manche meinen, das P am Wortanfang stehe für „password fishing." Das ist aber nicht eindeutig bewiesen.)

Dank des Phishing-Erfolgs konnten die Betrüger nun unter dem Namen legaler Nutzer über den E-Mail-Service der GWDG Spam verschicken. Dass die nächsten Phishing-E-Mails nun über GWDG-Konten an GWDG-Konten geschickt werden konnten, hat die Erfolgsaussichten der Phishing-E-Mails noch erhöht. Die Anfrage kam ja nun von einem Konto, das auf *gwdg.de*, *uni-goettingen.de* oder *mpg.de* endete. Das ließ die Phishing-E-Mails noch ein wenig glaubwürdiger wirken.

DAS KONTO ZURÜCKEROBERN

Die digitale Identität – zumindest eine – war den Opfern nun geraubt. Wie erobert man sie zurück?

Zunächst dürfte es immer die GWDG gewesen sein, die den Missbrauch kurzfristig unterbunden hat, indem die betroffenen Konten gesperrt wurden. Der Spam-Versand war damit unterbunden. Auch alle anderen Zugriffe auf Dienste, für die dieses Konto benötigt wurde, waren blockiert – für die Betrüger, aber auch für die wahren Kontoinhaber.

Die Kontoinhaber haben durch die Sperre meist schnell gemerkt, dass etwas nicht stimmte, und sich bei lokalen IT-Betreuern oder gleich bei der GWDG gemeldet. Über diese wurden sie über das Problem informiert und im Gespräch wurde nach einer Ursache gesucht. Dabei konnte dann meist recht schnell festgestellt werden, dass die Betroffenen Opfer von Phishing-E-Mails geworden waren (was nicht selbstverständlich war, denn Passwörter hätten auch auf anderen Wegen, z. B. über Keylogger, abgegriffen werden können). Nach der Klärung des Problems im Gespräch (und Tipps zum Umgang mit solchen E-Mails) wurden dann neue Passwörter gesetzt, die die Kontoinhaber bei der GWDG oder lokalen IdM-Administratoren abholen konnten, und das Konto wurde entsperrt.

Für die Opfer bedeutet ein solcher Angriff durch die – unbedingt notwendige – Sperre zunächst einen massiven Eingriff in ihre Arbeitsabläufe. Mit Glück hat die GWDG einen Weg gefunden, die Opfer über einen Ansprechpartner im Institut zu informieren. Wenn das nicht möglich war, mag es gedauert haben, bis man auf den Gedanken kam, die GWDG zu kontaktieren. Zudem bekommt man das neue Passwort verständlicherweise auch nicht einfach so, sondern muss sich erst einmal ausweisen – andernfalls könnte ja der nächste Betrüger das Konto wieder übernehmen, weil die GWDG ihm das einfach mitteilt.

Aber ist damit das ganze Ärgernis beseitigt?

Wenn die Betrüger auf ihrer gefälschten Webseite nur das Passwort abgegriffen und die übernommene Identität nur zum Versand von Spam genutzt hatten, wäre das Problem nach Änderung des Passworts höchstwahrscheinlich behoben. Aber stimmt die obige Voraussetzung?

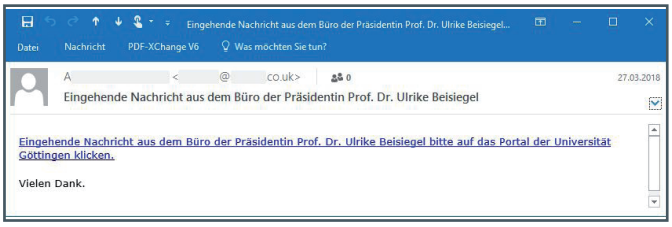
Die Phishing-Webseite hätte nicht nur zum Abgreifen des Passworts eingesetzt worden sein können. Die Angreifer hätten dort auch Schadsoftware hinterlegt haben können, mit der das Gerät, von dem die Webseite aufgerufen wurde, mit dieser hätte infiziert werden können.

Die Betrüger hatten die digitalen Identitäten der Kontoinhaber übernommen. Alle Dienste, die die Kontoinhaber mit ihren Identitäten nutzen können, hätten die Betrüger auch nutzen können – nicht nur den E-Mail-Versand. Sie hätten E-Mails lesen und auch unter Ausnutzung des Namens des Betroffenen schreiben, über VPN auf das interne Netz zugreifen, Dateien auf Servern auslesen, kopieren, ändern oder neu anlegen können (normale oder solche mit weiteren Schadfunktionen) usw.

Bisher gibt es keine Indizien, dass die Betrüger über den Spam-Versand hinaus die Konten missbraucht hätten. Aber wie immer, ist es schwierig, nachzuweisen, dass kein Ereignis stattgefunden hat, zumal viele Ereignisse theoretisch denkbar wären. Für die Opfer bleibt nur der Rat zu erhöhter Aufmerksamkeit und zur Überprüfung der verwendeten Geräte auf Schadsoftware.

PHISHING ERKENNEN

Früher waren Spam- und Phishing-E-Mails – zumindest für den Muttersprachler – an gebrochenem Deutsch und unrealistischen Bezügen leicht zu erkennen. Heute hat man es mit professionellen Kriminellen zu tun, die in fehlerfreiem Deutsch schreiben und häufig sehr genau die Umgebung ihrer potenziellen Opfer erkunden. Um Vertrauen zu erwecken und die Motivation zu erhöhen, wird auf prominente Personen oder Ereignisse in der Umgebung des Opfers Bezug genommen; im Beispiel in Abb. 2 ist es der klassische Bezug auf die oberste Leitungsebene. Auch die Webseiten der Betrüger enthalten mittlerweile meist passende Logos oder sind gleich komplette Kopien von realen Webseiten. Technisch ist das ja kein Problem, denn die Originalseiten stehen öffentlich im Internet bereit und lassen sich einfach kopieren (und für Zwecke der Kriminellen unmerklich modifizieren).



2_Neugier als Anreiz zum Klicken

Warum verhindert die GWDG nicht die Zustellung von Phishing-E-Mails?

Der Eindruck, dass die GWDG die Zustellung von Spam-E-Mails und Phishing-E-Mails nicht verhindert, täuscht. Der E-Mail-Service der GWDG nutzt den E-Mail-Support der des DFN-Vereins. Dort werden eingehende E-Mails auf Viren und Spam untersucht und eine Annahme von E-Mails, die virenverseucht sind oder als Spam klassifiziert wurden, wird dort verweigert, d. h. der Absender bekommt eine Fehlermeldung, dass seine E-Mail nicht angenommen wurde. Der DFN-Verein berichtete auf der letzten Betriebstagung im März, dass vom DFN-E-Mail-Support im Februar 2018 von ca. 3,2 Millionen eingelieferten E-Mails über 78 % direkt wegen Viren- oder Spam-Verdacht abgewiesen wurden. Ein

großer Teil der Gefährdungen wird also schon unterbunden.

Aber während diese Verfahren für E-Mails, die (bekannte) Viren enthalten, sehr gut funktioniert, ist die Erkennung von Spam und Phishing leider nicht so eindeutig möglich. Dadurch rutschen doch immer wieder solche E-Mails durch diesen Filter. Noch problematischer wird es, wenn Phishing-E-Mails von einem Konto des E-Mail-Service der GWDG zu einem anderen Konto dieses Dienstes geschickt werden. Diese internen E-Mails werden bisher auch nur intern verarbeitet und laufen gar nicht durch den Filter des DFN-E-Mail-Support. Da in den vergangenen Wochen leider wiederholt Phishing-Angriffe intern abliefen, konnten diese technisch nicht unterbunden werden.

Als Reaktion auf diese Problematik versucht die GWDG zudem auch intern im E-Mails-Service, zusätzliche Filter zu schalten. Die ersten Versuche haben bei der Bekämpfung der aktuellen Phishing-Welle bereits positive Ergebnisse gezeigt, werden aber (wie grundsätzlich alle Filtersysteme) nie einen vollständigen Schutz bieten können.

Künstliche Intelligenz scheint hier bisher der menschlichen noch unterlegen zu sein. Ihre Mithilfe ist daher weiter ein wichtiger Faktor. Bei der Erkennung von Phishing-E-Mails wollen wir Ihnen mit den folgenden Tipps helfen.

Passen Absender und Inhalt zusammen?

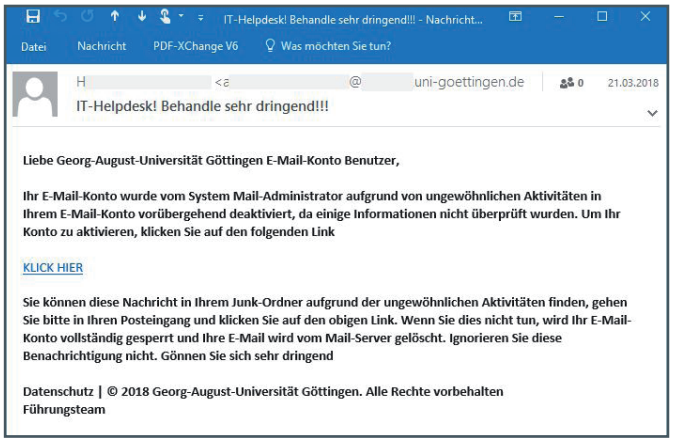
In der Hektik des Alltags kann man sich schnell durch Phishing-E-Mails zu übereilten und fatalen Handlungen verleiten lassen. Erster Rat muss daher sein: Ruhe bewahren und bei jeder E-Mail überlegen, ob Inhalt und (möglicherweise sogar nur vorgetäuschter) Absender zusammenpassen.

Am einfachsten ist das, wenn der Absender wie im ersten Beispiel nicht auf *gwdg.de*, *uni-goettingen.de* oder *mpg.de* endet (hier war es ein Absender aus dem United Kingdom *uk*). Aber vielleicht sind es auch nur Stil, die Art der Anrede oder die nicht den zu erwartenden Schlussformeln entsprechenden Formulierungen in der E-Mail, die misstrauisch machen könnten.

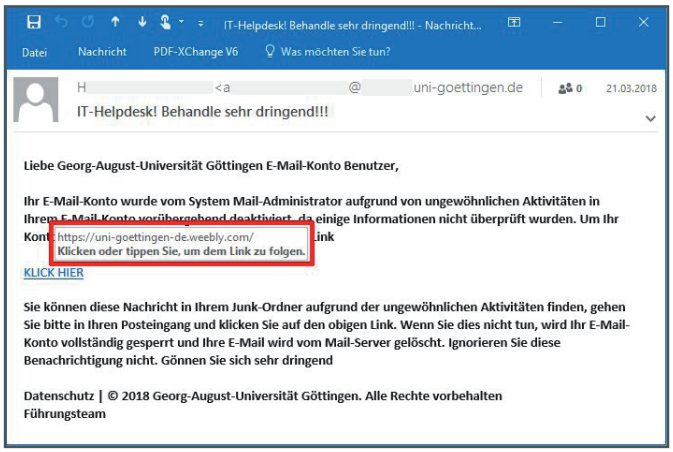
Wenn es, wie hier behauptet wird, eine E-Mail der GWDG (oder ihrer Bank u. a.) wäre, dann ist es schon verdächtig, dass ein Link in der E-Mail steht, der zu einer Anmeldeseite führen soll. Die GWDG wie auch Banken u. a. vermeiden es möglichst, solche Links zu verschicken. Soweit es die GWDG betrifft, gibt es nur wenige Ausnahmen (leider auch die Warn-E-Mails, bevor Konten gesperrt werden, wenn die Fristen für Passwortänderungen fast abgelaufen sind, aber z. B. auch E-Mails, die SharePoint oder GWDG ownCloud erzeugen, wenn Nutzer Inhalte teilen). Eine E-Mail von der GWDG, die einen Link zum Anmelden enthält, sollte man daher auf jeden Fall kritisch prüfen (dazu weiter unten mehr). Wenn die GWDG E-Mails mit Links verschickt, dann sollten diese zumindest nicht hinter einem undurchschaubaren „Klicken Sie hier“ verborgen sein.

Links prüfen

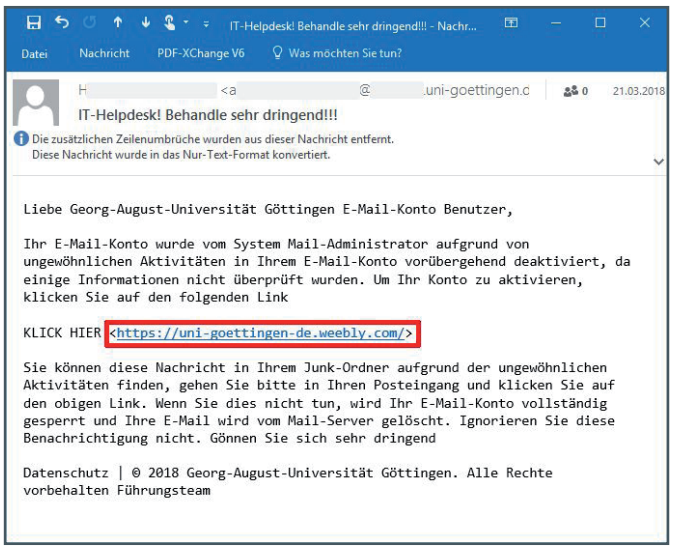
Das „Klicken Sie hier“ führt zum nächsten Rat: Sehen Sie sich genau an, wohin der Link Sie führen will. Was sich hinter dem Link versteckt, erfahren Sie, wenn Sie die Maus über den Link halten und kurz warten. Dann erscheint über dem Link (oder bei manchen Programmen auch in der Statuszeile am unteren Rand des Fensters) ein Hinweis auf das tatsächliche Ziel des Links. Das „KLICK HIER“ aus Abb. 3 offenbart dann das Ziel *https://uni-goettingen-de.weebly.com/* (siehe Abb. 4). Eine andere (vom Autor für sich



3_Drohung als Anreiz zum Klicken



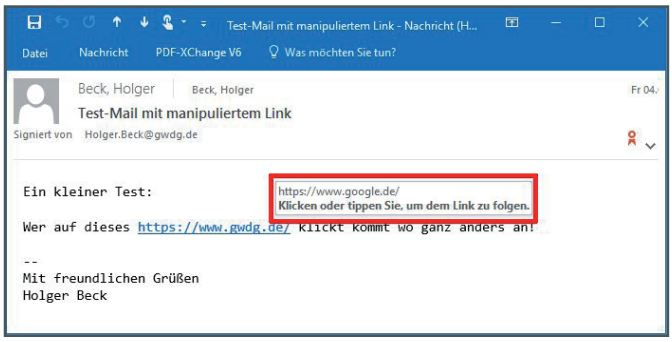
4_Phishing-E-Mail mit Maus-Over



5_Phishing-E-Mail in Nur-Text-Darstellung

persönlich bevorzugte) Variante wäre, das E-Mail-Programm so einzustellen, dass E-Mails zunächst immer im Nur-Text-Modus angezeigt werden (siehe Abb. 5). Hier sieht man das Link-Ziel sofort, ohne erst mit der Maus über dem Link anzuhalten.

Zu Links noch ein Hinweis: Selbst wenn in der E-Mail nicht nur „KLICK HIER“, sondern scheinbar direkt ein Link-Ziel der Form *https://...* erscheint, kann sich dahinter etwas ganz anderes verbergen. Wie beim „KLICK HIER“ ist auch ein E-Mail-Text, der wie ein Link aussieht, nur ein beliebiger Text. Den tatsächlichen Link sieht man auch dann nur, wenn man die Maus über den Link hält oder den Nur-Text-Modus verwendet. Das Beispiel in Abb. 6 illustriert das Problem.



6_Nur scheinbar leicht zu erkennendes Link-Ziel

Grundsätzlich empfiehlt es sich also, auch dann Links zu verifizieren, wenn es so aussieht, als ob man den ja schon so vertrauen kann.

Aber was macht man beim Lesen von E-Mails auf Smartphones? Die E-Mail-Apps haben meist keinen Nur-Text-Modus und eine Maus gibt es auch nicht. Der Trick ist hier, mit dem Finger nicht den Link kurz anzuklicken, sondern mit dem Finger länger auf den Link zu „drücken". Dann wird auch auf dem Smartphone das Link-Ziel angezeigt, bevor es im Browser geöffnet wird.

Ist damit dann alles klar, wenn man den wahren Link gesehen hat? Auf den ersten Blick könnte man meinen, dass ein Link z. B. zu einem System der Universität Göttingen gehört, wenn in diesem etwas von *uni-goettingen* und *de* steht. Aber ist das immer so? Damit zum nächsten Tipp.

Wer steckt hinter einem Link oder einer E-Mail-Adresse?

Links zu Webseiten können schon beliebig kompliziert und unübersichtlich sein. Hier ein konstruiertes Beispiel:
https://windturbinen.maschinenbau.uni-goettingen.de/turbine-einsatz/selbst-bei-tornados/php?id=34i2tbfu2iiu+name=suedlich-des-nordpols

Dieser Link könnte formal tatsächlich zu einem Webauftritt eines Projekts der Universität Göttingen gehören (wenn man davon absieht, dass es dort keine Maschinenbau-Fakultät gibt). Der nächste Link wäre aber wahrscheinlich ein Betrugsversuch:
https://windturbinen.maschinenbau.uni-goettingen.de-i.in/turbine-einsatz/selbst-bei-tornados/php?id=34i2tbfu2iiu+name=suedlich-des-nordpols

Auf den ersten Blick sieht man den Unterschied nicht. Mit etwas Suchen findet man den Fehler: Statt *uni-goettingen.de* steht da *uni-goettingen.de-i.in*. Das wäre dann keine Webseite von *uni-goettingen* im Land *de* (Deutschland) mehr, sondern eine Webseite der Domain *de-i* im Land *in* (Indien).

Ist das Geheimnis beim Prüfen eines Links nun ein Langgenug-auf-den-Link-Starren? Zum Glück nicht. Man kann das Problem, festzustellen, wer sich hinter einem Link verbirgt, d. h. den eigentlichen Domänennamen der Organisation zu finden, auch systematisch angehen. Die Organisation zu erkennen, auf die ein Link verweist, ist wesentlich für die Unterscheidung zwischen vertrauenswürdigen und nicht vertrauenswürdigen Links. Dazu benötigt es vier Schritte:

1. Fangen Sie nach dem *http://* oder *https://* an zu suchen, oder ganz am Anfang, wenn beides nicht angezeigt wird,
2. gehen Sie von da bis zum Ende der Adresse oder dem nächsten /,
3. gehen Sie dort zwei . zurück,
4. dann ist nur das, was zwischen diesem . und dem Ende oder dem / steht, die Domäne der Organisation!

Abb. 7 illustriert dieses Vorgehen. Auch in diesem Beispiel ist die Antwort auf das „Wer" ein *de-i.in* und damit führt ein solcher Link höchstwahrscheinlich nicht zu einer Original-Webseite der Universität Göttingen.

7_Wer ist das?

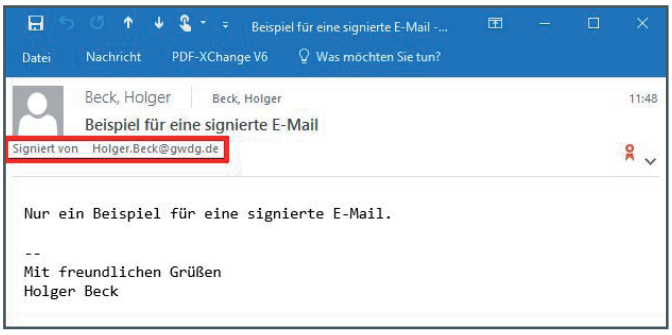
An dieser Stelle muss leider noch eine Warnung angebracht werden: Betrüger verwenden auch gerne „Tippfehlerdomänen", also Domänennamen, die sich nur durch kleine Abweichungen vom richtigen Domänennamen unterscheiden. Also nicht übersehen, wenn statt *gwdg.de gwdg.dk* oder statt *mpg.de mgp.de* stehen sollte. *uni-goettingen.de* bietet leider noch viel mehr Möglichkeiten für kleine, unauffällige Abweichungen.

Absender von E-Mails verifizieren

E-Mails sind bekanntlich wie Postkarten. Wie auf Postkarten oder Briefen kann man einen beliebigen Absender vorgeben. Man kann daher nicht davon ausgehen, dass der angegebene Absender tatsächlich der Absender ist. Experten können in (normalerweise von E-Mail-Programmen ausgeblendeten) E-Mail-Headern Indizien ablesen, ob der Absender gefälscht ist. Für die meisten E-Mail-Nutzer dürfte das keine Option sein.

Mittels digitaler Signatur gibt aber auch eine Art, E-Mails zu versenden, die es dem Empfänger ermöglicht, den Absender sicher festzustellen. Das Gute dabei ist, dass die für die Prüfung nötigen Kenntnisse relativ schnell zusammengefasst sind. (Zu erklären, wie man solche E-Mails versendet oder wie die Technik – Zertifikate oder PKI – dahinter aussieht, würde länger dauern. Wer dazu mehr wissen will, kann dieses Thema im Sonderheft 1/2014 „E-Mail-Verschlüsselung mit X.509.Zertifikaten" der GWDG-Nachrichten nachlesen.)

Mit einem Zertifikat signierte E-Mails erkennt man an einem speziellen Symbol (z. B. bei Outlook oder im Webinterface des E-Mail-Service der GWDG; andere Programme verwenden andere Symbole als Prüfsiegel). Solche E-Mails sind über kryptographische Funktionen so gesichert, dass die Angabe, welcher Absender die E-Mail signiert hat, vertrauenswürdig ist, d. h. die E-Mail kommt tatsächlich von dem unter „Signiert von" (rot umrahmter Bereich in Abb. 8) angegebenen Absender – zumindest solange, wie dieser den für diese Signatur nötigen Schlüssel geschützt hat.



8_Eine signierte E-Mail

Leider werden nur wenige E-Mails auf diese Weise signiert, sodass Absender häufig nicht verifiziert werden können.

Aber die wenigen E-Mails, die die GWDG sendet und Ihnen darin (vielleicht überraschend) Links zusendet, sind immer signiert. In den meisten Fällen kommen solche E-Mails auch vom Absender

support@gwdg.de (und sind von diesem signiert). Gerade wenn es um Phishing-E-Mails geht, die behaupten, von der GWDG zu kommen, haben Sie somit eine sehr gute, einfache und schnelle Prüfgrundlage. Unsignierten E-Mails, die angeblich von der GWDG kommen, sollten sie prinzipiell nicht trauen. Falls Sie diese nicht gleich ignorieren, sollten Sie zumindest bei der Service-Hotline der GWDG nachfragen, bevor Sie Anweisungen in solchen E-Mails folgen.

Noch eine Hilfe von Seiten der GWDG

Alle Phishing-E-Mails, die vorgeben, von der GWDG, der Universität Göttingen oder der MPG verschickt worden zu sein und von denen die GWDG Kenntnis erhalten hat, werden unter *https://www.gwdg.de/phishing-warnung* aufgelistet. Wenn Sie eine verdächtige E-Mail erhalten, können Sie dort zunächst prüfen, ob diese E-Mail der GWDG schon bekannt ist und von der GWDG als Phishing-E-Mail identifiziert wurde.

Der Umkehrschluss gilt leider nicht: Wenn Sie Ihre verdächtige E-Mail dort nicht finden, bedeutet dies leider nicht, dass diese als sicher anzusehen ist. Die GWDG kann nur Phishing-E-Mails veröffentlichen, die sie kennt. Daher sind wir auf Ihre Hilfe angewiesen: Sollten Sie eine verdächtige E-Mail mit Bezug zur GWDG, der Universität Göttingen oder der MPG erhalten haben, die auf der Warnungs-Webseite nicht aufgeführt ist, bitten wir Sie, diese an *support@gwdg.de* weiterzuleiten. So erhalten Sie Hilfe bei der Entscheidung über die Vertrauenswürdigkeit der E-Mail und die GWDG die nötige Information zur Pflege ihrer Warnungs-Webseite.

Mehr Informationen

Wer sich zum Thema Phishing-E-Mails weiterbilden will, dem kann die Webseite *https://secuso.org/nophish* empfohlen werden. Secuso ist eine Forschergruppe an der TU Darmstadt. Neben einem Video, das Phishing sehr anschaulich erklärt, finden Sie dort auch ein Trainingsprogramm, mit dem Sie üben können, Phishing-E-Mails zu erkennen.

Für IT-Sicherheit am Arbeitsplatz bietet die Informationssicherheitsgruppe am Göttingen Campus einen 90-minütigen Kurs an, der auf Anforderung in Einrichtungen in Göttingen vor Ort abgehalten werden kann.

Attention Phishing!

Phishing is a constant threat for users of e-mail communication. Regrettably, in the last time phishers successfully targeted users of GWDG's e-mail service and took over some accounts and misused them for sending more spam and phishing e-mails. This misuse cause minor disruption of GWDG's e-mail service (a few rejected e-mails and some delayed e-mails deliveries).

For protection against such attacks, users should carefully check there e-mails, verify the identity of the originator of the e-mails and the destinations of all links embedded in e-mails (or best never click on links in unexpected e-mails).

Before entering credentials on (possibly faked) websites, users must examine the website's domain. The last two parts of the domain name (parts meaning text between the dots in the name) will reveal the organization, to which the domain belongs. (The domain name is the text between the beginning *http(s)://* and the next */* or the end.)

E-mail certificates provide reliable identification of sender's identities. The GWDG will always sign official e-mails with such certificates. You will see a sign like  (the exact form is depending on the e-mail app) in or before the e-mail header as visualisation of a verification and certification of the sender.

Users, who receive suspicious e-mails should check *https://www.gwdg.de/phishing-warning*. The GWDG will document known phishing e-mails claiming to be sent by the GWDG, Göttingen University or MPG at his place. If you do not find such a suspicious e-mail documented there, please check carefully and inform the GWDG about such e-mails via *support@gwdg.de*.

Kurz & knapp

Öffnungszeiten des Rechenzentrums um Pfingsten

Das Rechenzentrum der GWDG ist an den beiden Pfingstfeiertagen 20.05. und 21.05.2018 geschlossen.

Falls Sie sich zu der Zeit, in der das Rechenzentrum geschlossen ist, in dringenden Fällen an die GWDG wenden wollen, schicken Sie bitte eine E-Mail an *support@gwdg.de*. Das dahinter befindliche Ticket-System wird auch während dieser Zeit von Mitarbeiterinnen und Mitarbeitern der GWDG regelmäßig überprüft.

Wir bitten alle Benutzerinnen und Benutzer, sich darauf einzustellen.

Grieger

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

MS SharePoint

KOLLABORATION LEICHT GEMACHT!

Ihre Anforderung

Sie möchten eine kooperative Kommunikations- und Informationsplattform für Mitarbeiter einrichten, die ständig und von überall verfügbar ist. Sie benötigen ein integriertes Dokumentenmanagementsystem und möchten gemeinsame Besprechungen und Termine planen und verwalten.

Unser Angebot

Wir bieten Ihnen SharePoint als Kollaborationsplattform. Wir können z. B. eine SharePoint Site Collection als gemeinsames Portal für Ihre Arbeitsgruppe oder Ihr Projektteam einrichten. Eine solche Site Collection kann sowohl in Englisch als auch in Deutsch präsentiert werden. Mit einer umfangreiche Auswahl an Schablonen, Apps und Layout-Vorlagen können Sie das Design Ihrer Site Collection anpassen. Der Zugriff erfolgt über GWDG-Benutzerkonten. Weitere Authentifizierungsverfahren sind möglich.

Ihre Vorteile

> Einheitliches Dokumenten-Managementsystem
> Umfangreiche Listen und Bibliotheksfunktionen für Dokumente, Bilder oder Dateien
> Steigern der Produktivität der Mitarbeiter durch vereinfachte tägliche Geschäftsaktivitäten.
> Einfaches Planen und Protokollieren von Besprechungen
> Führen nicht öffentlicher Diskussionsrunden
> Wissensmanagement: Aufbau eines Wikis für Ihre Mitarbeiter
> Bereitstellung von Informationen und Fachwissen für Mitarbeiter
> Geringer Entwicklungs- und Pflegeaufwand der SharePoint-Plattform für Benutzer
> Individuell anpassbares Layout und Design
> Optimale MS Office-Anbindung
> Einfache Benutzer- und Gruppenverwaltung

Interessiert?

Der Dienst steht allen Mitgliedern der Max-Planck-Gesellschaft und der Universität Göttingen zur Verfügung. Voraussetzung für die Nutzung ist die Benennung eines Ansprechpartners, der die Administration Ihrer Site Collection übernehmen soll. Wenn Sie SharePoint nutzen möchten, senden Sie bitte eine entsprechende E-Mail an support@gwdg.de. Nähere Informationen zu SharePoint sind auf der u. g. Webseite zu finden.

www.gwdg.de/sharepoint

Wer suchet, der findet – die neuen Webseiten zur Bioinformatik-Software bei der GWDG

Text und Kontakt:
Dr. Rainer Bohrer
rainer.bohrer@gwdg.de
0551 201-1829
Hans-Georg Sommer
hans-georg.sommer@gwdg.de
0551 201-1791
Eilika Wülfing
eilika.wuelfing@gwdg.de

Das umfangreiche Softwareangebot der GWDG im Bereich Bioinformatik (z. Zt. über 100 Programme) lässt sich jetzt wesentlich besser erschließen, da es nun möglich ist, in einer Auflistung der zur Verfügung stehenden Programme zu suchen. Vier neue Webseiten ermöglichen nicht nur die gezielte Suche nach einem Programm anhand seines Namens, sondern auch nach Verwendungszweck und Einsatzgebiet.

DIE ENTWICKLUNG DER LETZTEN JAHRE

Da sich in den letzten zehn Jahren durch die rasante Weiterentwicklung der Messtechnik im Bereich der Molekularbiologie die rechenintensive Auswertung von Messergebnissen, Next Generation Sequencing (NGS) und Massenspektrometrie zu einem festen Bestandteil der Forschungsarbeit entwickelt haben, ist die Anzahl der dafür notwendigen Programme, auch bei der GWDG als Rechenzentrum, sehr deutlich angestiegen. Die sich relativ schnell verändernden wissenschaftlichen Fragestellungen haben zusammen mit der Verbreitung von Linux und relativ einfachen Programmierwerkzeugen wie Java, Python und Perl zu einem deutlichen Anstieg von Bioinformatik-Programmen insbesondere in der Open-Source-Szene geführt. Durch die gestiegenen Rechenanforderungen sind die Bioinformatik-Programme auch in den Bereich des High Performance Computing vorgestoßen, so dass heutzutage ein Bioinformatiker vor einer deutlich komplexeren Hardware-Landschaft steht. Schon seit einiger Zeit ist die Anzahl der bei der GWDG zur Verfügung gestellten Programme auf über 100 angestiegen und es wurde dringend notwendig, unseren Kunden aus dem Bereich der Bioinformatik eine bessere Übersicht zu geben, welche Programme unter welchen Betriebssystemen und Zugriffsmöglichkeiten zur Verfügung stehen. Denn inzwischen stehen für die Arbeit mit den Programmen Rechner im Dialog- und Batchbetrieb, Linux- und Windows-Server, Webinterfaces und Lizenzserver zur Verfügung.

Darüber hinaus wurde es notwendig, trotz der Fülle der Programme seitens der Administration Wege zu finden, dass der Arbeitsaufwand zum Einpflegen der Informationen für neue Programme auf ein Minimum reduziert wird und dies zeitnah mit der Installation der Programme erfolgen kann.

DIE NEUEN WEBSEITEN

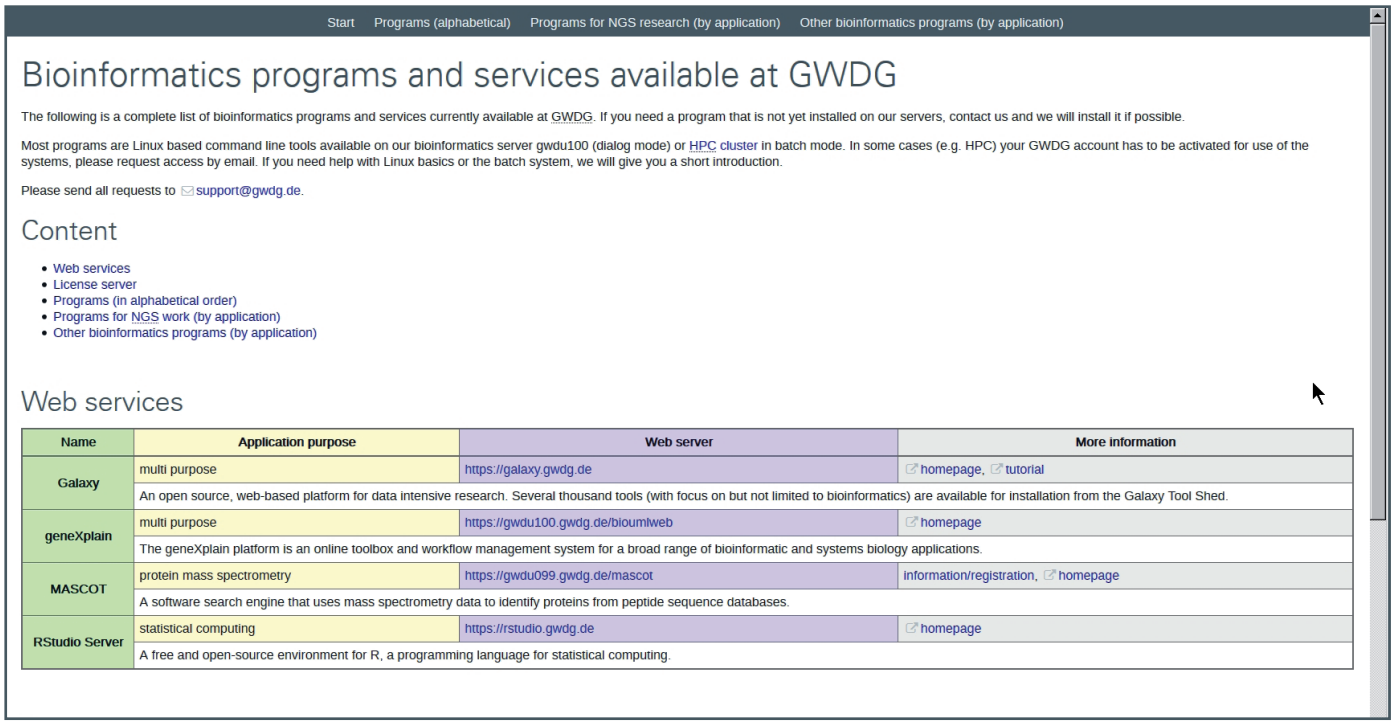
Da ständig immer wieder neue Programme installiert werden, lag es nahe, schon bei ihrer Installation die wichtigsten Informationen (Verwendungszweck, Verfügbarkeit, weiterführende Links) bereitzustellen, um dann die Webseiten automatisch generieren zu können. Diese generierten Webseiten werden sowohl über die *gwdu100* direkt bereitgestellt, als auch von der Webrepräsentanz der GWDG aus verlinkt. Bei dem Design wurde berücksichtigt, dass die generierten Webseiten auch auf mobilen Devices einschließlich Smartphones noch möglichst gut darstellbar sein sollen, weshalb die verschiedenen tabellarischen Listen auch auf vier Webseiten aufgeteilt wurden, um den Aufwand beim Scrollen nicht unnötig groß zu machen.

Die Startseite (*https://gwdu100.gwdg.de*) zeigt neben einer kurzen Einführung und der Verlinkung auf die eigentlichen Programmlisten zwei Tabellen mit den angebotenen **Webservices** und den vorhandenen **Lizenzservern** (siehe Abb. 1). Die drei anderen Webseiten enthalten tabellarische Programmlisten, die nach folgenden drei Kriterien erstellt wurden:

1. eine **alphabetische Sortierung** aller Programme *https://gwdu100.gwdg.de/proglist.html* (siehe Abb. 2)

New Websites for Bioinformatics Software

The GWDG's extensive range of bioinformatics software (currently more than 100 programs) can now be better exploited, as it is now possible to search in a list of available programs. Four new websites not only allow a targeted search for a program by name, but also by purpose and area of application.



Start Programs (alphabetical) Programs for NGS research (by application) Other bioinformatics programs (by application)

Bioinformatics programs and services available at GWDG

The following is a complete list of bioinformatics programs and services currently available at GWDG. If you need a program that is not yet installed on our servers, contact us and we will install it if possible.

Most programs are Linux based command line tools available on our bioinformatics server gwdu100 (dialog mode) or HPC cluster in batch mode. In some cases (e.g. HPC) your GWDG account has to be activated for use of the systems, please request access by email. If you need help with Linux basics or the batch system, we will give you a short introduction.

Please send all requests to support@gwdg.de.

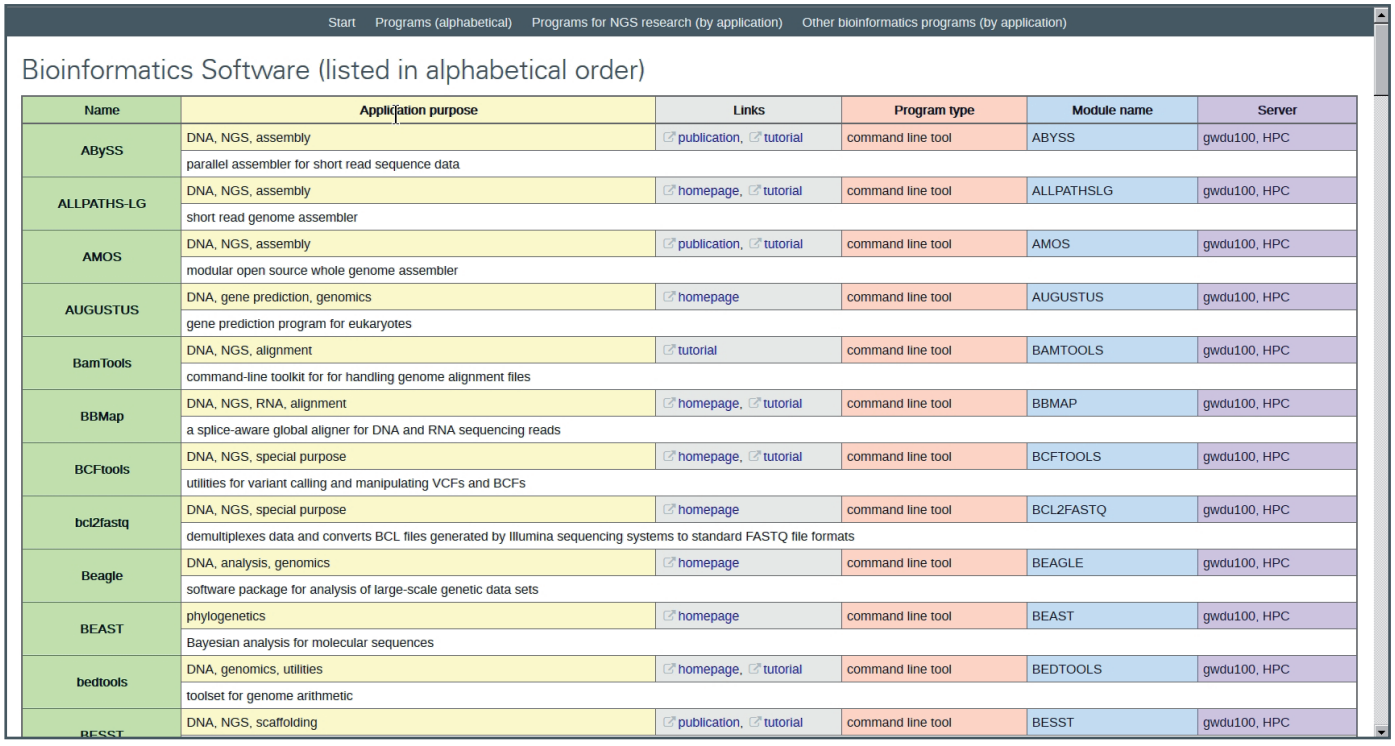
Content

- Web services
- License server
- Programs (in alphabetical order)
- Programs for NGS work (by application)
- Other bioinformatics programs (by application)

Web services

Name	Application purpose	Web server	More information
Galaxy	multi purpose	https://galaxy.gwdg.de	homepage, tutorial
	An open source, web-based platform for data intensive research. Several thousand tools (with focus on but not limited to bioinformatics) are available for installation from the Galaxy Tool Shed.		
geneXplain	multi purpose	https://gwdu100.gwdg.de/bioumlweb	homepage
	The geneXplain platform is an online toolbox and workflow management system for a broad range of bioinformatic and systems biology applications.		
MASCOT	protein mass spectrometry	https://gwdu099.gwdg.de/mascot	information/registration, homepage
	A software search engine that uses mass spectrometry data to identify proteins from peptide sequence databases.		
RStudio Server	statistical computing	https://rstudio.gwdg.de	homepage
	A free and open-source environment for R, a programming language for statistical computing.		

Abb. 1



Start Programs (alphabetical) Programs for NGS research (by application) Other bioinformatics programs (by application)

Bioinformatics Software (listed in alphabetical order)

Name	Application purpose	Links	Program type	Module name	Server
ABySS	DNA, NGS, assembly	publication, tutorial	command line tool	ABYSS	gwdu100, HPC
	parallel assembler for short read sequence data				
ALLPATHS-LG	DNA, NGS, assembly	homepage, tutorial	command line tool	ALLPATHSLG	gwdu100, HPC
	short read genome assembler				
AMOS	DNA, NGS, assembly	publication, tutorial	command line tool	AMOS	gwdu100, HPC
	modular open source whole genome assembler				
AUGUSTUS	DNA, gene prediction, genomics	homepage	command line tool	AUGUSTUS	gwdu100, HPC
	gene prediction program for eukaryotes				
BamTools	DNA, NGS, alignment	tutorial	command line tool	BAMTOOLS	gwdu100, HPC
	command-line toolkit for for handling genome alignment files				
BBMap	DNA, NGS, RNA, alignment	homepage, tutorial	command line tool	BBMAP	gwdu100, HPC
	a splice-aware global aligner for DNA and RNA sequencing reads				
BCFtools	DNA, NGS, special purpose	homepage, tutorial	command line tool	BCFTOOLS	gwdu100, HPC
	utilities for variant calling and manipulating VCFs and BCFs				
bcl2fastq	DNA, NGS, special purpose	homepage	command line tool	BCL2FASTQ	gwdu100, HPC
	demultiplexes data and converts BCL files generated by Illumina sequencing systems to standard FASTQ file formats				
Beagle	DNA, analysis, genomics	homepage	command line tool	BEAGLE	gwdu100, HPC
	software package for analysis of large-scale genetic data sets				
BEAST	phylogenetics	homepage	command line tool	BEAST	gwdu100, HPC
	Bayesian analysis for molecular sequences				
bedtools	DNA, genomics, utilities	homepage, tutorial	command line tool	BEDTOOLS	gwdu100, HPC
	toolset for genome arithmetic				
BESST	DNA, NGS, scaffolding	publication, tutorial	command line tool	BESST	gwdu100, HPC

Abb. 2

2. eine **systematische Sortierung** aller Programme, die **speziell für das NGS** gedacht sind *https://gwdu100.gwdg.de/ngs.html* (siehe Abb. 3)
3. eine **systematische Sortierung** weiterer Programme, die **nicht speziell für das NGS** gedacht sind *https://gwdu100.gwdg.de/non_ngs.html* (siehe Abb. 4)

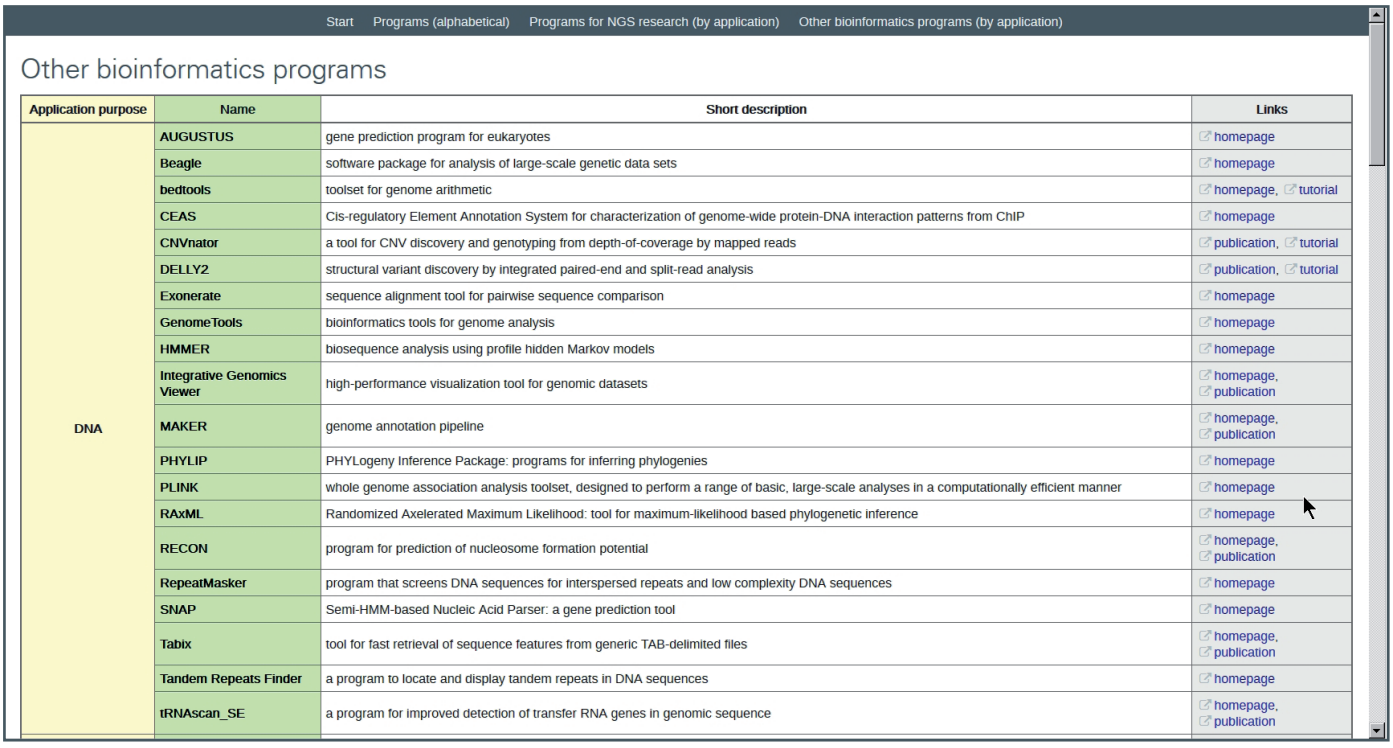
Gleichzeitig wurde den Programmen immer jeweils die besondere Eignung für bestimmte Forschungsobjekte und -gebiete zugeordnet, wie z. B. DNA, RNA, NGS, gene prediction, genomics, assembly, scaffolding, quality control, alignment, structural biology, documentation u. a. m. Selbstverständlich gibt es bei den meisten Programmen mehrere Einsatzgebiete, wobei es auch sein kann, dass wir eine Zuordnung etwas zu eng oder zu weit interpretiert haben. Wir haben die Aufteilung nach bestem Wissen vorgenommen und bitten um Rückmeldung, falls eine Zuordnung falsch ist oder eine wichtige fehlt. Diese systematische Zuordnung ist bewusst sehr grob gewählt, um die Übersichtlichkeit zu erhalten, und es empfiehlt sich deshalb dringend, mit Hilfe der weiterführenden Links (Homepage, Tutorial, Manpage etc.) genauer zu überprüfen, ob der beabsichtigte Verwendungszweck auch wirklich erfüllt werden kann. Bei jedem Programm erhält man darüber hinaus auch Informationen, auf welchem Server es zur Verfügung steht.

Start Programs (alphabetical) Programs for NGS research (by application) Other bioinformatics programs (by application)

Programs for NGS work

Application purpose	Name	Short description	Links
ChIP-Seq	MACS2	Model-based Analysis of ChIP-Seq data for identifying transcript factor binding sites	homepage
	ABySS	parallel assembler for short read sequence data	publication, tutorial
	ALLPATHS-LG	short read genome assembler	homepage, tutorial
	AMOS	modular open source whole genome assembler	publication, tutorial
	BamTools	command-line toolkit for for handling genome alignment files	tutorial
	BBMap	a splice-aware global aligner for DNA and RNA sequencing reads	homepage, tutorial
	BCFtools	utilities for variant calling and manipulating VCFs and BCFs	homepage, tutorial
	bcl2fastq	demultiplexes data and converts BCL files generated by Illumina sequencing systems to standard FASTQ file formats	homepage
	BESST	scaffolding of large fragmented assemblies	publication, tutorial
	Bowtie 1	ultrafast, memory-efficient short read aligner	homepage
	Bowtie 2	ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences	homepage
	BUSCO	quantitative measures for the assessment of genome assembly, gene set, and transcriptome completeness	homepage, publication
	Burrows-Wheeler Aligner	software package for mapping low-divergent sequences against a large reference genome	homepage
	Celera Assembler	de novo whole-genome shotgun (WGS) DNA sequence assembler	homepage
	Cutadapt	searches for the adapter in all reads and removes it when found	publication, tutorial
	deepTools	command-line tools to process and analyze deep sequencing data	homepage
	eXpress	streaming fragment assignment and quantification for high-throughput sequencing	homepage
	FastQC	quality control tool for high throughput sequence data	homepage
	FASTX	collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing	homepage
	GapFiller	stand-alone program for closing gaps within pre-assembled scaffolds	publication
	GATK	toolkit for variant discovery in high-throughput sequencing data	homepage
	Geneious	a powerful and comprehensive suite of molecular biology and NGS analysis tools with very good visualisation	homepage

Abb. 3



Start Programs (alphabetical) Programs for NGS research (by application) Other bioinformatics programs (by application)

Other bioinformatics programs

Application purpose	Name	Short description	Links
DNA	AUGUSTUS	gene prediction program for eukaryotes	homepage
	Beagle	software package for analysis of large-scale genetic data sets	homepage
	bedtools	toolset for genome arithmetic	homepage, tutorial
	CEAS	Cis-regulatory Element Annotation System for characterization of genome-wide protein-DNA interaction patterns from ChIP	homepage
	CNVnator	a tool for CNV discovery and genotyping from depth-of-coverage by mapped reads	publication, tutorial
	DELLY2	structural variant discovery by integrated paired-end and split-read analysis	publication, tutorial
	Exonerate	sequence alignment tool for pairwise sequence comparison	homepage
	GenomeTools	bioinformatics tools for genome analysis	homepage
	HMMER	biosequence analysis using profile hidden Markov models	homepage
	Integrative Genomics Viewer	high-performance visualization tool for genomic datasets	homepage, publication
	MAKER	genome annotation pipeline	homepage, publication
	PHYLIP	PHYLogeny Inference Package: programs for inferring phylogenies	homepage
	PLINK	whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner	homepage
	RAxML	Randomized Axelerated Maximum Likelihood: tool for maximum-likelihood based phylogenetic inference	homepage
	RECON	program for prediction of nucleosome formation potential	homepage, publication
	RepeatMasker	program that screens DNA sequences for interspersed repeats and low complexity DNA sequences	homepage
	SNAP	Semi-HMM-based Nucleic Acid Parser: a gene prediction tool	homepage
	Tabix	tool for fast retrieval of sequence features from generic TAB-delimited files	homepage, publication
	Tandem Repeats Finder	a program to locate and display tandem repeats in DNA sequences	homepage
	tRNAscan_SE	a program for improved detection of transfer RNA genes in genomic sequence	homepage, publication

Abb. 4

EINE KLEINE BITTE ZUM SCHLUSS

Last, but not least, bitten wir alle Kunden, uns Vorschläge für die Installation weiterer Software zu machen (per E-Mail an *support@gwdg.de*), wenn sie diese unbedingt brauchen oder glauben, dass diese auch für andere User besonders nützlich ist.

Neue Speicherstrategien für die gemeinsamen Laufwerke

Text und Kontakt:
Katrin Hast
katrin.hast@gwdg.de
0551 201-1808
Bjørn Nachtwey
bjoern.nachtwey@gwdg.de
0551 201-2181

Beobachtungen und einfache Auswertungen zeigen ohne große Überraschung, dass die im Rechenzentrum der GWDG gespeicherte Datenmenge exponentiell wächst; etwa alle zwei Jahre verdoppelt sich das Datenvolumen. Aus diesem Grunde sind neue Strategien zur Nutzung der unterschiedlichen Speicherbereiche erforderlich. In diesem GWDG-Nachrichten-Artikel soll über die aktuellen Probleme bei der Datenhaltung und der Speichernutzung sowie mögliche Lösungen berichtet werden. Eine kurze Einführung in die Datenschutzgrundverordnung (DSGVO) schließt den Artikel ab.

EINLEITUNG

Gegen Ende der 1990er-Jahre verfügten PCs in der Regel über Festplatten mit einer Größe von wenigen GByte, während heute Festplatten mit einem Volumen im TByte-Bereich üblich sind. Innerhalb eines Zeitraums von etwa 20 Jahren ist also das verfügbare Speichervolumen um einen Faktor von 1.000 gewachsen. Allerdings ist die Entwicklung bei der Speicherkapazität von Festplatten in den vergangenen Jahren deutlich verlangsamt. Die technischen Fortschritte erlauben nur noch ein langsames lineares Wachstum, die Zeiten der exponentiellen Fortschritte bei Festplattenvolumina sind vorbei. SSDs werden Festplatten als Speichermedium zunehmend ablösen, und hier ist mit hohen Zunahmen bei der Speicherkapazität sowie aber auch der Kosten zu rechnen.

Für die GWDG entsteht somit zum einen das Problem, dass das exponentielle Datenwachstum anhält, die Speichermedien aber nicht mehr so leicht in einem finanzierbaren Maß mitwachsen, so dass mehr auf die Kosten von Speichersystemen geachtet werden muss, und zum anderen das Problem, dass immer größere Datenmengen z. B. beim Backup verarbeitet werden müssen. Beide Fragen werden nachfolgend aufgegriffen.

ASPEKTE DER DATENVERWALTUNG

Backup & Restore

Das exponentielle Wachstum der Datenbereiche hat dazu geführt, dass in einigen Bereichen das Backup nicht mehr innerhalb von 24 Stunden durchgeführt werden kann und somit ein tägliches Backup zunehmend problematischer wird. Aber nicht nur die Gesamtmenge der Daten ist für die Backupsoftware problematisch, sondern auch der Umstand, dass es sich in einigen Fällen um sehr viele, sehr kleine Dateien handelt. Dies belastet die Backupsysteme noch zusätzlich.

Die Schattenkopien, eine Funktion einiger Speichersysteme, können mit diesem Problem deutlich besser umgehen. Besonders gut ist, dass die Nutzer die verlorenen Daten selber wiederherstellen können. Der Nachteil ist, dass die Schattenkopien auf demselben Speichersystem liegen wie die Originaldaten. Dies kann bei einem Totalverlust des Speichersystems auch zum Verlust der Daten führen. Deshalb können die Schattenkopien ein Backupsystem nie komplett ersetzen.

Wirtschaftlichkeit

Die Kosten eines Speichersystems richten sich nach Qualität, Kapazität und Funktionen des Systems. So ist man bestrebt, Daten in Kategorien einzuteilen und dem entsprechenden passenden Speichersystem zuzuordnen. So werden z. B. schon jetzt Daten, die nicht mehr im regelmäßigen Zugriff sind, archiviert. Hier sind die Kosten für die Datenhaltung deutlich geringer. Ein Grund dafür ist, dass archivierte Daten nur auf Magnetbändern gespeichert werden, die keine Energiekosten verursachen. In Zukunft sollen weitere Klassifikationen eingeführt werden; dazu später mehr.

Die Preise bzw. Bewertungen in Milli Arbeitseinheiten (mAE) für die drei Speicherbereiche Fileservice (einschließlich Backup),

New Strategies for Data Storage Areas

Observations and simple evaluations show, without much surprise, that the amount of data stored in the GWDG data center grows exponentially. About every two years, the data volume doubles. For this reason, new strategies for using the different storage areas are required. This GWDG news article aims to describe the current problems of data storage areas and will present some ideas about possible solutions. A short introduction to the General Data Protection Regulation (GDPR) concludes the article.

Backup und Archiv für 1 GByte pro Monat sind der folgenden Tabelle zu entnehmen. In den Preisen bzw. Bewertungen sind Bestandteile wie Strom, Wartung und Ersatzbeschaffungen enthalten.

SPEICHERBEREICH	mAE
Fileservice (einschließlich Backup)	1,30
Backup	0,27
Archiv	0,80

Kürzlich wurde für die gemeinsamen Laufwerke (Gruppenlaufwerke von Instituten) ein NetApp-Speichersystem angeschafft. Das System beinhaltet brutto 2 * 400 TByte redundant gespiegelte Speicherkapazität und wurde für mehrere 100.000 Euro beschafft. Hierbei sei erwähnt, dass sich diese Speichersysteme in Funktion und Qualität deutlich von den herkömmlichen Festplatten, die man in den üblichen bekannten Geschäften kaufen kann, unterscheiden. Die großen Vorteile liegen u. a. in einem ausgereiften Betriebssystem, das nicht nur die Speicherkapazität über Standard-Netzwerkprotokolle wie CIFS/SMB und NFS bereitstellen kann, sondern auch über mächtige interne Werkzeuge zur Datendeduplikation, Datenmigration, Datenspiegelung und Datensicherung verfügt. Auch ist die gesamte Hardware bis zu den Netzwerkadaptern hin auf hohe Leistungsfähigkeit hin ausgelegt. Nach den ersten praktischen Erfahrungen mit den Gruppenlaufwerken ermöglicht allein die Datendeduplikation eine Speicherplatzeinsparung von ca. 40 %, was letztlich einer entsprechenden Kapazitätserhöhung gleichkommt.

Bei einem möglichen zukünftigen weiteren Ausbau der NetApp-Umgebung können mit internen Datenmigrationswerkzeugen große Datenmengen weitaus schneller bewegt werden, als es mit traditionellen Kopiervorgängen zwischen Dateisystemen möglich wäre.

Speichersysteme

Die Eignung von Speichersystemen ist abhängig von den Funktionen, die ein Speichersystem mitbringt. Zum Beispiel werden bei der Archivierung die Daten auf ein Band gespeichert. Das ist besonders wirtschaftlich im Unterhalt, aber nicht für die tägliche Verwendung der Daten geeignet, da ein Bandroboter die Daten von den Bändern erst wieder verfügbar machen muss und somit der Zugriff für den Nutzer deutlich länger dauern würde als bei einem anderen Speichersystem.

In der GWDG verwenden wir für die gemeinsamen Dateibereiche zwei Speichersysteme: die *NetApp* und *StorNext*. Wie Sie und wir aus leidvoller Erfahrung feststellen mussten, eignet sich StorNext nicht für den hochfrequenten Nutzerzugriff unter Windows, ist aber aufgrund der extrem flexiblen Speicherverwaltung sehr gut für besonders große und auch stark wachsende Speicherbereiche geeignet, wie sie z. B. bei Messrechnern oder bilderzeugenden Verfahren anfallen können.

Die NetApp ist im Gegensatz dazu besonders für einen hochfrequenten Datenzugriff geeignet. Dies wird in besonderer Weise auch durch die Funktion der Schattenkopien unterstützt. Die Schattenkopien bieten den Vorteil, dass die Nutzer ihre Daten selber wiederherstellen können und die Intervalle der Erzeugung von Datenkopien besonders kurz gestaltet werden können. Um nun die Datenhaltung zu optimieren, ist es sinnvoll, die Daten zu klassifizieren und entsprechend ihrer Anforderungen auf einem adäquaten Speichersystem zu hosten. Die Aufteilung der Daten in verschiedene Speichersysteme hat auch den Vorteil, dass Backupintervalle dem Bedarf angepasst werden können. Das ist einer der Punkte, an dem wir ansetzen wollen, um künftig die Speichernutzung zuverlässig sicherzustellen.

KLASSIFIZIERUNG VON DATEN

Ein wichtiger Faktor, nach dem eine Datei klassifiziert werden kann, ist z. B. der Zeitpunkt, an dem das letzte Mal auf eine Datei zugegriffen wurde. Statistiken besagen, dass bis zu 60 % der Daten drei Jahre oder länger nicht genutzt wurden. Stichproben in den gemeinsamen Laufwerken bestätigen diese Statistiken. In diesem Fall sollte die erste Überlegung des lokalen Administrators sein, ob die Daten eventuell gelöscht oder archiviert werden können. (Leider braucht man zur Erstellung von Altersstatistiken für Datenbereiche eine spezielle und sehr teure Software, die der GWDG nur testweise zur Verfügung stand.)

Falls weder das Löschen der Daten noch das Archivieren in Frage kommen, bietet die GWDG als neuen Service ein weiteres Laufwerk an, in dem wenig genutzte Daten gespeichert werden können. Dieses zusätzliche Laufwerk soll aber keinesfalls das Archivieren von Daten ersetzen!

Die Verwendung von getrennten Speichersystemen für unterschiedliche Daten hat mehrere Vorteile:

- Die Menge der aktuellen und produktiven Daten ist deutlich kleiner, was dazu führt, dass der Datenbereich für die Nutzer insgesamt übersichtlicher wird. Daten werden schneller wiedergefunden.
- Es können auf die Nutzung und den Bedarf abgestimmte Backupstrategien verwendet werden. Durch die kleineren Datenbereiche der aktuell produktiven Daten kann das Backupintervall verkürzt werden.
- Zusätzlich sorgen Snapshots, die bis zu neun Mal täglich ausgeführt werden, für die Sicherheit der Daten. Die Schattenkopien bieten den Vorteil, dass der Nutzer seine Daten selber wiederherstellen kann. Man findet die Schnittstelle im Kontextmenü des Ordners mit dem Datenverlust unter der Registerkarte „Vorgängerversion“.
- Entsprechend wird in dem zusätzlichen Speichersystem das Backup auf das notwendige Maß reduziert. Geplant ist hier ein wöchentliches Backup mit TSM. Dieses Speichersystem bietet keine Schattenkopien.
- Durch die so entstehende Verkleinerung der Datenbereiche bei den bisherigen Gruppenlaufwerken können für den dortigen hochfrequenten Datenzugriff qualitativ hochwertigere und damit teurere Speichersysteme eingesetzt werden. Die Datenhaltung wird wirtschaftlicher.

Der zusätzliche Speicherbereich ist vom Pfadnamen dem aktuellen gemeinsamen Laufwerk sehr ähnlich, hier aufgezeigt am Beispiel des Departments für Agrarökonomie und Rurale Entwicklung der Universität Göttingen:

Aktueller Pfad für die gemeinsamen Laufwerke:
\\wfs-agrar.top.gwdg.de\uaao-all$\UAAO100
Zusätzlicher Pfad für spezielle Daten:
\wfs-agrar-spezial.top.gwdg.de\uaao-all$\UAAO100

Die jeweiligen zuständigen Administratoren können den zusätzlichen Speicherbereich per E-Mail an *support@gwdg.de* oder alternativ über die Webschnittstelle *https://support.gwdg.de* anfordern.

INFORMATIONEN ZUR SPEICHERNUTZUNG

Die Nutzung von Speicherbereichen wird durch technische Rahmenbedingungen reglementiert, die nicht allen Anwendern bekannt sind. Im Folgenden sind einige der häufigsten Probleme bei der Nutzung der Daten aufgelistet:

- Die Verwendung der Daten durch verschiedene Betriebssysteme führt immer wieder zu dubiosen Fehlern, deren Behebung zeitaufwendig oder oft auch nicht möglich ist. Dies ist besonders im Zusammenhang mit Mac- und Windows-Betriebssystemen aufgefallen. So können zum Beispiel Mac-Systeme die Benennung von Dateien mit Sonderzeichen ermöglichen, die bei einem Windows-Betriebssystem nicht zulässig sind.
- Der Umgang mit den NTFS-Zugriffsrechten ist sehr komplex und keinesfalls einfach zu verstehen. Wir haben deshalb Standards entwickelt, die wir bei Übergabe eines neuen Speicherbereichs setzen und als Empfehlung an die lokalen Administratoren weitergeben. Allerdings stellen wir immer wieder fest, dass von diesen Standards abgewichen wird. Besonders problematisch ist es, wenn die GWDG-Administratoren aus den Berechtigungen entfernt werden. Dann ist im Supportfall eine Unterstützung erstmal nur bedingt möglich. Abgesehen davon kann ohne die Zugriffsmöglichkeit auf die Daten auch kein Backup erfolgen.
- Problematisch ist ebenfalls, wenn besonders tiefe Ordnerpfade und/oder Dateinamen verwendet werden. Der sogenannte *fully qualified file name* setzt sich aus dem eigentlichen Dateinamen und dem vollständigen Pfad der Verzeichnisse zusammen. Dieser Pfad darf nicht länger als 260 Zeichen sein. Dies ist eine durch das Windows-System vorgegebene Grenze, die seit Windows 10 anpassbar ist. Verwendet man mehr Zeichen, kann das z. B. dazu führen, dass Dateien oder Ordner nicht mehr gelöscht werden können. Es wird empfohlen, die Ordnerpfade nicht so tief und die Ordner- und Dateinamen möglichst kurz zu gestalten.

DATENSCHUTZGRUNDVERORDNUNG (DSGVO)

Die ab dem 25. Mai 2018 geltende DSGVO ist eine Verordnung der Europäischen Union, die die Verarbeitung personenbezogener Daten durch private Unternehmen und öffentliche Stellen regelt. Die DSGVO hat unmittelbar Auswirkungen auf alle von der GWDG ihren Kunden angebotenen Dienste. Allerdings obliegen auch den Nutzern selbst Pflichten beim Umgang mit Daten, so dass wir hier auf einige Aspekte mit Bezug zu den „gemeinsamen Datenbereichen" eingehen wollen.

In der DSGVO werden sechs Grundprinzipien des Datenschutzes genannt, die vom jeweils „Verantwortlichen" einzuhalten sind. Mit der DSGVO sind sowohl der Nutzer als auch die GWDG als Auftragsverarbeiter in dieser Verantwortlichen-Rolle. Allerdings besitzt die GWDG meist keine ausreichenden Kenntnisse über die gespeicherten Daten, so dass die Einhaltung der Grundprinzipien teilweise durch die Nutzungsordnung der GWDG geregelt wird.

Die Grundprinzipien lassen sich mit den folgenden Schlagworten beschreiben:

- „Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz" – personenbezogene Daten dürfen nur gespeichert und verarbeitet werden, wenn dies zulässig ist. Die Einhaltung dieses Prinzips kann nur durch die Nutzer sichergestellt werden, da die GWDG keinerlei diesbezügliche Informationen zu den gespeicherten Daten hat.
- „Zweckbindung" – personenbezogene Daten dürfen nur zu den zuvor benannten Zwecken gespeichert und verarbeitet werden. Der Nutzungszweck ist der GWDG auch nicht bekannt, so dass eine Abweichung vom ursprünglichen Zweck für die GWDG nicht zu erkennen ist.
- „Datenminimierung" – es dürfen nur jene Daten verarbeitet werden, die benötigt werden. Die Festlegung zum Umfang der benötigten Daten ist der GWDG nicht möglich.
- „Richtigkeit" – gespeicherte und verarbeitete personenbezogene Daten müssen richtig sein. Die GWDG stellt durch technische und organisatorische Maßnahmen (TOM) sicher, dass keine Manipulation der Daten stattfindet bzw. diese wieder zurückgenommen werden kann. Eine inhaltliche Bewertung muss aber durch den Nutzer erfolgen.
- „Speicherbegrenzung" – personenbezogene Daten dürfen nur so lange gespeichert werden, wie es nötig ist. Die GWDG unterstützt die Einhaltung dieses Prinzips dahingehend, dass im Rahmen eines Löschkonzeptes offensichtlich inaktive Daten gelöscht werden. Ebenso unterliegen im Betrieb anfallende Logdaten dem Löschkonzept.
- „Integrität und Vertraulichkeit" – die betroffenen Personen erwarten, dass ihre Daten sicher und vertraulich verarbeitet werden und es einen Schutz gegen unbefugten Zugriff gibt. Die GWDG schützt die Daten im Rahmen der TOM. Allerdings obliegt der Schutz auch dem Nutzer, insbesondere bei sogenannten „Daten besonderer Kategorien", also besonders schützenswerten Daten.

Die TOM an dieser Stelle vollständig darzustellen und alle weiteren Schutzmechanismen aufzuzeigen, ließe den Umfang dieses Artikels um das Mehrfache wachsen; daher nur ein paar Beispiele:

- Der physische Zugang zu den Daten wird durch den kontrollierten und protokollierten Zugang zu den Maschinenräumen geschützt,; nur ausgewählte Personen der GWDG erhalten überhaupt Zugang.
- Ausgemusterte bzw. defekte Datenträger, die möglicherweise personenbezogene Daten enthalten, werden durch einen Dienstleister zerstört, der gemäß der DIN 66399 zertifiziert ist. Auch gilt für alle neu zu beschaffenden Systeme, dass defekte Datenträger nicht mehr zurückgegeben werden.
- Zugriffe auf Daten, die sich in den gemeinsamen Datenbereichen befinden, werden durch ACLs, also Zugriffsrechte beschränkt. Diese sind durch die dezentralen Administratoren nur in dem ihnen zugewiesenen Bereich veränderbar.

Da aber auch der Nutzer letztendlich für das Einhalten des Datenschutzes in der Pflicht steht, hier noch ein paar Hinweise:

- Die Nutzungsordnung der GWDG sieht bereits vor, dass besonders schützenswerte Daten nicht auf den GWDG-Systemen verarbeitet werden dürfen.

- Durch den Einsatz von Verschlüsselung auf Nutzerseite kann die Forderung aus der Nutzungsordnung erreicht werden, da die verschlüsselten Daten aus Sicht der GWDG nicht mehr zu den „besonderen Kategorien personenbezogener Daten“ gehören. Die DSGVO benennt in Artikel 32 Verschlüsselung auch explizit als probates Mittel zum Schutz personenbezogener Daten.
- Ebenso kann durch Pseudonymisieren, also das Ersetzen von personenbezogene Daten durch eindeutige Namen/Schlüssel, die aber nicht mehr der betroffenen Person zugeordnet werden können, ein weitreichender Schutz erreicht werden. Bei der Pseudonymisierung muss nur noch die Zuordnungstabelle besonders geschützt werden.

Sollten Sie weitere Fragen oder Anregungen zu dem Thema haben, verwenden Sie bitte unser Ticketsystem, zu erreichen unter *support@gwdg.de* oder *https://support.gwdg.de*.

Hands-on-Workshop „Data Mining“

Text und Kontakt:
Dr. Sven Bingert
sven.bingert@gwdg.de
0551 201-2164
Dr. Christian Köhler
christian.koehler@gwdg.de
0551 201-2193

Um eigenhändig Erfahrungen mit den Zukunftsthemen der Informatik zu sammeln, nahmen Studierende der Dr. Hans Riegel-Stiftung vom 1. bis 4. März 2018 an einem Workshop in Zusammenarbeit mit dem XLAB Göttingen teil. Am 3. März waren sie zu Gast am Institut für Informatik der Universität Göttingen und erhielten dort bei einem Vortrag von Prof. Dr. Ramin Yahyapour Einblicke in die Herausforderungen von Big Data und Data Analytics. Im Anschluss konnten die Studierenden in praktischen Übungen mittels der Software *Orange Data Mining* eigene Data-Mining- und Data-Analytics-Projekte durchführen.

PROGRAMM DES WORKSHOPS

Im März 2018 waren 20 ehemalige Sieger der Dr. Hans Riegel-Fachpreise [5] für die Teilnahme an einem Workshop zu den Zukunftsthemen der Informatik zu Besuch in Göttingen. Der Workshop beinhaltete u. a. Beiträge von Prof. Dr. Florentin Andreas Wörgötter (Universität Göttingen), Prof. Dr. Richard Neher (Max-Planck-Institut für Entwicklungsbiologie), der Schumann GmbH und von Prof. Dr. Ramin Yahyapour (GWDG) und wurde zusammen mit dem XLAB Göttingen [7] organisiert. Das Programm des Workshops ging über vier Tage und umfasste spannende Themen wie z. B. Robotik und die Echtzeitanalyse von Genomdaten von Influenzaviren. Am dritten Tag des Workshops waren die Teilnehmer zu Gast bei Prof. Yahyapour im Institut für Informatik und lernten dort einiges über die aktuellen Herausforderungen zum Thema Big Data und Data Analytics. Anschließend konnten sie in verschiedenen praktischen Übungen selbst Datenanalyse z. B. mittels maschinellem Lernen durchführen.

Jeder von uns hinterlässt in seiner täglichen Arbeit oder durch seine Kommunikation mittels digitaler Medien eine große Spur an kostbaren Daten. In vielen Bereichen werden diese Daten in ausgereiften Algorithmen verwendet, um unser Leben zu vereinfachen. Klassische Beispiele sind dedizierte Kaufvorschläge oder Playlisten bei verschiedenen Plattformen. Aber auch Gesichtserkennung oder die Auswertung der Kommunikation (z. B. Twitter-Tweets) sind aktuelle Themen.

Die GWDG hatte für die Teilnehmer des Workshops genau diese Themen aufgegriffen und zeigte, dass auch ohne große IT-Kenntnisse, aber mit den passenden Werkzeugen und den richtigen Fragestellungen, spannende Analysen betrieben werden können. Die Teilnehmer lernten u. a., aktuelle Twitter-Tweets oder Wikipedia-Artikel selbst zu beschaffen und verschiedene Analysen, z. B. zum Sentiment, durchzuführen. Dabei wurde den Teilnehmern offen gelassen, zu welchen Themen die Analyse durchgeführt werden sollte, sodass in der Vorstellung der Ergebnisse dann sowohl aktuelle politische Themen als auch Untersuchungen zur Dr. Hans Riegel-Stiftung präsentiert wurden. Für die Analyse

Hands-on Workshop “Data Mining“

In order to gain personal experience on the future topics of computer science, students of the Dr. Hans Riegel-Stiftung visited a workshop from March 1st to 4th 2018 in cooperation with the XLAB Göttingen. On March 3rd, they were invited to the Institute of Computer Science at Göttingen University, where they learned about the challenge of big data and data analytics in a lecture by Prof. Dr. Ramin Yahyapour. Afterwards, the students were able to carry out their own data mining and data analytics projects in practical exercises using the software *Orange Data Mining*.

wurde das Programm *Orange Data Mining* verwendet, für das schon viele Algorithmen und Schnittstellen zu Datenquellen implementiert sind. Dadurch konnte der Fokus der Veranstaltung auf die Anwendung gelegt werden, ohne dass fundierte Kenntnisse zu z. B. maschinellem Lernen oder Schnittstellen notwendig waren. Am Ende des Workshops wurde dann noch gezeigt, wie Gesichtserkennung in Bilddaten funktioniert und die erkannten Gesichter durch Ähnlichkeitsanalyse mit anderen Bilddaten verglichen werden können. Dieses Verfahren wird u. a. in aktuellen Mobiltelefonen als Authentifizierungsmethode angeboten.

ORANGE DATA MINING

Die Entscheidung für die Software *Orange Data Mining* [1] wurde aufgrund der bisherigen guten Erfahrungen, u. a. im Rahmen der GWDG-Teilnahme an der IdeenExpo 2017, getroffen. Vielfältige Möglichkeiten für den Import, die Weiterverarbeitung und Visualisierung sowie den Export eines Datensatzes werden dabei in Form sogenannter *Widgets* in einer grafischen Benutzeroberfläche zur Verfügung gestellt und können miteinander zu *Workflows* kombiniert werden. Durch diesen intuitiven Ansatz ist auch die Zugänglichkeit für Schüler und Studierende verschiedener Fachrichtungen gleichermaßen sichergestellt.

Ein einfaches Beispiel für einen Workflow ist in Abb. 1 dargestellt: *File* importiert einen im Dateisystem vorhandenen Datensatz, etwa eine CSV-Datei. Das Widget ist mit *Data Info* verbunden, so dass in der entsprechenden Detailansicht grundlegende Informationen über den Datensatz, bspw. dessen Umfang, angezeigt werden können. Die Daten werden weiterhin an eine interaktive Tabellenansicht *Data Table* weitergegeben. Diese gibt zusätzlich Informationen über die vom Anwender ausgewählten Daten an den angeschlossenen *Scatter Plot* weiter, so dass diese im Plot hervorgehoben werden. Schließlich wird die getroffene Auswahl durch *Save Data* abgespeichert.

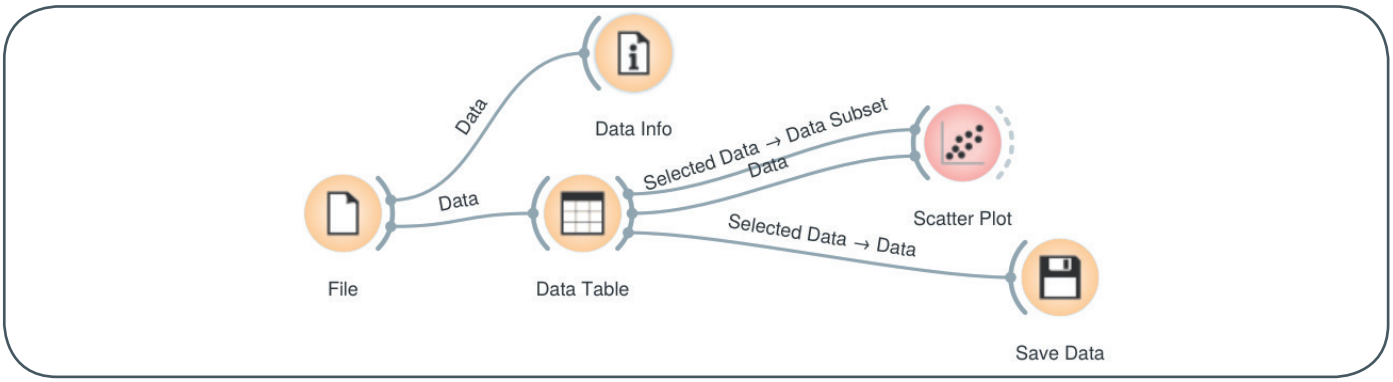
Für die Gruppen paarweise zusammenarbeitender Teilnehmer wurde je ein GWDG Cloud Server [2] vorbereitet. Vom CIP-Pool des Instituts für Informatik aus konnten dann die darauf installierten Orange-Instanzen per SSH mit X-Forwarding verwendet werden. Mit Hilfe der verwendeten Python-Paketverwaltung *Conda* [3] war die Installation der Umgebung und ihrer Abhängigkeiten jedoch auch auf dem eigenen System (Windows/Mac/Linux) einfach möglich.

BEISPIEL: KLASSIFKATION VERSCHIEDENER SCHWERTLILIEN-SPEZIES

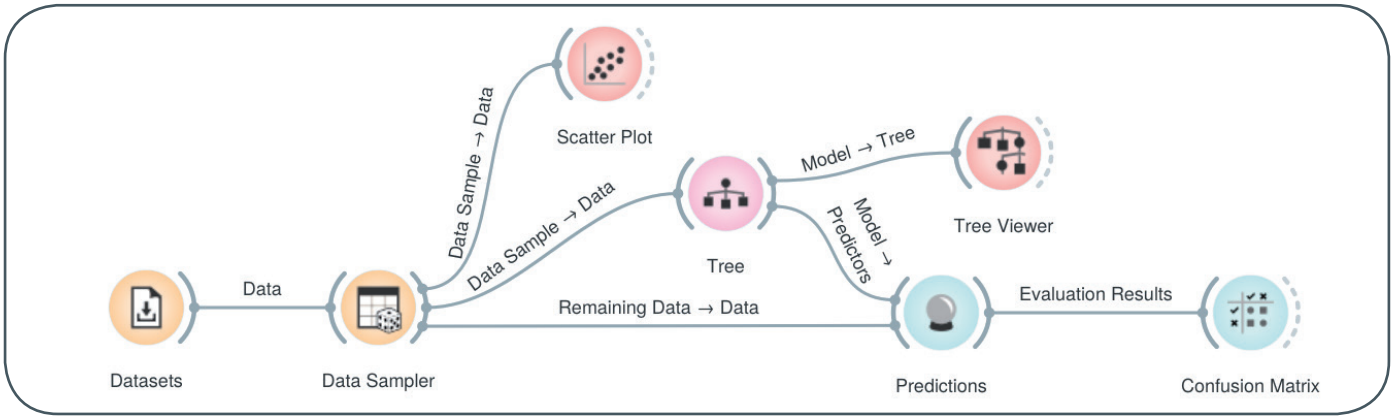
Eine der im Rahmen des Workshops bearbeiteten Aufgaben befasste sich mit überwachtem maschinellen Lernen auf Basis des klassischen *Iris*-Datensatzes von 1936 [4]. Für 150 Exemplare der Schwertlilie (*Iris*) sind darin jeweils die Länge und Breite (*length/width*) des Blütenblattes (*petal*) und des Kelchblattes (*sepal*) in cm sowie die jeweilige Spezies (*Iris-setosa*/*Iris-virginica*/*Iris-versicolor*) angegeben. Dabei wurden je Spezies 50 Exemplare vermessen.

In Abb. 2 ist der Orange-Workflow dargestellt, der zur Lösung dieser Aufgabe verwendet wurde: Auf den mitgelieferten *Iris*-Datensatz kann mit *Datasets* zugegriffen werden. Die Daten werden zunächst mit *Data Sampler* aufgeteilt: Hier werden 70 % des Datensatzes (was 105 Exemplaren entspricht) zufällig ausgewählt, als *Data Sample* weitergegeben und mit *Scatter Plot* visualisiert. Wie in Abb. 3 erkennbar ist, lässt sich auf Basis der Abmessungen des Blütenblattes (*petal*) bereits eine brauchbare Klassifikation (Vorhersage der Spezies) vornehmen, da die Datenpunkte in dieser Projektion größtenteils separiert sind.

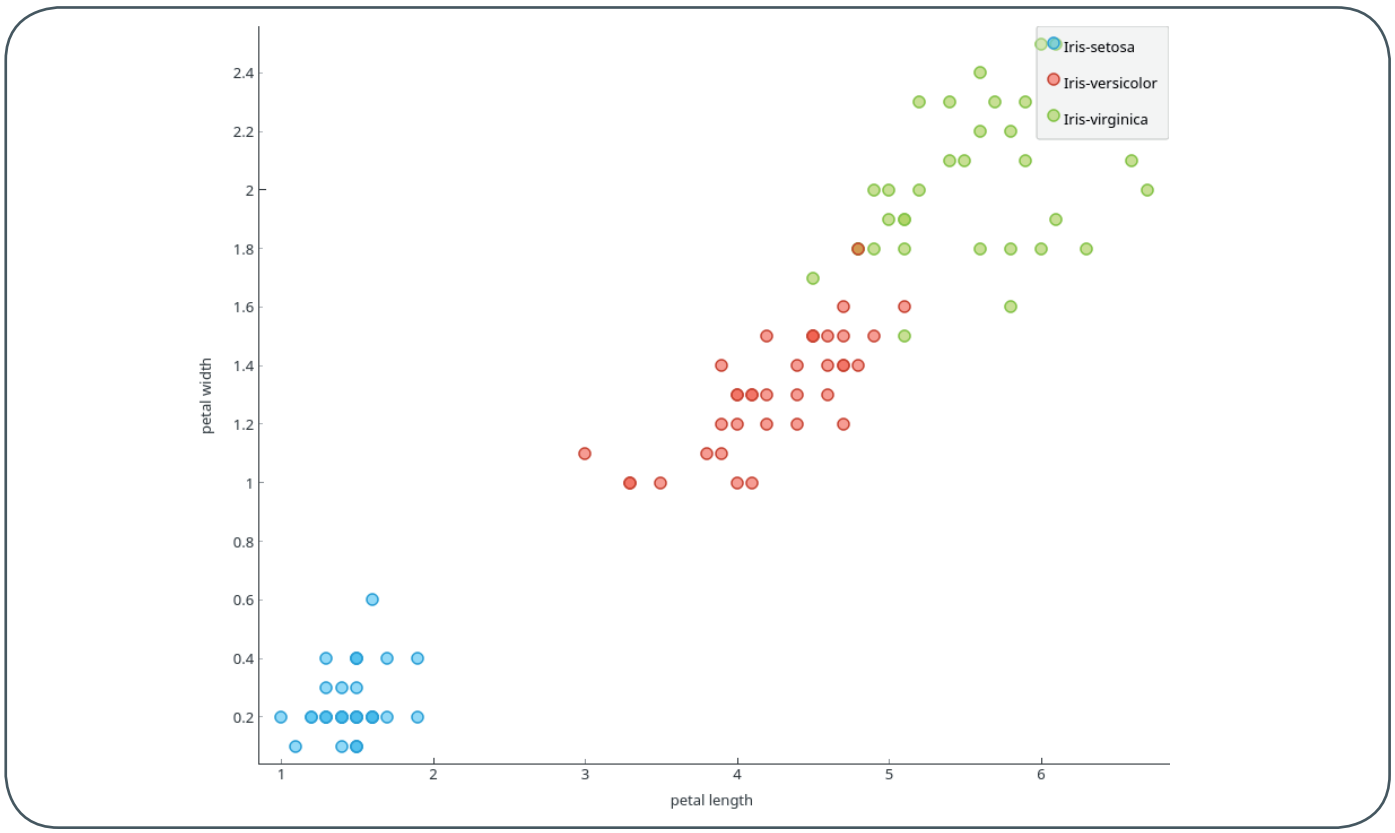
Die Trainingsdaten aus dem *Data Sample* werden weiterhin



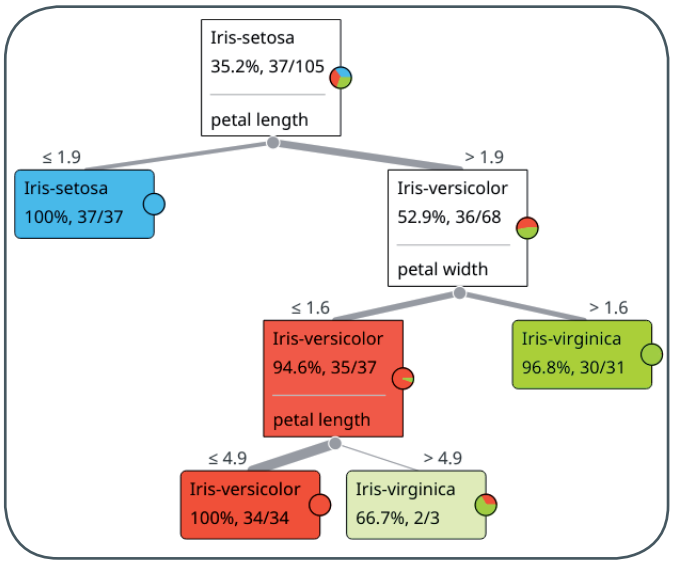
1_Grundlegender Workflow in Orange Data Mining



2_Workflow für beaufsichtigtes Lernen



3_Workflow für beaufsichtigtes Lernen: Plot der Abmessungen des Blütenblattes (petal)



4_Workflow für beaufsichtigtes Lernen: Entscheidungsbaum zur Klassifikation der Spezies

verwendet, um einen durch *Tree* symbolisierten Entscheidungsbaum zu trainieren, der in Abb. 4 dargestellt ist: Beginnend mit der Menge aller Datensätze lautet bei einem zufällig gewählten Exemplar die erste Vermutung, dass es sich mit 35,2 % Wahrscheinlichkeit um *Iris-setosa* handelt, da diese Spezies mit 37 von 105 Exemplaren die größte Gruppe darstellt. Wird nun nach der Länge des Blütenblattes (*petal length*) unterschieden, so ist bei einem Wert ≤ 1,9 cm die Vermutung bestätigt, da die 37 unter dieser Bedingung verbleibenden Exemplare alle der gleichen Spezies angehören. Für die 68 verbliebenen Exemplare ist mit einem Anteil von 52,9 % *Iris-versicolor* die größte Untergruppe und anhand der Breite des Blütenblattes (*petal width*) kann die Klassifikation fortgesetzt werden.

Auf diese Weise wird der Datensatz rekursiv in Untergruppen zerlegt, bis man bei den vorgegebenen Grenzen angelangt ist. In diesem Fall wurde die Baumtiefe dadurch limitiert, dass keine Gruppen mit weniger als fünf Exemplaren angelegt werden sollten.

Als *Remaining Data* werden nun die nicht für das Training verwendeten 45 Exemplare verwendet, um mit *Predictions* die durch den Entscheidungsbaum getroffenen Vorhersagen bzgl. der jeweiligen Spezies zu verifizieren. Das Widget *Confusion Ma-trix* gibt schließlich Auskunft darüber, wie viele Exemplare korrekt klassifiziert wurden. Im vorliegenden Beispiel wurde lediglich ein Exemplar *Iris-versicolor* fälschlicherweise als *Iris-virginica* klassifiziert, was plausibel erscheint, da es aus dem in Abb. 3 erkennbaren Übergangsbereich dieser beiden Spezies stammt. Eine genauere Klassifikation durch mehr Entscheidungsregeln, die auch diesen Bereich weiter auflösen würden, muss gegen die Maximaltiefe des Entscheidungsbaumes abgewogen werden.

LINKS

[1] Orange Data Mining: *http://orange.biolab.si/*
[2] GWDG Cloud Server: *https://www.gwdg.de/de/server-services/gwdg-cloud-server*
[3] Conda – Package, dependency and environment management for any language: *https://conda.io/*
[4] Iris Data Set: *https://archive.ics.uci.edu/ml/datasets/Iris*
[5] Dr. Hans Riegel-Stiftung: *http://www.hans-riegel-stiftung.com*
[6] Beitrag auf den Seiten der Dr. Hans Riegel-Stiftung: *http://www.hans-riegel-stiftung.com/news-detail-seite/news/einblicke-in-die-informatik/?tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Baction%5D=detail&cHash=d96da5036314255a084385dbcf01c27c*
[7] XLAB Göttingen: *http://www.xlab-goettingen.de*

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Bewerbungs-/Registrierungssysteme

WIR UNTERSTÜTZEN SIE IN IHRER ORGANISATIONSARBEIT!

Ihre Anforderung

Sie möchten ein Bewerbungs- oder allgemeines Registrierungsverfahren durchführen, z. B. für offene Stellen oder Tagungsmanagement. Bewerber sollen sich online bewerben und automatisiert per E-Mail benachrichtigt werden können. Gutachter sollen über das WWW auf die Bewerbungen bzw. Registrierungen zugreifen und Bewertungen online einstellen können.

Unser Angebot

Wir erstellen Ihnen nach Ihren Wünschen eine Lotus-Notes-Datenbank, die allen Kandidaten oder Registranten über einen Webbrowser offen steht. Die eingereichten Dokumente können aber nur von ausgewählten Gutachtern über das WWW und von speziellen Bearbeitern Ihres Instituts eingesehen, bearbeitet oder bewertet werden. Die Eingänge werden nach Ihren Kriterien sortiert und dargestellt. Weitere Workflows sind individuell gestaltbar.

Ihre Vorteile

> Leistungsfähiges ausfallsicheres System zum Aufnehmen von Bewerbungen oder Registrierungen über das WWW

> Datenschutzgerechte Speicherung und Verarbeitung der Daten

> Die Verteilung der Unterlagen auf Papier ist überflüssig, da der Zugriff der Gutachter oder sonstigen Bearbeiter über das WWW erfolgt.

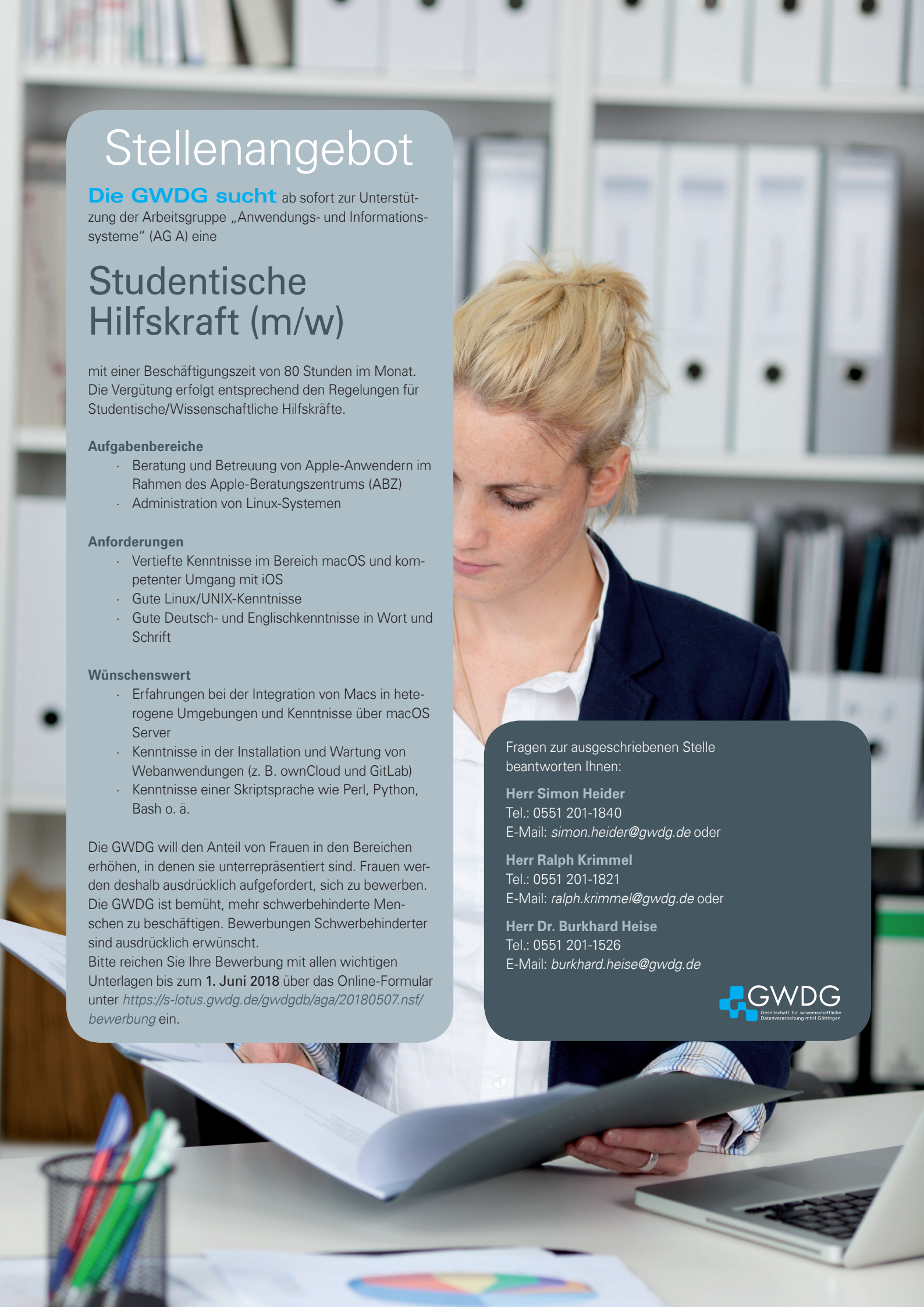
> Jeder Workflow ist an Ihre Situation anpassbar.

> Kandidaten können automatisiert per E-Mail benachrichtigt werden (z. B. Absagen).

Interessiert?

Der Service wie auch die individuelle Beratung können über support@gwdg.de angefordert werden. Nähere Informationen zum Workflow Management mit der Lotus-Software von IBM sind auf der u. g. Webseite zu finden.

www.gwdg.de/lotus



Stellenangebot

Die GWDG sucht ab sofort zur Unterstützung der Arbeitsgruppe „Anwendungs- und Informationssysteme“ (AG A) eine

Studentische Hilfskraft (m/w)

mit einer Beschäftigungszeit von 80 Stunden im Monat. Die Vergütung erfolgt entsprechend den Regelungen für Studentische/Wissenschaftliche Hilfskräfte.

Aufgabenbereiche

- Beratung und Betreuung von Apple-Anwendern im Rahmen des Apple-Beratungszentrums (ABZ)
- Administration von Linux-Systemen

Anforderungen

- Vertiefte Kenntnisse im Bereich macOS und kompetenter Umgang mit iOS
- Gute Linux/UNIX-Kenntnisse
- Gute Deutsch- und Englischkenntnisse in Wort und Schrift

Wünschenswert

- Erfahrungen bei der Integration von Macs in heterogene Umgebungen und Kenntnisse über macOS Server
- Kenntnisse in der Installation und Wartung von Webanwendungen (z. B. ownCloud und GitLab)
- Kenntnisse einer Skriptsprache wie Perl, Python, Bash o. ä.

Die GWDG will den Anteil von Frauen in den Bereichen erhöhen, in denen sie unterrepräsentiert sind. Frauen werden deshalb ausdrücklich aufgefordert, sich zu bewerben. Die GWDG ist bemüht, mehr schwerbehinderte Menschen zu beschäftigen. Bewerbungen Schwerbehinderter sind ausdrücklich erwünscht.
Bitte reichen Sie Ihre Bewerbung mit allen wichtigen Unterlagen bis zum **1. Juni 2018** über das Online-Formular unter *https://s-lotus.gwdg.de/gwdgdb/aga/20180507.nsf/bewerbung* ein.

Fragen zur ausgeschriebenen Stelle beantworten Ihnen:

Herr Simon Heider
Tel.: 0551 201-1840
E-Mail: *simon.heider@gwdg.de* oder

Herr Ralph Krimmel
Tel.: 0551 201-1821
E-Mail: *ralph.krimmel@gwdg.de* oder

Herr Dr. Burkhard Heise
Tel.: 0551 201-1526
E-Mail: *burkhard.heise@gwdg.de*

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

INFORMATIONEN:
support@gwdg.de
0551 201-1523

Mai bis
Dezember 2018

Kurse

KURS	VORTRAGENDE/R	TERMIN	ANMELDEN BIS	AE
QUICKSTARTING R: EINE ANWENDUNGSORIENTIERTE EINFÜHRUNG IN DAS STATISTIKPAKET R	Cordes	15.05. – 16.05.2018 9:00 – 12:00 und 13:00 – 15:30 Uhr	08.05.2018	8
ADMINISTRATION VON PCS IM ACTIVE DIRECTORY DER GWDG	Quentin	24.05.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	17.05.2018	4
SHAREPOINT – EINFÜHRUNG FÜR ANWENDER	Buck, Kasper	30.05.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	23.05.2018	4
SHAREPOINT –EINFÜHRUNG FÜR SITECOLLECTION-BESITZER	Buck, Kasper	31.05.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	24.05.2018	4
INDESIGN – AUFBAUKURS	Töpfer	05.06. – 06.06.2018 9:30 – 16:00 Uhr	29.05.2018	8
OUTLOOK – E-MAIL UND GROUPWARE	Helmvoigt	14.06.2018 9:15 – 12:00 und 13:00 – 16:00 Uhr	07.06.2018	4
ANGEWANDTE STATISTIK MIT SPSS FÜR NUTZER MIT VORKENNTNISSEN	Cordes	20.06. – 21.06.2018 9:00 – 12:00 und 13:00 – 15:30 Uhr	13.06.2018	8
SHAREPOINT – EINFÜHRUNG FÜR ANWENDER	Buck, Kasper	27.06.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	20.06.2018	4
SHAREPOINT – EINFÜHRUNG FÜR SITECOLLECTION-BESITZER	Buck, Kasper	28.06.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	21.06.2018	4

KURS	VORTRAGENDE/R	TERMIN	ANMELDEN BIS	AE
STATISTIK MIT R FÜR TEILNEHMER MIT VORKENNTNISSEN – VON DER ANALYSE ZUM BERICHT	Cordes	03.07. – 04.07.2018 9:00 – 12:00 und 13:00 – 15:30 Uhr	26.06.2018	8
SHAREPOINT – EINFÜHRUNG FÜR ANWENDER	Buck, Kasper	15.08.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	08.08.2018	4
SHAREPOINT – EINFÜHRUNG FÜR SITECOLLECTION-BESITZER	Buck, Kasper	16.08.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	09.08.2018	4
GRUNDLAGEN DER BILDBEARBEITUNG MIT PHOTOSHOP	Töpfer	21.08. – 22.08.2018 9:30 – 16:00 Uhr	14.08.2018	8
SHAREPOINT – EINFÜHRUNG FÜR ANWENDER	Buck, Kasper	12.09.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	05.09.2018	4
SHAREPOINT – EINFÜHRUNG FÜR SITECOLLECTION-BESITZER	Buck, Kasper	13.09.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	06.09.2018	4
INDESIGN – GRUNDLAGEN	Töpfer	18.09. – 19.09.2018 9:30 – 16:00 Uhr	11.09.2018	8
SHAREPOINT – EINFÜHRUNG FÜR ANWENDER	Buck, Kasper	17.10.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	10.10.2018	4
SHAREPOINT – EINFÜHRUNG FÜR SITECOLLECTION-BESITZER	Buck, Kasper	18.10.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	11.10.2018	4
PHOTOSHOP FÜR FORTGESCHRITTENE	Töpfer	23.10. – 24.10.2018 9:30 – 16:00 Uhr	16.10.2018	8
EINFÜHRUNG IN DIE STATISTISCHE DATENANALYSE MIT SPSS	Cordes	13.11. – 14.11.2018 9:00 – 12:00 und 13:00 – 15:30 Uhr	06.11.2018	8
ADMINISTRATION VON PCS IM ACTIVE DIRECTORY DER GWDG	Quentin	15.11.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	08.11.2018	4
SHAREPOINT – EINFÜHRUNG FÜR ANWENDER	Buck, Kasper	21.11.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	14.11.2018	4
SHAREPOINT – EINFÜHRUNG FÜR SITECOLLECTION-BESITZER	Buck, Kasper	22.11.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	15.11.2018	4
INDESIGN – AUFBAUKURS	Töpfer	27.11. – 28.11.2018 9:30 – 16:00 Uhr	20.11.2018	8
OUTLOOK – E-MAIL UND GROUPWARE	Helmvoigt	06.12.2018 9:15 – 12:00 und 13:00 – 16:00 Uhr	29.11.2018	4
ANGEWANDTE STATISTIK MIT SPSS FÜR NUTZER MIT VORKENNTNISSEN	Cordes	11.12. – 12.12.2018 9:00 – 12:00 und 13:00 – 15:30 Uhr	04.12.2018	8

KURS	VORTRAGENDE/R	TERMIN	ANMELDEN BIS	AE
SHAREPOINT – EINFÜHRUNG FÜR ANWENDER	Buck, Kasper	19.12.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	12.12.2018	4
SHAREPOINT – EINFÜHRUNG FÜR SITECOLLECTION-BESITZER	Buck, Kasper	20.12.2018 9:00 – 12:30 und 13:30 – 15:30 Uhr	13.12.2018	4

Teilnehmerkreis
Das Kursangebot der GWDG richtet sich an alle Mitarbeiterinnen und Mitarbeiter aus den Instituten der Universität Göttingen und der Max-Planck-Gesellschaft sowie aus einigen anderen wissenschaftlichen Einrichtungen.

Anmeldung
Anmeldungen können schriftlich per Brief oder per Fax unter der Nummer 0551 201-2150 an die GWDG, Postfach 2841, 37018 Göttingen oder per E-Mail an die Adresse *support@gwdg.de* erfolgen. Für die schriftliche Anmeldung steht unter *https://www.gwdg.de/antragsformulare* ein Formular zur Verfügung. Telefonische Anmeldungen können leider nicht angenommen werden.

Kosten bzw. Gebühren
Unsere Kurse werden wie die meisten anderen Leistungen der GWDG in Arbeitseinheiten (AE) vom jeweiligen Institutskontingent abgerechnet. Für die Institute der Universität Göttingen und der Max-Planck-Gesellschaft erfolgt keine Abrechnung in EUR.

Absage
Sie können bis zu acht Tagen vor Kursbeginn per E-Mail an *support@gwdg.de* oder telefonisch unter 0551 201-1523 absagen. Bei späteren Absagen werden allerdings die für die Kurse berechneten AE vom jeweiligen Institutskontingent abgebucht.

Kursorte
Alle Kurse finden im Kursraum oder Vortragsraum der GWDG statt. Die Wegbeschreibung zur GWDG sowie der Lageplan sind unter *https://www.gwdg.de/lageplan* zu finden.

Kurstermine
Die genauen Kurstermine und -zeiten sowie aktuelle kurzfristige Informationen zu den Kursen, insbesondere zu freien Plätzen, sind unter *https://www.gwdg.de/kursprogramm* zu finden.

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

FTP-Server

Eine ergiebige Fundgrube!

Ihre Anforderung

Sie möchten auf das weltweite OpenSource-Softwareangebot zentral und schnell zugreifen. Sie benötigen Handbücher oder Programmbeschreibungen oder Listings aus Computerzeitschriften. Sie wollen Updates Ihrer Linux- oder FreeBSD-Installation schnell durchführen.

Unser Angebot

Die GWDG betreibt seit 1992 einen der weltweit bekanntesten FTP-Server, seit sieben Jahren mit leistungsfähigen Ressourcen für schnellen Service.

Ihre Vorteile

> Großer Datenbestand (50 TByte), weltweit verfügbar
> Besonders gute Anbindung im GÖNET
> Aktuelle Software inkl. Updates der gebräuchlichsten Linux-Distributionen
> Unter pub befindet sich eine aktuell gehaltene locatedb für schnelles Durchsuchen des Bestandes.
> Alle gängigen Protokolle (http, https, ftp und rsync) werden unterstützt.

Interessiert?

Wenn Sie unseren FTP-Server nutzen möchten, werfen Sie bitte einen Blick auf die u. g. Webseite. Jeder Nutzer kann den FTP-Dienst nutzen. Die Nutzer im GÖNET erreichen in der Regel durch die lokale Anbindung besseren Durchsatz als externe Nutzer.

>> www.gwdg.de/ftp-server


GWDG
Gesellschaft für wissenschaftliche
Datenverarbeitung mbH Göttingen