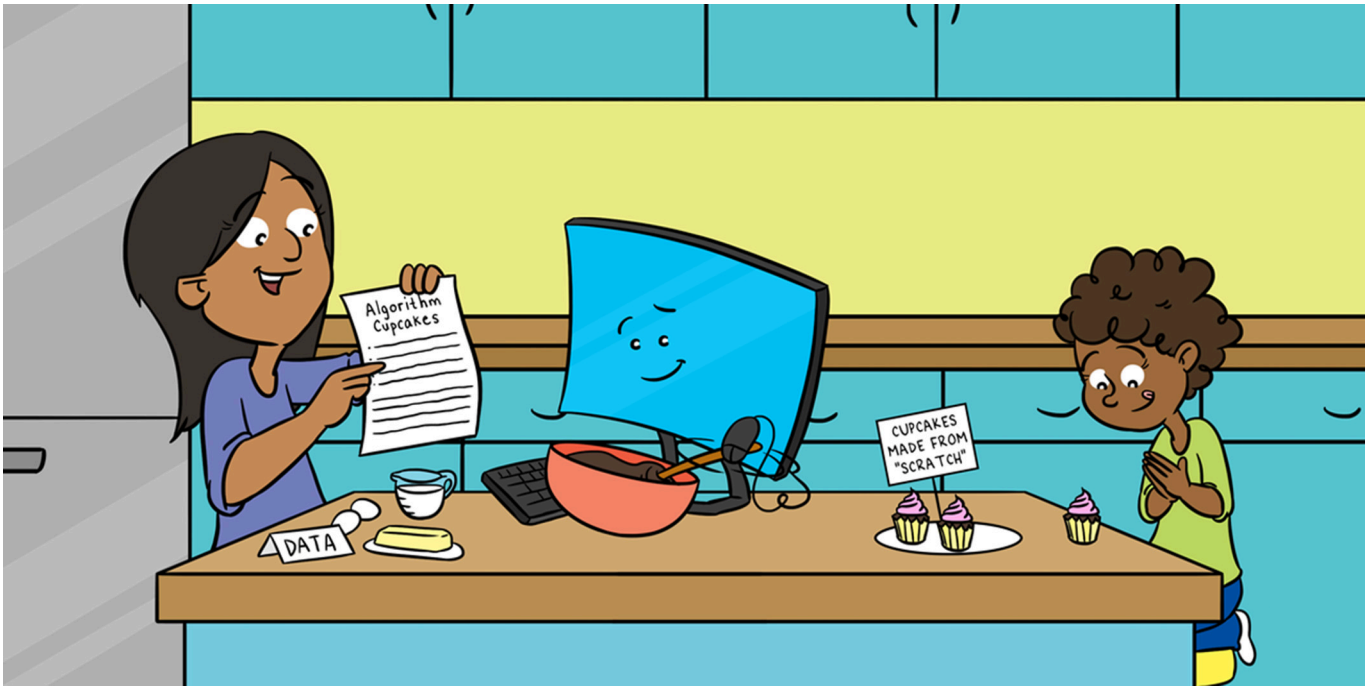
الرياضيات
نُشر على الإنترنت بتاريخ: 28 فبراير 2022
doi: 10.3389/frym.2021.632923-ar



تشكيل المعرفة من البيانات الضخمة باستخدام علم البيانات

Andrea Rau *
Université Paris-Saclay, INRAE, AgroParisTech, GABI, Jouy-en-Josas, France

تُعرف البيانات المجمعة بكميات ضخمة باسم "البيانات الضخمة"، وهي تُغيّر من الطريقة التي نفكر ونجيب بها عن الأسئلة في العديد من المجالات المختلفة، مثل التنبؤ بالطقس وعلم الأحياء. وفي ضوء وجود جميع هذه المعلومات، فإننا نحتاج إلى أجهزة الحاسوب لمساعدتنا على تخزينها ومعالجتها وتحليلها وفهمها. يضم علم البيانات أدوات من مجالات علمية مختلفة؛ مثل علم الإحصاء والرياضيات وعلوم الحاسوب لإيجاد أنماط مثيرة للاهتمام في البيانات الضخمة. إذ يكتب علماء البيانات تعليمات (توجيهات) تدريجية تعرف باسم الخوارزميات لإخبار الحاسوب بكيفية التعلم من البيانات. ولمساعدة الحاسوب على فهم هذه التوجيهات، يجب ترجمة الخوارزميات من السؤال الأصلي الذي وجهه عالم البيانات إلى إحدى لغات البرمجة؛ ثم تجب إعادة ترجمة النتائج حتى يتسنى للبشر فهمها، وهو ما يعني أن علماء البيانات هم محققو بيانات ومبرمجون ومترجمون في آن واحد!

المراجعون الصغار:


JASMINE
العمر:11

بيانات من حولنا في كل مكان

البيانات هي مجموعة منسقة من المعلومات المرتبطة —مثل الأرقام والقياسات والكلمات والأوصاف - والتي جُمعت وخُزِنَت لغرض معين. طورت العديد من الأدوات الجديدة مؤخرًا، وهو

ما سهل إلى حد ما عملية جمع كميات ضخمة للغاية من البيانات. فعندما تتاح البيانات بكميات هائلة، فإنها غالبًا ما تعرف باسم **البيانات الضخمة**. غيرت البيانات الضخمة من الطريقة التي نفكر بها ونجيب بها عن العديد من الأسئلة المختلفة، مثل التنبؤ بالطقس وإيجاد طرق مختصرة لتجنب التعثر في الازدحام المروري، أو اقتراح مسلسل تليفزيوني جديد قد تحبه بناء على المسلسلات التي شاهدتها من قبل.

البيانات الضخمة (BIG DATA)

مجموعات البيانات الضخمة جدًا والمعقدة للغاية والتي تشكل تحديًا للعلماء فيما يتعلق بتخزينها ومعالجتها وتحليلها وتفسيرها. ويحتاج علماء البيانات في الغالب إلى استخدام أدوات وطرق متخصصة للتعامل مع البيانات الضخمة.

البيانات الضخمة: تحدٍّ كبير في علم الأحياء!

ساعدت البيانات الضخمة أيضًا على تقدم الأبحاث في علم الأحياء، وهو علم معني بدراسة الكائنات الحية مثل الإنسان والحيوان والنبات والبكتيريا. وتتيح العديد من الأدوات المتخصصة جدًا الآن تجميع البيانات البيولوجية من مختبرات الأبحاث والمستشفيات، ومن الطبيعة، وحتى من المنزل! على سبيل المثال، يمكن أن تحتوي الأجهزة التي نرتديها على مستشعرات تنقل البيانات بشكل آني ومباشر لمساعدة الأطباء على مراقبة مدى جودة نومك. كما يمكن أيضًا للطائرات بدون طيار أن تحلق فوق المزارع والحقول وتلتقط صورًا للحقول لتعطي رؤية شاملة عن نمو المحاصيل الزراعية. ويمكن للتقنيات المختبرية الجديدة حاليًا أن تقرأ بسهولة المجموعة الكاملة من التعليمات الجينية لشخص ما، والتي تتكون من ثلاثة مليارات حرف (لإعطائك فكرة حول مقياس هذه الحروف، فإن ثلاثة مليارات ثانية تساوي 90 عامًا!). وفي ضوء وجود جميع هذه المعلومات، تمثل عمليات تخزينها ومعالجتها وتحليلها وفهمها تحديًا كبيرًا، كما أننا بحاجة إلى أجهزة الحاسوب للمساعدة.

علم الرياضيات + علم الإحصاء + علم الحاسوب + البيانات الضخمة = علم البيانات

البيانات الضخمة كبيرة جدًا لدرجة أنها قد أدت إلى تطوير مجال جديد نسبيًا ومثير للاهتمام يعرف باسم **علم البيانات**. يضم علم البيانات أدوات من علم الإحصاء والرياضيات وعلم الحاسوب لإيجاد أنماط مدهشة من قواعد البيانات المعقدة؛ مثل قواعد البيانات الضخمة. يجب أن يقضي علماء البيانات الكثير من الوقت في تنظيم البيانات قبل أن يبدأوا في العمل عليها. وللإجابة على سؤال معين، يحتاج عالم البيانات إلى إيجاد **مجموعة بيانات** أو تكوينها، أو إيجاد تشكيلة من مجموعات البيانات. وبعض مجموعات البيانات متاح للعامة للاستخدام، ومن الممكن أن تساعدك محركات البحث مثل "محرك بحث مجموعة بيانات جوجل[1]" في هذا الأمر باستخدام الكلمات المفتاحية. بينما هناك مجموعات أخرى من البيانات، مثل تلك التي تحتوي على معلومات طبية عن المرضى، لا تكون متاحة إلا لمجموعة محددة من الأشخاص فقط. وربما يحتاج عالم البيانات إلى جمع بيانات جديدة للإجابة على سؤال ما. على سبيل المثال، إذا أردت أن تعرف اللون المفضل لزميلك في الصف الدراسي، فيمكنك كتابة استبيان لجمع الإجابات من الطلاب الآخرين.

علم البيانات (DATA SCIENCE)

مجال علمي بيني يضم أدوات من علم الإحصاء والرياضيات وعلم الحاسوب لإيجاد أنماط مثيرة للاهتمام من قواعد البيانات المعقدة، مثل البيانات الضخمة.

[1] https://datasetsearch.research.google.com

مجموعة البيانات (DATASET)

مجموعة منسقة من المعلومات المرتبطة - مثل الأرقام والقياسات والكلمات والأوصاف - والتي جُمعت وخزنت لغرض معين.

من الفوضى إلى البيانات المُنظَمة

يعد تنظيم البيانات في صيغة قابلة للاستخدام من أكبر المهام التي يجب على عالم البيانات القيام بها. وإحدى طرق فعل ذلك هو تخيل البيانات الضخمة باعتبارها خليطًا يحتوي على جميع قطع "الليجو" التي لديك مبعثرة هنا وهناك في جميع أرجاء منزلك. فقبل أن تبدأ في تصنيف هذه القطع لبناء شيء ما، يجب أن تقوم أولًا بترتيبها وتجميعها كلها في كومة واحدة في نفس الغرفة! إن معظم مجموعات البيانات الحقيقية تقبع في حالة من "الفوضى" الشديدة، بمعنى أنها قد تشتمل على أخطاء مطبعية أو حتى قيم مفقودة. وكمثال على ذلك، ربما تشمل بعض الردود على الاستبيان

الذي قمت به حول اللون المفضل لزميلك إجابات مثل: ``أزرق''، و``الأزرق''، و``أزرق''. ولجعل هذه البيانات أسهل في الفهم، ستحتاج إلى ترتيبها من خلال تغيير كل هذه الاختلافات إلى قيمة واحدة مثل ``أزرق''، حيث إنها جميعها تشير إلى نفس اللون.

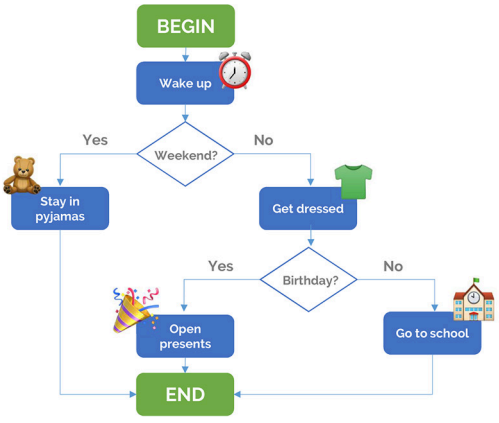
الخوارزميات: وصفات علم البيانات

بمجرد أن تكون جميع قطع الليجو الخاصة بك في مكان واحد، يكون أمامك الكثير من الأهداف، مثل تصنيف مكعبات الليجو في مجموعات أو التنبؤ بنوع مجموعة المكعبات التي ربما تحبها لاحقًا. وإذا كان لديك عدد صغير من قطع هذه اللعبة، فربما يكون من السهل عليك القيام بهذا الأمر يدويًا. أما في حال البيانات الضخمة، فإننا بحاجة إلى أدوات خاصة لمساعدتنا في إتمام المهمة. ويعتبر **تعلم الآلة** إحدى الأدوات القوية للتعامل مع البيانات الضخمة، وهو ما يحدث عندما نأمر الحاسب الآلي بالتعلم من البيانات دون أن نزوده بالإجابة أولًا. ولفعل هذا، يجب أن يعطي علماء البيانات الحاسب الآلي مجموعة من التوجيهات التفصيلية التدريجية تعرف باسم **الخوارزمية** (الشكل 1). ويجب أن تكون هذه الخوارزميات مكتوبة بطريقة يستطيع الحاسب الآلي أن يقرأها، وهو ما يعرف **بالتشفير**. يمكنك أن تفكر في الخوارزمية باعتبارها وصفة لخبز كعكة. تبدأ الوصفة بمجموعة من العناصر (بياناتك)، والتي تخبرك بالضبط بكيفية مزج الزبد وتسخين الفرن وطهو الكعكة (الخوارزمية الخاصة بك) للحصول على حلوى لذيذة (نتائجك). ومع ذلك، فالفرق بين الوصفة والخوارزمية يتمثل في أن توجيهات الخوارزمية يجب أن تكون دقيقة جدًا حتى يعلم الحاسب الآلي ما يجب عليه القيام به بالضبط.

في الوصفة الغذائية، نقول: ``اخلط العجين السائل مع القليل من الملح''، ولكن في الخوارزمية سيكون الأمر كالآتي: ``أضف جرامًا واحدًا من الملح إلى العجين السائل وقلّبهما ثلاث مرات باستخدام ملعقة خشبية''.

ما اللغة التي تجيدها أنت وحاسوبك على حد سواء؟

التشفير طريقة لترجمة السؤال العلمي إلى لغة يتحدثها الحاسب الآلي. هناك العديد من اللغات المختلفة التي يتحدثها الناس في كل أرجاء العالم (الإنجليزية والفرنسية والإيطالية والألمانية وغيرها




```
BEGIN
wake up
if today is a weekend then
 | stay in pyjamas
else
 | get dressed
 | if today is your birthday then
 |  | open presents
 | else
 |  | go to school
END
```

شكل 1

تعلم الآلة (MACHINE LEARNING)

استخدام الخوارزميات في تعليم الحاسب الآلي كيفية التعلم بشكل أوتوماتيكي من البيانات وتحسين مستواه عن طريق الخبرة والتجربة دون الحاجة إلى تدخل بشري.

الخوارزمية (ALGORITHM)

مجموعة من التوجيهات أو القواعد التفصيلية التدريجية التي يجب على الحاسب الآلي اتباعها.

التشفير أو الترميز (CODING)

استخدام لغة برمجة للتواصل مع جهاز الحاسب الآلي، وتزويده بالتوجيهات والتعليمات المشار إليها باسم الخوارزمية.

شكل 1

الخوارزمية هي مجموعة من الأوامر التدريجية للحاسب الآلي. وبعد رسم مخطط انسيابي لتوصيل كل نقطة بالأخرى إحدى الطرق الفعالة لتصور خوارزمية معينة وبنائها. في المخططات الانسيابية، يمكن أن تمثل المستطيلات أفعالًا أو خطوات، بينما يمثل المعين القرارات. يمكنك في الصباح استخدام مخطط انسيابي، مثل هذا على يسار الشكل، لتقرير ما إذا كان بمقدورك البقاء مرتديًا بيجامة النوم، أو فتح هدايا عيد الميلاد، أو الذهاب إلى المدرسة أم لا. وبعد رسم المخطط الانسيابي، يمكنك حينها ترجمة خطوات الخوارزمية الخاصة بك إلى وصف أكثر تفصيلًا، كما هو موضح على اليمين.

الكثير). وبالمثل، يوجد العديد من لغات التشفير التي يمكن استخدامها لكتابة خوارزمية معينة (الشكل 2). وعلى غرار الوصفة المكتوبة بالإنجليزية والفرنسية والتي قد تعبر عن الشيء ذاته بطريقتين مختلفتين، تزود لغات التشفير المختلفة الحاسب الآلي بالتوجيهات بطرق مختلفة. يبتكر العلماء لغات تشفير جديدة كل عام! وهناك أيضًا لغة تشفير تم اختراعها خصيصًا للأطفال في المرحلة العمرية من 8 أعوام إلى 16 عامًا وتعرف بـ Scratch [2] [1]. وتوجد لغتا تشفير شائعتان يستخدمهما علماء البيانات حاليًا بصورة معتادة لكتابة الخوارزميات، وتعرفان باسم R و Python. وكلتا اللغتان **مفتوحة المصدر**، وهو ما يعني أن علماء البيانات الذين يكتبون هذه الخوارزميات بهاتين اللغتين يشاركونهما مع الغير مجانًا. وهو ما يسهل على علماء البيانات العمل سويًا والمساعدة على تحسين الرموز التي يبتكرونها!

https://scratch.mit.edu[2]

مفتوح المصدر (OPEN SOURCE)

نوع من برمجيات الحاسب الآلي المطورة والمدعومة مجتمعيًا. وتكون الرموز والبرمجيات مفتوحة المصدر متاحة للاستخدام والمشاركة والتعديل المجاني من قبل أي شخص.

جمع وصفات الحاسوب في كتاب وصفات علم البيانات

ربما يتوجب على عالم البيانات كتابة العديد من الخوارزميات وجمعها للحصول على الإجابة التي يبحث عنها. وكما يجمع الطاهي العديد من الوصفات في كتاب طهي واحد، فإن عالم البيانات أحيانًا ما يبتكر أو يستخدم مجموعة من الخوارزميات تعرف بحزم البرمجيات. وعندما تُكتب **حزم البرمجيات** بلغة مفتوحة المصدر مثل R أو Python، فإن هذا يساعد عالم البيانات على إيجاد عمل قابل للتكرار. نعني بعلم البيانات القابل للتكرار أنه يمكن للأشخاص الآخرين إعادة تشغيل عمل عالم آخر وتكراره وإعادة استخدامه. وهو ما يساعد الجميع على العمل على نحوٍ أكثر كفاءة ومشاركة النتائج التي يتوصلون إليها مع الآخرين بسهولة أكبر.

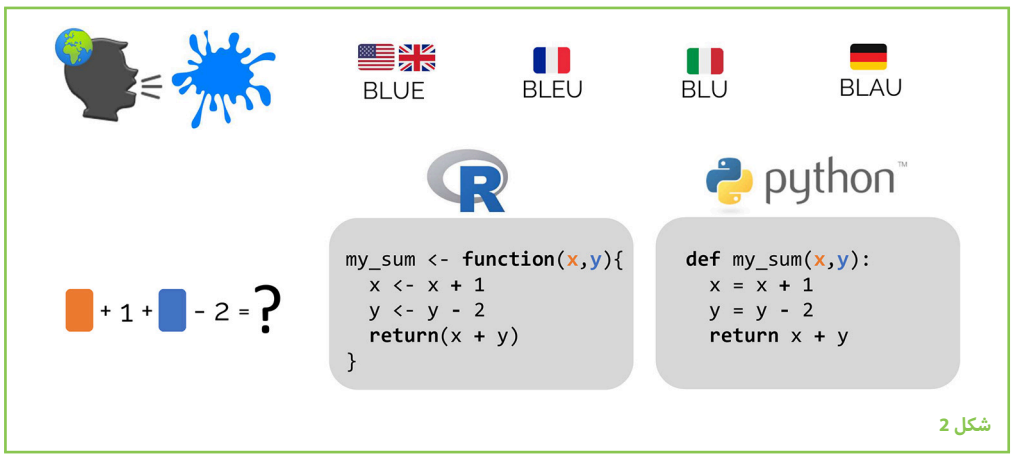
حزم البرمجيات (SOFTWARE PACKAGE)

مجموعة منظمة من الخوارزميات المرتبطة والتي تعمل سويًا لأداء مهمة معينة أو القيام بوظيفة متشابهة.

تساعد عملية إعادة التكرار أيضًا على خلق شعور من الثقة حول صحة هذه الخوارزميات وموثوقيتها. وبنفس الطريقة، يمكنك إعطاء كتاب الطهو المفضل لك إلى أحد أصدقائك حتى يمكنه صناعة هذه الكعكة اللذيذة لنفسه!

الخلاصة

البيانات الضخمة آخذة في الازدياد، سواء في علم الأحياء أو المعاملات المصرفية أو التسويق، كما سيستمر تأثيرها الضخم على حياتنا في شتى المجالات. ولكن هناك قلق متزايد يتعلق بعواقب تجميع البيانات الضخمة على الخصوصية، مثلما يحدث عندما تسجل في خدمة مجانية أو تطبيق



شكل 2

يمكن تشفير الخوارزميات باستخدام لغات تشفير مختلفة، تمامًا كما يمكن أن نعبر عن الأفكار باستخدام اللغات المختلفة. دعونا نقول إننا نريد أن نكتب خوارزمية ستأخذ أي رقمين، أضف 1 إلى الرقم الأول واطرح 2 من الرقم الثاني، ثم أجمعهما سويًا. فلو بدأنا بـ2 و4، سنكون بحاجة إلى أن نعلّم الحاسب الآلي أن يعطينا (2 + 1) + (4 − 2) = 5 كإجابة. تبدو الخوارزمية هنا، والتي تسمى my_sum، متشابهة في لغتيّ التشفير R وPython؛ ولكن إذا دققت النظر، فستجد بعض الاختلافات.

مجاني (مثل مواقع التواصل الاجتماعي، أو البريد الإلكتروني، أو البث المباشر للفيديوهات، أو خدمات مشاركة الموقع)، أو في تبادل الموافقات لجعل شركة ذات ملكية خاصة تجمع بيانات عنك. وربما تشمل البيانات الكلمات المفتاحية التي تبحث عنها، والمواقع الإلكترونية التي تتصفحها، والفيديوهات التي تحبها، أو الأماكن التي تزورها في الحي الذي تسكن فيه. تستخدم الشركات هذه البيانات لتصميم إعلانات ودعاية تستهدفك أنت خصيصًا، ويكون الهدف من ذلك عادة بيع أكبر قدر ممكن من المنتجات لك! يمكنك أخذ الخطوات لتدرك أنواع البيانات التي يتم تجميعها عنك من خلال الاطلاع على خصائص التطبيق، على سبيل المثال. وهو ما يساعدك على فرض قيود على عملية جمع بعض أنواع البيانات، مثل معلومات عن الموقع، كما يساعدك ذلك أيضًا على تحديد التطبيقات والخدمات التي تثق فيها، وتلك التي ربما تفكر في إزالة تثبيتها من على جهازك.

وخلال السنوات القادمة، سنكون بحاجة إلى الكثير من علماء البيانات الجدد لمساعدتنا في فهم البيانات الضخمة باستخدام تقنيات تعلم الآلة. وسيكون من الضروري جدًا للناس من مختلف الخلفيات أن يتأكدوا من حصول جميع الأطراف على استفادة متساوية من هذه التحليلات. إنه وقت مناسب حقًا كي تصبح عالم بيانات؛ فنحن مثل المحققين وعلماء الرياضيات والفنانين ومبرمجي الحاسوب والمترجمين، ونؤدي جميع هذه المهن مدمجة في آن واحد!

المراجع

1. Maloney, J., Resnick, M., Rusk, N., Silverman, B., and Eastmond, E. 2010. The scratch programming language and environment. *ACM Trans. Comput. Educ.* 10:1–15. doi: 10.1145/1868358.1868363

نُشر على الإنترنت بتاريخ: 28 فبراير 2022

حرره: Norma Ortiz-Robinson

مرشدو العلوم: Jason Anema

الاقتباس: Rau A (2022) تشكيل المعرفة من البيانات الضخمة باستخدام علم البيانات. Front. Young Minds doi: 10.3389/frym.2021.632923-ar

مُترجَم ومقتبس من: Rau A (2021) Cooking Up Knowledge From Big Data Using Data Science. Front. Young Minds 9:632923. doi: 10.3389/frym.2021.632923

إقرار تضارب المصالح: يعلن المؤلفون أن البحث قد أُجري في غياب أي علاقات تجارية أو مالية يمكن تفسيرها على أنها تضارب محتمل في المصالح.

المراجعون الصغار

JASMINE, العمر: 11

أُدعى ياسمين. أحب ألعاب الألواح التعاونية ذات الطابع الاستراتيجي. كما أحب أيضًا قراءة روايات أجاثا كريستي وغيرها من الروايات البوليسية الكلاسيكية. أحب التزلج على الجليد والسباحة والتجديف بالقوارب والفنون القتالية. وأنا حاصلة على حزام أسود في لعبة الكاراتيه، وهو ما يعد أفضل إنجازاتي لأنه كان تحديًّا كبيرًا حيث تطلب الأمر مني ست سنوات حتى أنهي مقرري التعليمي في هذه الرياضة.

المؤلفون

ANDREA RAU

أنا عالمة إحصاء حيوية وعالمة بيانات وباحثة في المعهد الوطني الفرنسي للبحوث الزراعية والغذاء والبيئة في جوي أون جوساس، بفرنسا. أطور النماذج الإحصائية وأكتب رموز الحاسوب لمساعدة علماء الأحياء على العثور على أنماطٍ مثيرة للاهتمام في بياناتهم حول علم الجينوم. وبالإضافة إلى كتابة رموز الحاسوب بلغة برمجة تسمى R، فإنني أتحدث الإنجليزية والفرنسية في العمل. وفي وقت فراغي، أحب طهي الوصفات الجديدة واللعب مع ابنتي إليز وكلبي بيلا. andrea.rau@inrae.fr*

النسخة العربية مقدمة من
Arabic version provided by

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology