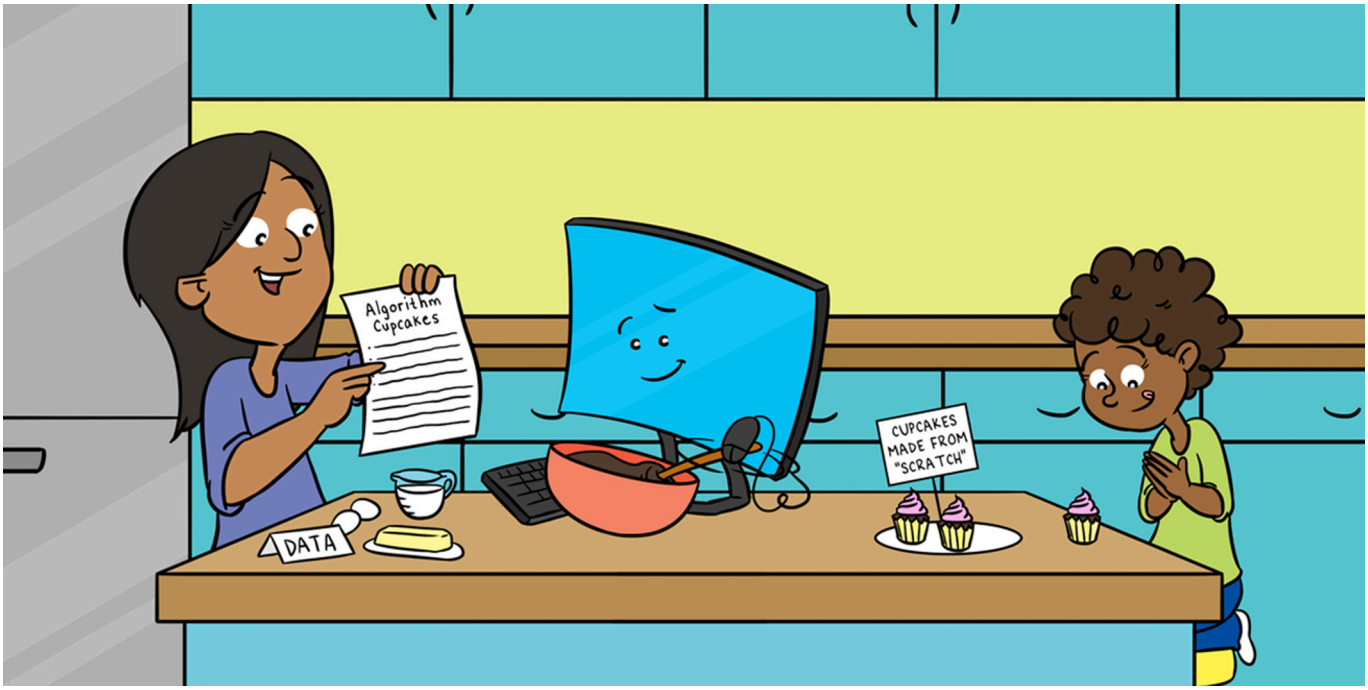
frontiers | פרונטירז מדע לצעירים

מתמטיקה וכלכלה
פורסם אונליין: 05 בינואר 2024
doi: 10.3389/frym.2021.632923-he



ניתוח יֶדַע מִבִּיג דָאטָה באמצעות מדעי הנתונים

Andrea Rau *

אוניברסיטת פריז-סקליי, AgroParisTech, INRAE, GABI, Jouy-en-Josas, צרפת

נתונים שנאספים בכמויות גדולות מאוד נקראים בִּיג דָאטָה (Big Data, בעברית: נְתוּנֵי עָתֵק). אלה שינו את האופן שבו אנו חושבים על שאלות במגוון תחומים כמו חיזוי מזג האוויר וביולוגיה, ועונים עליהן. זמינותו של המידע הרב הזה מצריכה מַחְשְׁבִים שיסייעו לנו לאחסן את המידע, לעבד אותו, לנתחו ולהבינו. מדעי הנתונים משלבים כלים מתחומים כמו סטטיסטיקה, מתמטיקה ומדעי המחשב במטרה למצוא דפוסים מעניינים בביג דאטה. מדעני נתונים כותבים הוראות בסגנון 'צעד אחר צעד', שמכונות אלגוריתמים, במטרה ללמֵד מחשבים כיצד ללמוד מנתונים. כדי לסייע למחשבים להבין את ההוראות האלה, אלגוריתמים צריכים לתרגם את השאלה המקורית שנשאלה על ידי מדעני הנתונים לשפת תכנות. התוצאות צריכות להיות מתורגמות חזרה, כך שבני אדם יוכלו להבינן. כלומר, מדעני נתונים הם בלשי נתונים, מתכנתים ומתרגמים – הכול בחבילה אחת!

נתונים, נתונים בכל מקום

נתונים הם אוסף של מידע – מִספרים, מדידות, מילים, או תיאורים – שנאספו ואוחסנו למטרה מסוימת. לאחרונה, פותחו מגוון כלים חדשים שהפכו איסוף של כמויות גדולות

סוקרים צעירים

JASMINE
גיל: 11

במיוחד של נתונים למשימה קלה למדי. כאשר נתונים זמינים בכמויות עצומות, לעיתים קרובות הם מכונים בִּיג דָאטָה. ביג דאטה שינו את האופן שבו אנו חושבים על מגוון שאלות שונות ועונים עליהן, לדוגמה בתחומי חיזוי מזג האוויר; מציאת נתיבים חלופיים כדי להימנע מעמידה בפקק תנועה, או הצעת סדרת טלוויזיה חדשה שהצופים עשויים לאהוב בהתבסס על תוכניות קודמות שבהן צפו.

ביג דאטה (Big data)

מערכים גדולים ומורכבים במיוחד של נתונים, שהם מאתגרים לאחסון, לעיבוד, לניתוח ולפירוש. לעיתים קרובות מדעני נתונים צריכים להשתמש בכלים ושיטות מיוחדים כדי לעבוד עם ביג דאטה.

ביג דאטה: אתגר גדול בביולוגיה!

ביג דאטה אף סייעו לקדם מחקר בביולוגיה – ענף העוסק בחקר דברים חיים כמו בני אדם, חיות, צמחים וחיידקים. כיום, מגוון כלים ייעודיים מאפשרים איסוף של ביג דאטה בביולוגיה במעבדות מחקר; בבתי חולים; בטבע, ואפילו בבית! לדוגמה, מכשירים לְבִישִׁים כמו שעונים חכמים יכולים להכיל חיישנים הפועלים בזמן אמת, ולסייע לרופאים לְנַטֵר למשל כמה טוב אתם יְשֵׁנִים.

רַחְפָנִים יכולים לעוף מעל חוֹות ולצלם תמונות של השדות במטרה לסַפֵּק מבט ממעוף הציפור לגבי מצבם של יבולים. שיטות מעבדה חדשות מאפשרות כעת לקרוא בקלות את מערַך ההוראות הגנטיות השלם של אדם, המורכב מכ-3 מיליארד אותיות ברצפים שונים (כדי לתת לכם מושג לסדר הגודל – 3 מיליארד שניות שוות ל-90 שנים!). זמינותו של המידע הרב הזה מציבה אתגר הכרוך באחסון נתונים; עיבודם, ניתוחם ופירושם. כאן נכנסים לתמונה מַחשבים, המסייעים לנו לעשות זאת.

מדע נתונים (Data science)

ענף בין-תחומי המשלב כלים מסטטיסטיקה, מתמטיקה ומדעי המחשב במטרה למצוא דפוסים מעניינים במערכי נתונים מורכבים, לרבות ביג דאטה.

מערך נתונים (Dataset)

אוסף מובנה של מידע קשור – מִספרים, מדידות, מילים, או תיאורים – שנאסף ואוחסן עבור סיבה מסוימת.

מתמטיקה + סטטיסטיקה + מדעי המחשב + ביג דאטה = מדע הנתונים

ביג דאטה הם כה גדולים בהיקפם ובכמותם, עד שהובילו לפיתוח תחום חדש יחסית ומְרגֵש, שנקרא מדע הנתונים. מדע הנתונים משלֵב כלים ממגוון תחומים אחרים, לרבות סטטיסטיקה, מתמטיקה ומדעי המחשב, במטרה למצוא דפוסים מעניינים בנתונים מורכבים. מדעני נתונים נדרשים להקדיש זמן רב לארגון נתונים לפני שביכולתם להתחיל לעבוד. כדי לענות על שאלה מסוימת, מדעני נתונים צריכים למצוא מערַך נתונים (סֶט נתונים), או אוסף של מערכי נתונים, או ליצור כאלה. חלק ממערכי הנתונים זמינים לציבור ומורשים לשימוש כולם. מנוע חיפוש כמו Google Dataset Search[1] יכול לסייע למצוא מערך כזה באמצעות מילות מפתח. מערכי נתונים אחרים, כמו אלה שכוללים מידע רפואי אישי, עשויים להיות זמינים רק למערך מצומצם של אנשים. כדי לענות על שאלה מסוימת, מדעני נתונים עשויים להצטרך לאסוף נתונים חדשים. לדוגמה, אם ברצונכם לדעת מהם הצבעים האהובים על חבריכם לכיתה, תוכלו לכתוב שאלון במטרה לאסוף תשובות מהתלמידים האחרים.

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

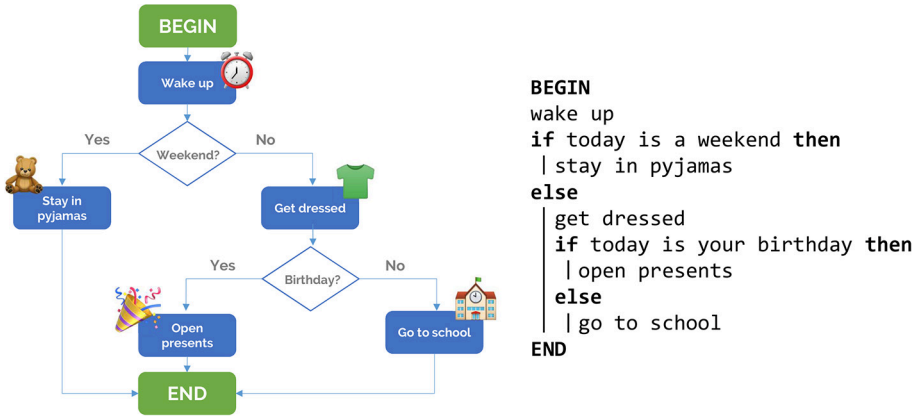
מנתונים מבולגנים לנתונים מסודרים

חלק גדול מעבודתם של מדעני נתונים כרוך בארגון הנתונים שהם מעוניינים להשתמש בהם בפורמט שָׁמִיש. דרך אחת לחשוב על כך היא לדמיין ביג דאטה בתור אוסף גדול של לְבֵנֵי לֶגוֹ® המפוזרות בכל רחבי הבית. לפני שאתם מתחילים למיין את הַלְּבֵנִים שלכם במטרה לִבְנות

משהו, עליכם לסדֵּר אותן בערימה בחֶדֶר אחד! מַרבית מערכי הנתונים האמיתיים 'מבולגנים' מאוד, כלומר עשויים לכלול שגיאות דפוס או אפילו ערכים חסרים. לדוגמה, חלק מהתגובות לַסֶּקֶר שערכתם על הצבעים האהובים על חבריכם לכיתה יכול לכלול את התשובות: 'כחול'; 'קחול'; 'ככול' ו'כחוול'. כדי להפוך את הנתונים האלה קלים יותר להבנה, עליכם לסדר אותם על ידי שינוי כל הגרסאות הללו לערך יחיד- 'כחול', מאחר שכולן מתייחסות לאותו הצבע.

אלגוריתמים כמתכונים של מדע הנתונים

כאשר כל לבני הלגו® שלכם נמצאות במקום אחד, עשויות להיות לכם מטרות רבות. לדוגמה, לקבֵּץ את הלבנים למערכים, או לחזות מהו סוג המערך שבו תהיו מעוניינים להשתמש בהמשך. אם יש לכם כמות קטנה של לבני לגו®, יכול להיות קל לבצע זאת ידנית, אך עבור ביג דאטה, אנו זקוקים לסיוע של כלים מיוחדים. כלי עוצמתי להתמודדות עם ביג דאטה נקרא למידת מכונה, שבה מלמדים את המחשב כיצד ללמוד מנתונים בלי לקבל ראשית את התשובה. כדי לעשות זאת, מדעני נתונים נדרשים לספֵּק למחשב מערך של הוראות מפורטות, צעד אחר צעד, שנקראות אלגוריתם (איור 1). הוראות אלה צריכות להיכתב באופן שהמחשב יכול להבין, מה שמכונה קידוד (תכנות). תוכלו לחשוב על אלגוריתם כמו מתכון לאפיית עוגה. המתכון מתחיל עם מערך של רכיבים (הנתונים שלכם), והוא מנחה אתכם בדיוק כיצד (האלגוריתם שלכם) לערבב את הבלילה, לחמם את התנור ולאפות את העוגה כדי לקבל קינוח טעים (התוצאות שלכם). אולם, ההבדל בין מתכון לבין אלגוריתם הוא שההוראות של אלגוריתם צריכות להיות מדויקות מאוד כך שהמחשב יֵדַע בדיוק מה לעשות. כך, בדוגמת המתכון, במקום לכתוב "ערבבו קומץ מלח לבלילה", ייכתב באלגוריתם משהו בסגנון "הוסיפו גרם אחד של מלח לבלילה וערבבו שלוש פעמים בעזרת כף עץ".



איור 1

מציאת שפה משותפת לבני אדם ולמַחשבים

קידוד הוא דרך לתרגם שאלה מדעית לשפה שהמחשב יכול לדבר בה. ברחבי העולם ישנן הרבה שפות שונות שאנשים מדבֵּרים בהן (אנגלית, צרפתית, איטלקית, עברית...), ובאופן דומה, ישנן הרבה שפות קוד שניתן להשתמש בהן כדי לכתוב אלגוריתם (איור 2). ממש כשם שמתכון שכתוב באנגלית ובצרפתית יכול לבטא את אותו הדבר בשתי דרכים ייחודיות, שפות

למידת מכונה (Machine learning)

השימוש באלגוריתמים במטרה ללמד מחשב כיצד ללמוד באופן אוטומטי מנתונים ולהשתפר מתוך הניסיון, בלי לקבל סיוע מבן אדם.

אלגוריתם (Algorithm)

מערך של הוראות או חוקים מפורטים, צעד אחר צעד, שמַחשב עוקב אחריהם.

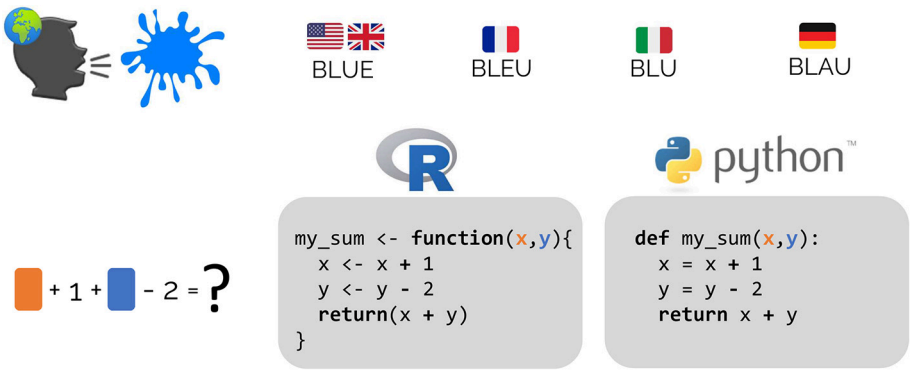
קידוד (Coding)

שימוש בשפת תכנות כדי לתקשר עם מחשב ולספק לו הוראות, שנקראות אלגוריתם.

איור 1

אלגוריתם הוא מערך של הוראות 'צעד אחר צעד' למחשב. דרך שימושית לדמיין אלגוריתם ולבנותו היא לצייר תרשים זרימה ולחבֵּר בין צעד אחד לבא אחריו. בתרשימי זרימה, מלבנים יכולים לייצג פעולות, ומעוינים החלטות. בבוקר, תוכלו להשתמש בתרשים זרימה כמו זה שבחלק השמאלי באיור כדי להחליט אם להישאר בפיג'מה, לפתוח מתנות יומולדת, או ללכת לבית הספר. לאחר ציור תרשים הזרימה, תוכלו לתרגם את הצעדים עבור האלגוריתם שלכם לתיאור מפורט יותר, כפי שמוצג בחלק הימני באיור.
מקרא החלק הימני (מקביל לתרשים הזרימה):
התחלה – התעוררו – אם היום חל סוף שבוע אז: הישארו בפיג'מה, אחרת: התלבשו – אם היום חל יום הולדתכם אז: פִּתחו מתנות, אחרת: לכו לבית הספר – סוף. Yes = כן, No = לא.

תכנות שונות מנחות את המחשב בהוראות בדרכים שונות. שפות קידוד חדשות מומצאות בכל שנה! ישנה אפילו שפת תכנות שנוצרה במיוחד עבור ילדים בני 8-16, שנקראת סקראץ'[2] [1]. כיום, שתי שפות תכנות פופולריות המשמשות מדעני נתונים לעיתים קרובות לכתיבת אלגוריתמים, הן 'R' ו'פייתון'. שתי השפות הן מסוג קוד פתוח, כלומר מדעני נתונים שכותבים את האלגוריתמים שלהם בשפות אלה יכולים לחלוק אותם עם כולם, חינם. זה מֵקֵל על מדעני נתונים לעבוד יחד ולסייע לשפר את הקודים שהם מפתחים, באופן הדדי!



איור 2

שילוב מתכוני מחשב לסֵפֶר הבישול של מדעי הנתונים

מדעני נתונים עשויים להצטרך לכתוב כמה אלגוריתמים ולשלֵב ביניהם כדי לקבל את התשובה שהם מחפשים. ממש כשם ששֶפִים עשויים לאסוף כמה מתכונים יחד בספר בישול, מדעני נתונים לעיתים יוצרים מְקַבְּצֵי אלגוריתמים, שמכונים חבילות תוכנה, או משתמשים בהם. כשחבילות תוכנה נכתבות בשפות קוד פתוח כמו R או פייתון, הדבר יכול לסייע למדעני נתונים ליצור עבודה הָדִירָה. מדע נתונים הָדִיר משמעותו שאנשים אחרים יכולים בקלות להריץ מחדש עבודה של מדעני נתונים, לחזור עליה, ולהשתמש בה מחדש. זה מסייע לכולם לעבוד ביעילות רבה יותר, ולחלוק בקלות עם אחרים את מה שמצאו. הדירות גם מסייעת לבנות אמון בנכונות האלגוריתמים. באותו האופן, תוכלו לתת את ספר הבישול האהוב עליכם לחברים, כך שיוכלו לאפות את העוגה הטעימה בעצמם!

מסקנות

ביג דאטה נעשים גדולים יותר, בין אם בביולוגיה, בבנקאות, או בשיווק, וצפויים להמשיך להשפיע מאוד על חיינו. אולם, לצד היתרונות ישנן גם דאגות גדֵלות לגבי ההשלכות של איסוף ביג דאטה על פרטיות. זאת כאשר אתם נרשמים לשירותים או ליישומים חינמיים (כמו רשת חברתית, דואר אלקטרוני, וידיאו סטרימינג, או שירותים מבוססי מיקום), בתמורה להסכמתכם לְאַפשר לחֶברה פרטית לאסוף עליכם נתונים. נתונים אלה עשויים לכלול את מילות המפתח שאתם מחפשים; את אתרי האינטרנט שאתם גולשים בהם; סרטונים שאתם אוהבים, או מקומות בשכונתכם שאתם מבקרים בהם. חֲבָרוֹת משתמשות בנתונים כדי ליצור פרסומות שמיועדות במיוחד עבורכם, לעיתים קרובות במטרה למכור לכם כמה שיותר! ביכולתכם לנקוט בצעדים כדי להיות מודעים לסוגי הנתונים שנאספים עליכם, לדוגמה על

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

קוד פתוח (Open source)

סוג של תוכנת מחשב המפותחת על ידי קהילה ונתמכת על ידה. באופן טיפוסי, קוד פתוח ותוכנת מקור פתוח פתוחים לשימוש, לחלוקה, ולשינוי על ידי כולם.

איור 2

ניתן לקודד אלגוריתמים בשפות תכנות שונות, ממש כשם שניתן לבטא רעיונות בשפות שונות. החלק העליון של האיור: המילה 'כחול' (Blue) באנגלית, כפי שמבוטאת על ידי לאומים שונים, משמאל לימין: ארה"ב ואנגליה, צרפת, איטליה וגרמניה. החלק התחתון של האיור: נאמר שאנו רוצים לכתוב אלגוריתם שלוקח כל שני מספרים, מוסיף 1 למספר הראשון ומחסיר 2 מהמספר השני, ואז מחבר ביניהם. אם נתחיל עם המספרים 2 ו-4, אנו מעוניינים ללמד את המחשב לתת לנו (2 + 1) + (2 - 4) = 5 בתור תשובה. האלגוריתם שלנו, אשר נקרא my_sum (הסכום שלי), נראה דומה בשפות התכנות R ופייתון (python), אולם אם תסתכלו על שני האלגוריתמים מקרוב, תבחינו בהבדלים מסוימים.

חבילת תוכנה (Software package)

אוסף מאורגן של אלגוריתמים קשורים הפועלים יחד עבור מטלה מסוימת, או שיש להם תפקוד דומה.

ידי הסתכלות בהגדרות של יישׂומים. זה יכול לסייע לכם להגביל את איסוף הנתונים לסוגי נתונים מסוימים, כמו מידע על מיקומכם, וכן לעזור לכם להחליט על אֵילוּ יישומים ושירותים אתם סומכים, ואֵילוּ אתם צריכים לשקול להסיר.

בשנים הקרובות, נזדקק להרבה מדעני נתונים חדשים שיוכלו למצוא היגיון בביג דאטה באמצעות שיטות של למידת מכונה. זה יהיה חשוב במיוחד עבור אנשים ממגוון רקעים, כדי לסייע לוודא שכולם יכולים להרוויח באופן שווה מהניתוחים הללו. זֶהוּ זמן מרגֵּש להיות מדעני נתונים הפועלים כמו בלשים, מתמטיקאים, אומנים, מתכנתים ומתרגמים – הכול באריזה אחת!

מקורות

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

פורסם אונליין: 05 בינואר 2024

נערך על ידי: Norma Ortiz-Robinson

מנחים מדעיים : Jason Anema

ציטוט: Rau A (2024) ניתוח יֶדַע מבִּיג דָּאטָה באמצעות מדעי הנתונים. Front. Young Minds. doi: 10.3389/frym.2021.632923-he

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

הצהרת ניגוד אינטרסים: המחברים מצהירים כל המחקר נערך בהעדר כי קשר מסחרי או פיננסי שיכול להתפרש כניגוד אינטרסים פוטנציאלי.

זכויות יוצרים © 2021 © 2024 Rau. זהו מאמר בגישה פתוחה שמופץ תחת תנאי רישיון Creative Commons Attribution License (CC BY). השימוש, ההפצה או ההעתקה מותרים לשימוש בפורומים אחרים ובלבד שיינתן קרדיט למחברים המקוריים ולבעל זכויות היוצרים, ושהפרסום המקורי בעיתון זה מצוטט בהתאם למקובל באקדמיה. השימוש, ההפצה או ההעתקה אינם מותרים אם הם אינם עומדים בתנאים אלה.

סוקרים צעירים

JASMINE, גיל: 11
קוראים לי Jasmine. אני אוהבת משחקי לוח שיתופיים מבוססי אסטרטגיה, וגם לקרוא את ספרי אגאתה כריסטי ונובלות בלשיות קלאסיות אחרות. נהנית לעשות סקי, לשחות, לחתור בקיאק ולהתאמן באומנויות

לחימה. יש לי חגורה שחורה בקרטֶה, וזה ההישג המועדף עליי מאחר שהיה מאתגֵר מאוד, ולקח לי 6 שנים להשלים את הכשרתי.

הכותבים

ANDREA RAU

אני בִּיוֹסטטיסטיקאית, מדענית נתונים, וחוקרת במכון המחקר הצרפתי הלאומי לחקלאות, מזון וסביבה (INRAE) ב-Jouy en Josas, צרפת. אני מפתחת מודֵלים סטטיסטיים וכותבת קוד מחשב במטרה לסייע לביולוגים למצוא דפוסים מעניינים בנתוני הגֶנומיקה שלהם. נוסף על כתיבת קוד מחשב בשפת התכנות R, במסגרת עבודתי אני מדברת אנגלית וצרפתית. בזמני הפנוי, אוהבת לבשל מתכונים חדשים ולשחק עם בתי Elise ועם הכלבה שלי Bella. andrea.rau@inrae.fr*

הוצאת פרונטירז מדע לצעירים ישראל
Hebrew version provided by

מוזיאון המדע ע"ש בלומפילד ירושלים
متحف العلوم على اسم بلومفيلد القدس
Bloomfield Science Museum Jerusalem

THE SAGOL NETWORK