COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Peter Rose (UC San Diego), May 5,2020

Title: COVID-19-Net : Intégration des données sur la santé, les agents pathogènes et l'environnement dans un graphique de connaissances pour le suivi, l'analyse et la prévision des cas

Peter Rose CIC Database Profile

NSF Award #: 2028411

Youtube Recording with Slides

July 2020 CIC Webinar Information

Transcript Editor: Rhyley Vaughan

Transcript

Florence Hudson:
Merci, Peter.

Peter Rose:
Slide 1
D'accord, je vous remercie, Florence, de nous avoir invités. Merci, Florence, de nous avoir invités. Permettez-moi de vous parler un peu de l'historique de COVID-19-Net. Nous faisions partie du programme de réseau de connaissances ouvert de la NSF. En particulier, nous étions intéressés par la mise en relation de trois types de données : les données biomédicales, les données environnementales et les données sociodémographiques. Puis, en janvier, COVID est arrivé. Nous avons pensé qu'il s'agissait d'un excellent cas d'utilisation où nous devions intégrer des données multidisciplinaires, alors si nous pouvons passer à la diapositive suivante.

Slide 2
Cette diapositive vous montre ce que nous essayons d'accomplir. Si vous pensez à COVID, il y a vraiment trois domaines principaux. Le premier est l'hôte, qui peut être humain ou animal, etc. Ensuite, il y a le pathogène, le virus, et tout le reste, que nous appelons l'environnement. Nous voulons permettre aux chercheurs d'étudier l'interaction entre ces différents domaines. Par exemple, l'interaction hôte-pathogène : comment l'environnement affecte-t-il les infections,

etc. Voilà pour le contexte, et nous avons commencé. Prenez la diapositive suivante, s'il vous plaît.

Slide 3
Lorsque nous avons demandé cette subvention RAPID, nous avions des collaborateurs de l'Open Knowledge Network de la NSF. Nous avions des collaborateurs de l'UC Santa Barbara. C'est en vert. L'UCSF est en orange, et nous en bleu. Nous nous répartissons en quelque sorte les tâches dans des domaines particuliers. Nous nous concentrons sur les domaines bleus tels que les caractéristiques de la population, les données sanitaires, les informations sur les agents pathogènes et les informations environnementales, tandis que nos collaborateurs de l'UC Santa Barbara (il s'agit d'un autre RAPID) se concentrent sur le transport et la chaîne d'approvisionnement. L'UCSF se concentre sur la biomédecine. Diapositive suivante s'il vous plaît.

Slide 4
Voici, en quelque sorte, notre prototype de graphe de connaissances. Dans un graphe de connaissances, il n'y a pas de silos d'informations, mais des liens entre toutes les données. Ce n'est qu'un prototype. Nous n'avons pas encore beaucoup d'informations ici. Sur le côté droit, vous voyez des informations sur le pathogène, l'hôte. Nous disposons d'informations sur l'épidémiologie, sur cette épidémie ici. Nous nous sommes ensuite beaucoup concentrés sur le domaine biomédical. Nous avons plus de 30 000 souches différentes dans ce graphe de connaissances, mais elles sont toutes reliées entre elles. Pour chaque souche, nous connaissons toutes ses variantes de mutations, l'effet sur les gènes et les protéines. Nous connaissons les interactions protéine-protéine. Nous relions tout cela aux publications. Enfin, et c'est le plus important, nous établissons également un lien avec la géolocalisation. Nous avons cartographié l'ensemble de la hiérarchie géographique du monde, de sorte que nous pouvons associer des souches ou des cas à n'importe quel endroit du monde - jusqu'au niveau du secteur de recensement. Ok, diapositive suivante s'il vous plaît.

Slide 5
Lorsque nous avons lancé ce projet, nous voulions qu'il soit automatisé et qu'il puisse être développé par d'autres, de sorte que nous disposons d'un flux de travail très transparent et reproductible. Tout d'abord, nous commençons par des données en libre accès. Nous voulons être en mesure de redistribuer l'information, donc nous commençons par des dépôts de données publiques fiables, puis nous avons créé un processus qui extrait automatiquement l'information et l'intègre. C'est vraiment la clé - l'intégration de toutes ces informations. Nous passons beaucoup de temps [...] c'est là que se fait la plus grande partie du travail. Avec COVID, les choses changent tous les jours, nous avons donc un processus de mise à jour quotidienne. Nous disposons d'un logiciel open source et, chaque jour, nous mettons à jour les informations, nous les intégrons, puis nous les téléchargeons dans un graphe de connaissances qui peut ensuite être interrogé. Nous essayons de suivre les principes d'équité, et tout est ouvert et facilement accessible. Vous pouvez accéder à tous nos logiciels. Il est réutilisable et ainsi de suite. Sur ce, nous pourrions peut-être passer à la diapositive suivante.

Slide 6

Vous savez, une fois que nous avons créé ce graphe de connaissances, il y a plusieurs choses que nous pouvons faire (ou que l'utilisateur final peut faire). Tout d'abord, il est possible d'interroger et de parcourir les connaissances où je trouve des informations. Ce qui est montré ici en haut à gauche. Il s'agit d'explorer l'interaction protéine-protéine entre les protéines virales et les protéines humaines, par exemple. Ensuite, comme il s'agit d'un graphique, vous pouvez interagir si nous l'explorons. En vert, les différentes souches du virus. Dans la géolocalisation particulière, vous pouvez examiner des mutations spécifiques et voir comment elles sont partagées entre les différentes souches, par exemple. Il s'agit d'une analyse plus interactive qu'exploratoire, mais si vous souhaitez effectuer une analyse quantitative plus approfondie, vous pouvez également accéder facilement à toutes ces données dans des carnets de calcul, tels que notre studio ou les carnets Jupyter, pour un type d'analyse plus reproductible. Comme je l'ai mentionné, la cartographie sur la géolocalisation est évidemment très importante pour les corvettes, c'est pourquoi nous accédons également à ces informations par le biais de tableaux de bord. En haut à droite, nous montrons, par exemple, les cas actuels du comté de San Diego ou le nombre de cas prévus. Au centre, nous nous concentrons sur des villes spécifiques et examinons, par exemple, diverses conditions préexistantes dans ces zones : quelle est la prévalence du cancer, du diabète, des maladies cardiaques, etc. et comment cela affecte-t-il la population à risque ? Nous pouvons approfondir la question à l'aide de l'outil "swell". Nous pouvons descendre, par exemple, jusqu'au niveau du secteur de recensement et examiner plus en profondeur les populations à risque, la structure d'âge, etc. Nous vous remercions. Prochaine diapositive, s'il vous plaît.

Slide 7

Nous tenons à remercier la NSF pour son financement, ainsi que le programme Open Knowledge Network auquel nous avons participé et ses collaborateurs. Il est évident que nous recherchons également des collaborateurs. Nous recherchons un certain nombre de collaborateurs, donc si vous avez des ensembles de données ouvertes que vous souhaitez partager avec nous, nous aimerions en discuter. Nous aimons les intégrer. Si vous avez un code qui extrait des données de diverses sources de données, cela nous intéressera, et bien sûr si vous voulez utiliser nos données (et je pense que nous avons déjà parlé à un certain nombre de personnes dans le cadre de ce programme), vous savez que nous voulons aussi entendre parler de vous. Je pense que c'est tout. Dans la fenêtre de chat, je vais coller un lien vers notre graphique COVID pour qu'il soit disponible en ligne et que vous puissiez l'explorer vous-même. Je vous remercie de votre attention.