MIT COVID-19 Datathon: data without boundaries

Eva M Luo ,[1,2] Sarah Newman,[3] Maelys Amat,[4] Marie-Laure Charpignon ,[5] Erin R Duralde,[4] Shrey Jain,[6] Aaron R Kaufman,[7] Igor Korolev,[8] Yuan Lai,[9] Barbara D Lam,[4] Megan Lipcsey,[4] Alfonso Martinez,[10] Oren J Mechanic,[1,11] Jack Mlabasati,[4] Liam G McCoy,[6] Freddy T Nguyen,[12] Matthew Samuel,[13] Eric Yang,[10] Leo Anthony Celi[1,4,14]

For numbered affiliations see end of article.

Correspondence to
Dr Eva M Luo, OB/GYN, Harvard Medical Faculty Physicians at Beth Israel Deaconess Medical Center Inc, Boston, MA 02215, USA; eluo1@bidmc.harvard.edu

Received 30 June 2020
Accepted 14 August 2020

To cite: Luo EM, Newman S, Amat M, *et al*. *BMJ Innov* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjinnov-2020-000492

The COVID-19 virus is a formidable global threat, impacting all aspects of society and exacerbating the existing inequities of our current social systems.[1 2] As we battle the virus across multiple fronts, data are critical for understanding this disease and for coordinating an effective global response. Given the current digitisation of so many aspects of life, we are amassing data that can be extrapolated and analysed for the effective forecasting, prevention and treatment of COVID-19. With responsible stewardship, the tools and data-driven solutions currently in development for the COVID-19 pandemic will serve in the present while providing a much-needed foundation for a data-based response to future outbreaks and disasters.

In response to COVID-19, and using data generated thus far, groups at the Massachusetts Institute of Technology (MIT) in partnership with the American Civil Liberties Union (ACLU) of Massachusetts, Google Cloud, Beth Israel Deaconess Medical Center (BIDMC) Innovations Group and Harvard Medical Faculty Physicians at BIDMC came together to host the MIT Challenge COVID-19 Datathon (COVID-19 Datathon) from 10–16 May 2020. A 'datathon' adopts the 'hackathon' model, with a focus on data and data science methodologies, which promotes collaboration, design thinking and problem solving.[3] In a typical hackathon, participants with disparate but complementary backgrounds work together in small groups for a prescribed and intensive 'sprint', typically over the course of one weekend, to develop a new concept, product or business idea. Subject matter expert 'mentors" oversee and advise the teams. At the conclusion of the event, the teams present to a panel of judges. Winners are selected and are typically awarded seed funding. Datathons differ from hackathons in that the output is data analysis. MIT Critical Data, one of the organising groups of the COVID-19 Datathon, has hosted 36 international healthcare datathons.[4–7]

Building on the successes of the 'MIT COVID-19 Challenge' virtual hackathons, the COVID-19 Datathon was organised as a week-long event with the goal of investigating various data sources to glean insights about the pandemic. The event was divided into five research tracks: (1) Measuring policy impact; (2) Misinformation; (3) Disparities in health outcomes; (4) Epidemiology; (5) 'Megacity' Pandemic Response in New York City (NYC). While datathons and hackathons are typically in-person events, the COVID-19 Datathon was conducted virtually. Using digital communication tools such as Zoom (an online videoconferencing platform), Slack (an online messaging platform), Google Drive (a cloud-based storage platform) and email, the COVID-19 Datathon still managed to generate the creative synergy that is a hallmark of such events. The virtual format even had certain advantages over an in-person event, such as allowing for asynchronous connections between mentors and teams, reducing perceptions of hierarchy and encouraging more democratic participation overall.

The COVID-19 Datathon was advertised through partner organisations and

Figure 1 Map of MIT Challenge COVID-19 Datathon participants (44 countries represented). MIT, Massachusetts Institute of Technology.

personal and academic networks. The organising team selected 297 participants and 77 mentors from 44 countries (figure 1) with balanced representation across self-identified genders, as well as diverse expertise across participants. Teams were created by the organisers to balance team composition across data scientists, clinicians, engineers, designers, project managers and subject matter experts. Open COVID-19 datasets were curated by research track and uploaded on Google Cloud. Potential research questions were crafted by the organising team and mentors. Forty-seven teams of approximately three to six participants each were spread across the five research tracks, and each team identified and refined a research question on which they would focus. Mentors checked in with teams daily to provide feedback and guidance. The event also included midpoint presentations for both immediate and asynchronous feedback from additional mentors. At the end of the week, all teams presented their analyses. Ten teams were selected as semi-finalists to present their work to a panel of judges composed of domain experts from partner organisations and the organising team, with the full datathon cohort as an audience.

Using publicly available datasets (table 1), teams processed, linked and harmonised data, conducted analyses and built models. Such analysis required significant work to unpack, interpret, validate and reconcile data across heterogeneous sources. With emphasis on reproducibility, teams were required to submit their code repositories and notebooks for review. The COVID-19 Datathon projects reflected a wide breadth of research outputs. Project 'COVID-19 Patient Severity Index' evaluated 4000 patients across four datasets and developed a way to stratify patients based on comorbidities and other demographics to predict risks for mortality and hospital length-of-stay while identifying biomarkers that best correlate with mortality predictions. Another project, 'Reopening of super-spreader businesses and risk of COVID-19 transmission',[8] classified businesses as 'super-spreaders' through the development of a Transmission Risk Index based on data that captured both frequency and duration of visits to businesses pre-pandemic. The team

Table 1 Select publicly available datasets used in the MIT challenge COVID-19 Datathon

Source	Dataset
Johns Hopkins University	Center for Systems Science and Engineering (CCSE) COVID-19 Epidemiological Data Repository
European Centre for Disease Preventionand Control (ECDC)	Epidemiological Data
WHO	Case and Death Data
World Bank	Healthcare Indicators of Interest
New York Times	US State-Level and County-Level COVID-19 Count Data
Safegraph	Open Census Data
US Census Bureau	American Community Survey
New York City Metropolitan Transportation Authority (NYC MTA)	Mobility Data
NYC Department of Health	Community Health Survey Public Use Data
NYC Department of Health	Facility Database
NYC Department of Health	Emergency Medical Services (EMS) Incident Dispatch Data
Google	Search Data
University of California, Los Angeles (UCLA) Law	COVID-19 Behind Bars Project
Vera Institute of Justice	COVID-19 Jail Dataset
Citibike	Mobility Data
GDelt Project	COVID-19 News Dataset
The COVID Tracking Project	COVID Racial Data Tracker
ProPublica	Clinical Trials: Participant Demographic Data
University of Southern California	COVID Tweet IDs
University of California, Berkeley	COVID Exposure Indices

MIT, Massachusetts Institute of Technology.

then tested the association between super-spreader businesses and rates of COVID-19 cases. In a project entitled 'Can your zip code affect your chances of getting COVID-19?', the team employed unsupervised learning to cluster zip codes in New York based on 240 features including commuting, family composition and income data, and evaluated the clusters with respect to number of cases and deaths. A number of projects will continue beyond the COVID-19 Datathon and will continue to share their code repositories.

Many of the projects had immediate policy implications for the public and private sector. One of the projects cited above, 'Reopening of super-spreader businesses and risk of COVID-19 transmission', has already, only 1 week after the datathon, been incorporated into predictive models at Beth Israel Deaconess Medical Center, an academic medical centre in Boston, to help prepare for a possible second wave of infections as social distancing measures are relaxed. The ACLU of Massachusetts also plans to direct findings from the COVID-19 Datathon to policy and activism organisations.

The COVID-19 Datathon is one example of how data scientists, healthcare professionals and engineers from around the global community can gather, virtually, to pool their resources and successfully collaborate on analyses using publicly available data. The virtual nature of the COVID-19 datathon permitted certain benefits, including the ability to reach a broader range of experts, and allowing busy frontline clinicians and public health practitioners to participate and connect with data scientists asynchronously as their schedules allowed. We are currently living in an unprecedented time; this is not the first global pandemic, but it is the first one with real-time global interconnection, communication and the collection of massive amounts of data. Learning from the data, responsibly and across disciplines, in combination with communication, education, treatment and policy decisions, are our best ways forward to defeat this virus while laying the groundwork for collaborative data science in the face of future calamity.

Author affiliations
[1]Harvard Medical Faculty Physicians at Beth Israel Deaconess Medical Center Inc, Boston, Massachusetts, USA
[2]OB/GYN, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[3]metaLAB, Berkman Klein Center, Harvard University, Cambridge, Massachusetts, USA
[4]Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[5]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[6]Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
[7]New York University—Abu Dhabi Campus, Abu Dhabi, UAE
[8]HealthDSA: Health Data Science and Analytics Community, Boston, Massachusetts, USA
[9]Urban Science and Planning, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[10]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[11]Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[12]Massachusetts Institute of Technology, Boston, Massachusetts, USA
[13]Harvard University T H Chan School of Public Health, Boston, Massachusetts, USA
[14]Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Twitter Eva M Luo @EvaMLuo, Shrey Jain @@shreydjain13 and Freddy T Nguyen @freddytn

Acknowledgements We thank our global COVID-19 Datathon mentors for donating their time and expertise to fighting COVID-19 with us.

Contributors All authors listed meet ICMJE criteria for authorship. All authors listed contributed to the planning, writing and editing of this manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Eva M Luo http://orcid.org/0000-0002-0729-3665
Marie-Laure Charpignon http://orcid.org/0000-0002-5786-2627

REFERENCES

1 Dorn Avan, Cooney RE, Sabin ML. COVID-19 exacerbating inequalities in the US. *Lancet* 2020;395:1243–4.
2 Gould E, Shierholz H. Not everybody can work from home: black and Hispanic workers are much less likely to be able to telework. Economic Policy Institute. Available: https://www.epi.org/blog/black-and-hispanic-workers-are-much-less-likely-to-be-able-to-work-from-home/ [Accessed 23 May 2020].
3 Aboab J, Celi LA, Charlton P, *et al*. A "datathon" model to support cross-disciplinary collaboration. *Sci Transl Med* 2016;8:333ps8.
4 Badawi O, Brennan T, Celi LA, *et al*. Making big data useful for health care: a summary of the inaugural MIT critical data conference. *JMIR Med Inform* 2014;2:e22.
5 Organizing Committee of the Madrid 2017 Critical Care Datathon, Núñez Reiz A, Martínez Sagasti F, *et al*. Big data and machine learning in critical care: opportunities for Collaborative research. *Med Intensiva* 2019;43:52–7.
6 Serpa Neto A, Kugener G, Bulgarelli L, *et al*. First Brazilian datathon in critical care. *Rev Bras Ter Intensiva* 2018;30:6–8.
7 Li P, Xie C, Pollard T, *et al*. Promoting secondary analysis of electronic medical records in China: summary of the PLAGH-MIT critical data conference and health Datathon. *JMIR Med Inform* 2017;5:e43.
8 O'Donoghue A, Dechen T, Pavlova W, *et al*. Super-spreader businesses and risk of COVID-19 transmission. *medRxiv* 2020.