MIT COVID-19 Datathon: data without boundaries


The COVID-19 virus is a formidable global threat, impacting all aspects of society and exacerbating the existing inequalities of our current social systems. 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.

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.

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 video-conferencing 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
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’, 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

---

**Figure 1** Map of MIT Challenge COVID-19 Datathon participants (44 countries represented). MIT, Massachusetts Institute of Technology.
that 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 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
1Harvard Medical Faculty Physicians at Beth Israel Deaconess Medical Center Inc, Boston, Massachusetts, USA
2OB/GYN, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
3metaLAB, Berkman Klein Center, Harvard University, Cambridge, Massachusetts, USA
4Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
5Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
6Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
7New York University—Abu Dhabi Campus, Abu Dhabi, UAE
8HealthDTS: Health Data Science and Analytics Community, Boston, Massachusetts, USA
9Urban Science and Planning, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
10Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
11Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
12Massachusetts Institute of Technology, Boston, Massachusetts, USA
13Harvard University T H Chan School of Public Health, Boston, Massachusetts, USA
14Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
15School of Public Health, Harvard University, Cambridge, Massachusetts, USA
16Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA
17ProPublica
18Northwestern University, Evanston, Illinois, USA
19Vera Institute of Justice COVID-19 Jail Dataset
20City of Los Angeles Department of Public Health
21San Francisco Department of Public Health
22Los Angeles County Department of Public Health
23Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
24School of Medicine, University of California, Los Angeles (UCLA), Los Angeles, California, USA
25Department of Biostatistics, University of California, Berkeley
26University of Southern California COVID Tweet IDs
27Department of Epidemiology, University of Southern California
28Palantir Technologies, New York, New York, USA
29MIT, Massachusetts Institute of Technology.

University of California, Berkeley COVID Exposure Indices

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

<table>
<thead>
<tr>
<th>Source</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Johns Hopkins University</td>
<td>Center for Systems Science and Engineering (CSE) COVID-19 Epidemiological Data Repository</td>
</tr>
<tr>
<td>European Centre for Disease Prevention and Control (ECDC)</td>
<td>Epidemiological Data</td>
</tr>
<tr>
<td>WHO</td>
<td>Case and Death Data</td>
</tr>
<tr>
<td>World Bank</td>
<td>Healthcare Indicators of Interest</td>
</tr>
<tr>
<td>New York Times</td>
<td>US State-Level and County-Level COVID-19 Count Data</td>
</tr>
<tr>
<td>Safegraph</td>
<td>Open Census Data</td>
</tr>
<tr>
<td>US Census Bureau</td>
<td>American Community Survey</td>
</tr>
<tr>
<td>New York City Metropolitan Transportation Authority (NYC MTA)</td>
<td>Mobility Data</td>
</tr>
<tr>
<td>NYC Department of Health</td>
<td>Community Health Survey Public Use Data</td>
</tr>
<tr>
<td>NYC Department of Health</td>
<td>Facility Database</td>
</tr>
<tr>
<td>University of California, Los Angeles (UCLA) Law</td>
<td>COVID-19 Behind Bars Project</td>
</tr>
<tr>
<td>Vera Institute of Justice</td>
<td>COVID-19 Jail Dataset</td>
</tr>
<tr>
<td>Citbike</td>
<td>Mobility Data</td>
</tr>
<tr>
<td>GDelt Project</td>
<td>COVID-19 News Dataset</td>
</tr>
<tr>
<td>The COVID Tracking Project</td>
<td>COVID 19 Case Counts Dataset</td>
</tr>
<tr>
<td>ProPublica</td>
<td>Clinical Trials: Participant Demographic Data</td>
</tr>
<tr>
<td>University of Southern California</td>
<td>COVID Tweet IDs</td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>COVID Exposure Indices</td>
</tr>
<tr>
<td>MIT, Massachusetts Institute of Technology</td>
<td></td>
</tr>
</tbody>
</table>
covid-19 pandemic or until otherwise determined by BMJ. You may use, download and print the article for any lawful, non-commercial purpose (including text and data mining) provided that all copyright notices and trade marks are retained.

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

REFERENCES