Wellcome Data Science Ideathon

Participant Pack

26 June–12 July 2023

Table of contents

Instructions

This challenge pack contains all the information you need to know to tackle one of the Wellcome Ideathon challenges. In this document we will go through all the requirements that must be met for you to submit an eligible solution to your chosen challenge as well as the judging criteria that will be used to evaluate your solution.

You may select **one** challenge from those listed in the challenge pack to tackle as part of the Ideathon. Your challenge must be aligned with the health challenge you selected when applying to the event.

Your solution to the challenge should consist of:

1. Reproducible, open-source code that can be run to demonstrate your **prototype**

2. A fifteen-minute **presentation**, which should include a demonstration of your prototype

3. Slides to accompany your **presentation**

4. **For researcher teams only:** A **proposal** (maximum 3 pages) outlining your future plans

By 17:00 (BST) on 11 July 2023 you must have [tagged](#) your GitHub repository with the tag "submission" and this should include your final prototype code, slides, and proposal (if you are a researcher team). In addition, you must **email your slides** to ideathon@wellcome.org.

You can start working on your solution immediately.

Solutions, presentations and proposals

GitHub

All members of your team will have been sent invitations to join a private GitHub repository, which will be named after your team. The content in this repository should be licenced under MIT and CC-BY 4.0. Wellcome staff will have access to this repository, but other teams will not. Content in this repository will be made public after the event.

Prototype

Full details of the type of prototype you may want to create are given in the individual challenges. However, there are some limitations and restrictions that must be considered:

1. **Reproducibility** – All code you produce should be reproducible – you can use notebooks or scripts (with set seeds) to achieve this.

2. **Real world datasets** – Most challenges during the event will have associated datasets that have been taken from real world data. Make sure you have clearly read the challenge GitHub repository (linked with the challenge) clearly so you know which datasets can be used and which must be used. For some challenges, no real-world data is provided. You cannot use any real-world data that we do not provide.

3. **Simulated datasets** – For all challenges, you are permitted to generate your own datasets for the purposes of developing, testing, and demonstrating your prototype. However, there are further conditions attached to how you may do this, see WellcomeIdeathon2023/simulated_data for full details.

Presentation and slides

You can choose to structure your slides and presentation however you wish, but you may want to consider spending:

- The first half presenting your prototype and discussing how it addresses the "Challenge" heading of your chosen challenge.
- The second half presenting slides that answer each of the questions under the "Questions to answer in your presentation" heading of your chosen challenge.

We do not expect all participants to present however we do expect you to address how all members contributed to your solution.

Presentation accessibility: When choosing your slide template, please consider your choice of font, and the size, colour and contrasts used in both the text and in any figures, ensuring that all text and figures will be clearly readable to all members. We recommend using a widescreen aspect ratio (16:9). We do not allow movies/animations and/or transitions and they should not be included. We will display the content from one laptop using the submitted slides.

Proposal (researcher teams only)

Your proposal must be a maximum of two pages (including figures and tables) and should outline how you would spend £100,000 to further develop and implement your proposal. Your budget will be non-binding and if you win you will be expected to submit a full budget up to the winning award amount (£80K, £100K, or £120K). There is no requirement about how the proposal is structured but as a helpful guide you may consider:

1. **Background (max 0.5 pages):** Short paragraph summarising the background to your chosen challenge in your own words

2. **Prototype (max 0.5 pages):** A description of your submitted prototype (remember this document will be read before you have given your presentation)

3. **Plans for future work (max 2 pages):** A description of the work you will cover over the one-year grant period, which may consist of answers from your presentation.

   a. **For teams tackling mental health challenges:** Include a description of how you would work with people with lived experience to test and co-develop your prototype at multiple stages of the project (e.g. design, delivery, governance, dissemination).

   b. **For teams tackling climate and health challenges:** Include how you would communicate with policymakers and other key stakeholders to ensure your developed tools are accessible and user friendly (or 'co-produced').

   c. **For teams tackling infectious disease challenges:** Include how your developed platform/pipeline/dashboard will be applicable/usable in low-resource contexts and how you'll work with researchers in LMICs to co-develop the platform.

4. **Budget:** A table providing a non-binding budget for how you would spend £100K to develop your prototype further – we have provided a template you can use on the next page.

Budget Template

Salaries:

Name	Justification	Period on project (months)	% time	Total

Materials and consumables:

Description	Justification	Total

Equipment:

Description	Justification	Total

Access charges:

Description	Justification	Total

Travel and subsistence:

Description	Justification	Total

Miscellaneous:

Description	Justification	Total

Summary of costs requested:

Description	Total
Salaries	
Materials & Consumables	
Equipment	
Access charges	
Travel and subsistence	
Miscellaneous – other	
Grand Total	

Evaluation

Solutions will be judged according to the following criteria:

Criteria	Prototype	Presentation	Proposal (Researcher teams only)
Feasible			Plans for future work are feasible given presented budget and timeline.
Innovative	Prototype is innovative and novel.		
Accessible	Prototype is user-friendly and aligns with Wellcome's trustworthy data science goals.	Aligns with Wellcome's diversity, equity and inclusion, and trustworthy data science goals.	
Thoughtful	Does the prototype clearly address the set challenge?	Clear and well-thought-out responses to the "Questions to answer in your presentation".	Plans are well-thought-out with a clear direction for future work.
Cohesive		Is it clear how all team members contributed to the solution?	Every member of your team will have clear involvement in future development.

Ideathon challenges

Mental health

Mental health challenge 1

Develop a platform for long-term participant retention

Background

Retaining participants with mental health problems in mental health and in clinical research studies is notoriously difficult. Participant dropout may result in trials underpowered to detect a treatment effect and can be a particular issue when there is differential dropout among treatment arms. This problem is compounded in longitudinal studies that require participants to remain in the study for long time periods (for example, 10+ years). Potential solutions to this must consider individual participant needs around communication, for example some participants may prefer telephone calls, others may prefer messaging on an app, and others may prefer in-person conversations.

Challenge

[Github repository](#)

Design an approach to improve long-term retention of participants with mental health problems in a study. In the long-term, we are interested in both researcher- and participant-facing ideas that can help with long-term (10+ years) retention. However, for your prototype you should focus on the front-end for either researchers or participants. For researchers this could include (but is not limited to), logging of participant information and requirements, notifications, and reminders. For participants this could include (but is not limited to) machine learning (for example, for adaptive user-design), gamification, preferences, and much more. You should focus particularly on tools that can be used in early intervention studies in adolescents and young adults.

Questions to answer in your presentation

1. How does your design improve the long-term retention of participants with mental health problems in a clinical research study?

2. How would you evaluate the impact of your application? How could you discover early utility of the application and likelihood of success in ten years within a shorter time period?

3. How have you considered the needs and requirements of research participants and researchers?

Mental health challenge 2

Develop an interactive application for finding and accessing clinical trials

Background

At any point in time there are many clinical trials seeking volunteer participants. These inform scientific progress and provide participants with access to cutting-edge, experimental treatments as well as payment for their time. However, uptake in mental health trials is not as high as for other conditions, such as cancer. Furthermore, if a participant did want to participate in a trial, current platforms are not particularly visible or easy to find and are limited in scope. For example, the MQ Mental Health platform is limited in the number of trials it offers and the NIHR platform is not user-friendly. Other prominent examples include NHS DigiTrials and NIHR's Be Part of Research – whilst both these platforms are user-friendly, they are not mental health-focused and are restricted to the UK. Moreover, they do not cover the entire end-to-end user journey, and do not consider the range of possible users, including both research participants and researchers. To accelerate research, there needs to be a cultural shift towards more participants with mental health problems wanting to participate in, and having access to, ongoing clinical trials.

Challenge

[Github repository](#)

Design an application (app) to help people with mental health problems search for ongoing and upcoming clinical trials that they can participate in. As with the previous challenge, in the long-term we are interested in an app that can be valuable to the full range of researchers and participants listed above, but for now please just focus on meeting the needs of participants from diverse backgrounds. The app should enable a stronger connection between participants and researchers by removing dependencies on third parties (for example, social media sites) and allowing potential participants to filter by preference (for example type of intervention, frequency of visits/data collection sessions). We are particularly interested in solutions that can be utilised for the entire user journey from discovering trials and signing up, to updates/notifications from the research team and follow-up surveys and more. For your prototype you can focus on developing a platform for finding and accessing trials, but you should also consider how your application could be broadened into a comprehensive resource that can be used for: public engagement, communications and outreach, educational tools and resources, and the platform for finding, accessing, and taking part in trials.

Questions to answer in your presentation

1. How does your design improve upon existing platforms?

2. How will your platform ensure it attracts people from diverse backgrounds and with heterogenous symptoms and interests?

3. How have you considered the needs and requirements of research participants and researchers?

Mental health challenge 3

Design an intervention builder for the development of digital interventions

Background

Experiments in behavioural science (psychology, neuroscience, etc.) often ask participants to complete computer-based behavioural tasks. Developing these tasks is made possible for researchers with relatively little programming experience by task builders that simplify the development process (for example, jsPsych, PsychoPy, and psychtoolbox). Unfortunately, there are no equivalent tools to aid mental health researchers in the design and development of digital mental health interventions, such as interventions using anything from mobile applications, to VR and digital sensors. Where behavioural tasks typically present a straightforward and linear sequence of screens in a single session, interventions often include more involved features (such as an online chat with a therapist) and are presented over multiple sessions (requiring features like notifications and progress tracking). Given these added requirements, existing task builders cannot be applied to building interventions in a straightforward way. This limits intervention development to mental health research groups that have resources to hire software developers or researchers with relevant expertise. Therefore, there is a need to develop a tool that will lower the barrier to entry to digital intervention development.

Challenge

[Github repository](#)

Design an **intervention builder** that will aid mental health researchers in developing digital interventions. You should not reinvent the wheel by building features that are already offered by existing task builders. Instead, where possible, you should build on what already exists and focus on developing features that digital interventions need but existing tools do not offer. A simple example would be integrating existing task builders to design individual sessions and then sequencing these together over time with your intervention builder. Your tool must consider data privacy where required (for example, for progress tracking applications) and provide (or integrate with) safe data storage solutions. As well as thinking about building new interventions, your tool should also be able to store (locally or online) and share workflows/pipelines that are developed on your platform (but should not force users to do this). For your prototype we recommend focusing on features that are most lacking to showcase the power of your tool.

Questions to answer in your presentation

1. How does your tool help mental health researchers develop digital interventions?

2. How have you minimised the programming expertise required to interact with your tool?

3. How does your tool integrate with existing tools, frameworks, or platforms that researchers might want to use when developing digital interventions?

Ideathon challenges

Climate and health

Climate and health challenge 1

Accessible development of metrics and models for assessing the economic impacts of climate change

Background

Climate change has direct impacts on human health, for example due to global heating and extreme weather events, as well as indirect impacts by increasing the spatial distribution and incidence of climate-sensitive infectious diseases. Climate change can also influence human health by economic impacts.
For example, by inflation, job loss, and diversion of financial resources to restore infrastructure destroyed by global heating – all of which have direct knock-on effects on human health. To help set and optimise targets that drive action to minimise the harmful direct effects of climate change on human health, Wellcome is funding a project to standardise climate-health metrics that can be used for agenda setting and policymaking. Now we are looking to tackle the negative effects on human health that result from the economic impacts of climate change. We are interested in understanding how to develop standardised metrics for measuring the economic impacts of climate change that can be utilised for policymaking and model development.

Challenge

[Github repository](#)

Use the provided datasets to demonstrate how data science approaches could be used to identify correlations and metrics that can be used for inference/prediction of the economic impacts of climate change on health. You might consider representation learning for generating new relationships between variables, reinforcement learning for identifying objectives to optimise, unsupervised learning for clustering, supervised learning for identification of predictors, or any other method, including non-ML methods.

As well as identifying the potential of these models, demonstrate how these could be selected by people without coding experience on a no-code open-access platform where users can bring their own economic and climate datasets, run models and workflows on top of these, and then visualise, analyse, and download results. For any identified climate-economic metrics (which need not be verified in practice for this event), demonstrate how these could practically be utilised in model development and optimisation and for long-term model deployment.

Questions to answer in your presentation

1. How would your platform allow users to demonstrate that any metrics/models developed on the platform are useful in the real-world?

2. How will you ensure your results generalise to continually updated climate and economic data as well as new data sources?

3. How will your platform, in particular analysis and visualisation tools, handle uncertainty and variability in climate data?

Climate and health challenge 2

Develop a platform for accessing and augmenting climate datasets with other data sources

Background

Many climate datasets exist however these are not necessarily easy for researchers to use. One difficulty is that climate data are found over many different spatiotemporal scales which means that most datasets are not immediately compatible with each other. This is especially true when considering how climate data can be used to augment health data, financial data, and other data sources. Data cleaning is the starting point of any analysis however for climate data even this can provide a barrier to entry, we want to remove this barrier.

Challenge

Github repository

Create a sophisticated data infrastructure platform for multiple data sources across differing spatiotemporal scales that includes Extract Transform Load pipelines for loading different data types and for data augmentation. Your platform should be as flexible as possible and should provide generic pipelines for custom pre-processing. Your platform should enable storing and sharing pipelines (but should not force users to do so) and should allow users to upload their own data. As an example (and you do not need to do this as part of the challenge) your platform could enable workflows such as:

- Upload Electronic Health Record (EHR) data with hospital-level data measured on a daily scale to a local environment ensuring the data will not be shared further;

- Collapse the data to be on a monthly scale;
- Pull methane emission data from an online data source measured at a national level on an annual scale and then interpolate this to a monthly scale;
- Augment the EHR data with the methane data by using hospital look-up information to match hospital locations;
- Visualise the augmented data on the platform and perform basic analysis using a graphical interface;
- Download processed data for further analysis and save pipeline.

Questions to answer in your presentation

1. How can your platform handle multiple types of data and how can users choose how data can be collapsed and interpolated? How will you trade-off user-choice against design decisions that favour optimal methods?

2. How would you visualise complex data across multiple spatiotemporal scales?

3. How can users search and access data and be informed about benefits and limitations (including possible biases) of data and data types across different scales?

Climate and health challenge 3

Develop a dashboard for tracking the effects of methane emissions on health

Background

Methane emissions are the biggest contributor to the formation of [ground-level ozone](). Methane is also a greenhouse gas that causes global heating and is [80 times more potent than CO2](). Reducing methane can be driven by societal and agricultural changes (such as farming less meat) but technology also has a big role to play, for example, by utilising machine learning to [optimise infrastructure repair]() or to [quickly detect and prevent methane leaks](). Despite advances and interest in the area, there has not been a lot of research linking methane and health data in an accessible way that can be used by clinicians or policymakers to adapt to and mitigate the increasing public health pressures from methane emissions.

Challenge

[Github repository]()

Use the provided datasets to create a dashboard that can be used by policymakers and clinicians to understand and track the impact of methane on health – in particular [respiratory disease]() and [mental health](). The platform should be user-friendly and not require expert climate knowledge but should provide enough information to enable evidence-driven mitigation strategies to improve the health of populations. For example, how might a clinician or policymaker use your platform to prepare for increases in number of patients to respiratory wards in hospitals? How might a policymaker leverage the platform to lobby companies to quickly fix leaky pipelines? In addition, your platform should be modular and interoperable so multiple data sources can be incorporated for health and methane data of distinct types, for example suitable for satellite and on-the-ground methane sites. Moreover, your platform should also allow users to upload their own data that could be overlayed onto downloaded methane data.

Questions to answer in your presentation

1. How has your platform presented complex methane and health data (and their relationship) in a way that can be understood by key stakeholders but without 'dumbing down' the information?

2. How would you ensure your platform is accessible, usable, and sustainable and that data can be shared safely and well credited?

3. How will your platform integrate datasets with different time and geographical scales?

Ideathon challenges

Infectious disease

Infectious disease challenge 1

Develop a platform for discovery and analysis of correlates of protection

Background

Vaccine development takes 10-15 years on average [1], though for some infections that are either rare, slow to progress to disease, or difficult to track, traditional vaccine development methods are extremely challenging. One prominent example is Nipah virus (NiV), for which traditional vaccination efficacy trials are not possible and in fact one study estimated it would either take (depending on trial type) 516 years and 163,000+ vaccine doses or 43 years and 1.83million doses or 7 years and 2.5million doses to run a vaccination trial for NiV. One method to speed up the vaccine development pipeline is to consider correlates of protection as surrogate endpoints in trials. Correlates of protection are typically measurable immune responses that are statistically correlated with protection against disease or infection by a pathogen. Finding correlates of protection and being able to use them as surrogate endpoints can allow the vaccine development pipeline to be accelerated by either utilising a surrogate endpoint to allow a trial to be concluded early, or by informing future decisions or likelihood of success of a vaccine for individual immune responses (thus allowing more targeted future development). Platforms such as ImmPort provide access to study data that can be used to identify correlates of protection. The existing analysis tools in ImmPort Galaxy and in ImmuneSpace (another similar platform) aid identification of immune correlates **within a study dataset**. They could be made more user-friendly for non-coders / non-statisticians or could incorporate more advanced methods based on machine learning to identify and analyse correlates **across study datasets**.

Challenge

Github repository

Use the provided studies that have been downloaded from ImmPort to develop and demonstrate the power of a platform that can take data in a standardised format and provide visualisations and analysis methods to identify correlates of protection. In the long run (for instance, post-Ideathon), this platform should include: 1) user-friendly tools (in a no-code plug and play style) for identifying correlates within connected study datasets; 2) tools for identify correlates across unconnected study datasets to identify how correlates are similar and how they differ according to key variables (for example, participant and/or vaccine age). For your prototype, you only need to demonstrate the ability to do one of these. This platform should provide more analysis methods than the current status quo platforms (above) and should be able to offer both classical statistical methods (for example, Ms) and machine learning solutions (for example, reentation learning, reinforcement learning, supervised learning, and clustering) – you are not expected to create bespoke models or methods but to interface with existing packages. Your platform should be user-friendly and provide outputs that can be easily understood by downstream stakeholders including regulatory authorities such as MHRA, FDA and others. In your presentation you should demonstrate your prototype in action by uploading data, using your interface to aggregate data and identify potential correlates, and visualise and validate your findings.

Questions to answer in your presentation

1. How will you work with biostatisticians (those who advise regulatory authorities) in the long-term so your platform can help to formalise the discovered markers?

2. How will you ensure your modelling pipelines generalises to new diseases/vaccines?

3. How would you propose working with government stakeholders (including policymakers and regulators) to ensure that ML methods are embraced, and findings are trusted? How would you balance this against community needs (think about how the public was hesitant about steps being skipped in the accelerated COVID vaccine pipeline)?

Infectious disease challenge 2

Develop NLP pipeline for vaccine uptake sentiment analysis

Background

Social media has demonstrated its potential to reflect and influence public opinion through the spreading of information, malinformation, misinformation, and disinformation. Several research groups have explored the use of social media to monitor and evaluate public sentiment towards vaccines but there are challenges to overcome to do this in a standardised and robust manner, and it's unclear how it could translate through to the design and implementation of interventions to increase vaccine confidence and uptake. Natural language processing (NLP) is a cutting-edge machine learning method that has repeatedly been demonstrated to be a powerful technique for sentiment analysis, knowledge extraction, and other methods related to natural language / free text. We are interested in understanding how NLP can be utilised to analyse social media posts to understand and track the association between vaccine uptake and public sentiment, and to inform interventions that can increase uptake.

Challenge

Github repository

Develop an NLP pipeline to analyse public sentiment and messages on social media to understand the association with real-time vaccine uptake. As well as real-time uptake, the algorithm should include predictive methods for alerting stakeholders to decreasing sentiment (or when a critical mass of 'negativity' has been reached) in order to predict short-term and long-term 'shocks to immunisation systems'. You may also want to consider baseline heuristics as

well as advanced machine learning solutions, for example triggering warnings for negative sentiment increasing rapidly (for example, increase of X tweets/day). Insights from the system should be useful for real-time modelling to understand current sentiment, as well as for informing the design and monitoring impact of interventions to increasing vaccine uptake and/or preventing shocks to the immunisation system. For example, you may want to consider automated suggestions of interventions and/or causal inference methods for analysing the success of interventions and/or monitoring of sentiment based on interventions (for example, alysing replies to influencer's tweets). The developed pipeline should include a public-facing platform that includes a clear description of the underlying algorithm, visualisations of findings, and lay explanations of the observed effects and any proposed interventions.

Questions to answer in your presentation

1. How would you engage with the public to ensure that the platform is minimally invasive and not perceived as a negative tool?

2. How does your proposed pipeline and platform handle predictions in real-time as well as longer time? How would you incorporate uncertainty/ confidence of predictions in triggering alerts?

3. How will you validate any classifier in your platform to distinguish genuine sentiment and deliberate anti-vaccination malinformation ('trolls') accurately?

Infectious disease challenge 3

Develop a surveillance pipeline to optimise multiple data sources

Background

Once the scale of the COVID-19 pandemic became clear, the UK Government convened the Scientific Advisory Group for Emergencies (SAGE). SAGE is comprised of multiple working groups, the largest of which are the Scientific Pandemic Insights Group on Behaviours (SPI-B) and the Scientific Pandemic Influenza Group on Modelling (SPI-M). SPI-M ran complex computation models to advise the government on possible future trajectories of COVID-19 and how various interventions may affect the course of the virus. In contrast, SPI-B performed behavioural science experiments and analysis. These working groups remained separate within SAGE and there was minimal interaction between the two. However, closer monitoring of non-traditional 'big data' sources, such as behavioural data, retrospective analyses, and qualitative analyses, may improve predictive ability and policymaking for future disease outbreaks.

Challenge

Github repository

Build tools to integrate data from multiple sources for case-based surveillance. Use the provided datasets to demonstrate how quantitative and qualitative data can be combined to increase the predictive potential of models to better inform public health response. This should cover everything from integrating with APIs, data aggregation, modelling, all the way to public communication. You may consider anything from NLP for qualitative analysis and incorporation into a quantitative machine learning model, to guidelines for communicating qualitative data alongside quantitative data without diminishing the importance of both

forms of data, to causal analysis to understand how quantitative and qualitative data affect each other. Your platform should be scalable and extensible and in the long run, (for instance, developed after the Ideathon) should be able to incorporate multiple quantitative and qualitative data sources such as interview data, behavioural analysis, genomic data, climate data, and more. Your platform should also allow pipelines to be created for people without coding experience and pipelines using public APIs should be sharable (with creators being citable). Finally, you should consider how your pipeline could incorporate, and adapt to, new data should there be changes in external factors, for example it was observed during the pandemic that when models forecasted high-levels of infections then people would act more cautiously thereby ensuring the forecasts did not come true – how would you incorporate this in your model (note you do not need to consider this exact scenario)?

Questions to answer in your presentation

1. How do you propose to transparently evaluate your model and to demonstrate its performance (good or bad) over time?

2. In the long run, how would you engage with the public to ensure that forecasts from the model are not viewed as scientific jargon or overly pessimistic?

3. How would you broaden your prototype to include data from other sources ranging from quantitative databases to more social media?

Wellcome supports science to solve the urgent health challenges facing everyone. We support discovery research into life, health and wellbeing, and we're taking on three worldwide health challenges: mental health, infectious disease, and climate and health.