Name	Folder	Charter	Minutes	Calendar feed	Mail-list	Contacts
Data Science	wg-data-science-folder	wg-data-science-charter	wg-data-science-minutes	/cal-wg-data-science.ics	wg-data-science@lists.hubmapconsortium.org	Cole Trapnell - Co-chair | coletrap@uw.edu Nils Gehlenborg - Co-chair | nils@hms.harvard.edu Maggie Vella - Knowledge Manager | Margaret_Vella@hms.harvard.edu

Please use this time to best suit the needs of your Working Group. We recommend introducing team members that have not met, reviewing past achievements of the group, upcoming plans, and ensuring an action plan is in place with members assigned. At a minimum please review your charter and capture below the following information for NIH. Please take notes in this document as a record of the discussion.

Suggested agenda:

- Roll call; introduce new members
- Progress during the past year (please summarize key achievements; a few sentences will suffice)
- Plans for the upcoming year
  - What are they key activities planned for the upcoming year? (a few bullet points will suffice)
  - Are there action items from the WG that need to be presented to the SC?
  - Will additional resources be needed for any activities?
- Administrative issues

  - Internal communication - are channels working?
  - Any changes to charter? (https://drive.google.com/open?id=1kqCJ13KNwLj0g2IDpQukoopqyM5JGy7sX-scSXPsEro)
  - Meeting schedule changes?
  - Do you want to rotate co-chairs ./ knowledge manager?
- Any Other Business

Notes:

- We imagined the kind of questions a cell biologist coming to the HuBMAP portal would ask
  - What are all of the cell types that the given cell type comes into contact with in this anatomical region?
  - biology language, and Bob has been trying to come up with a query language
  - He has written up a document, more formal and precise than our initial anecdotal questions
  - Bob imagines that the first thing written down in any query is the types of records you want to fetch (chromatin accessibility patterns, cell types, etc)
  - Next, you will add constraints from metadata, or anatomical in nature, for example
- Questions:
  - The metadata issue: which metadata, etc?
    - Outsourced to the data release teams? Seems unclear, were initially charged to not focus on metadata
    - We need definitions of metadata
  - How do you imagine the query language interacting with data that is not directly available (e.g. cell type)?
    - Suppose you want a list of differentially expressed genes. What tests?
    - We need to make decisions on the types of statistical tests applied
    - Others outside may not be satisfied with the given decisions made
  - Is there a way to make a list? (cell type, etc)
    - We hope to have a formal specification written down in parallel to the work getting done in the DRT's
  - Is the raw data query going to be done on NIH servers? (worry of computational intensity). Who does computational heavy lifting?
    - No
    - Processing - can be done, doing a raw image query can take hours to run, do we limit anybody because it can block everybody else
    - Data depth too large unless you have spun up VMs
    - Raw data query Nils envisions to be a query of the files, not the content - defer to IEC team to tell us what is possible/not

        - Expectation of more granular search - build in as something that can be cued and run later (makes human refinement community expectation a challenge)
        - TCGA model - very structured queries with downloads, raw queries would be off the books. Is this the way to go? Stephen (consultant) cannot advocate for downloading the data - NCI initiative, download clogs up everything
        - Can assume that users will come in and download the whole thing (maybe initially feasible), maybe this is fine for first release and we may need to support this behavior
    - How will the user interact with the query language?
        - Could be a query they type in
        - Could also have UI components that allow you to assemble query on the fly
    - Is this achievable in the following 9 months?
        - Probably not to a thorough extent by june
        - Will not be able to query across the whole database by the release
    - There will be a need for creating UI to lower the barrier to entry for users just getting acquainted with HuBMAP
- Most people will not have the experience to navigate IP codes
- Why not have a simple python API?
    - There are people who are not familiar
    - We will define a language that could be implemented as a python package
    - We could also launch Jupyter Notebook
- What is the net output that the data science WG is supposed to deliver for JUne?
    - Need to come up with a better draft of the query language
    - Push to have some of these queries implementation as part of the initial data portal
    - The expected output is that initial draft query language
    - Everything that is needed so an initial user can go in and see what {proteins} they want to start looking for in {cell type}, very basic and modest list of queries we could support
- Do you feel compelled to release every data type, or just expand on the ones with depth?
    - TMCs were tasked with listing the data expected to be available by June - find that list
    - Focus less on what data, and build this broader, define a language, then focus on implementation and go deeper
    - Inf that goes deep on one aspect, and on the surface across the top, you create a matrix and fill in gaps in coming years
- Executive Summary
    - Points of consensus

- Agreement that a structured query language serving up HuBMAP data would serve a constituency of users
  - Some would not want to use it, but many would
- Points of contention
  - Many questions regarding the scope of the query language, and how it will grow in the coming months
  - What will it do (if anything) come june
  - Who is really responsible for defining the metadata?
  - What is needed to be implemented? CMU TC HIVE group will implement
  - Infrastructure requirements - TCGA model or more expansive? Will we allow a structure for download?
  - Stories to drive initial queries, decision on what timeline and basis?
  - Maybe each DRT should have a top priority, and one more that cuts across the DRTs
  - Can we make geographic queries across the CCF? Cannot rely on this being available for all tissue types

12 organs