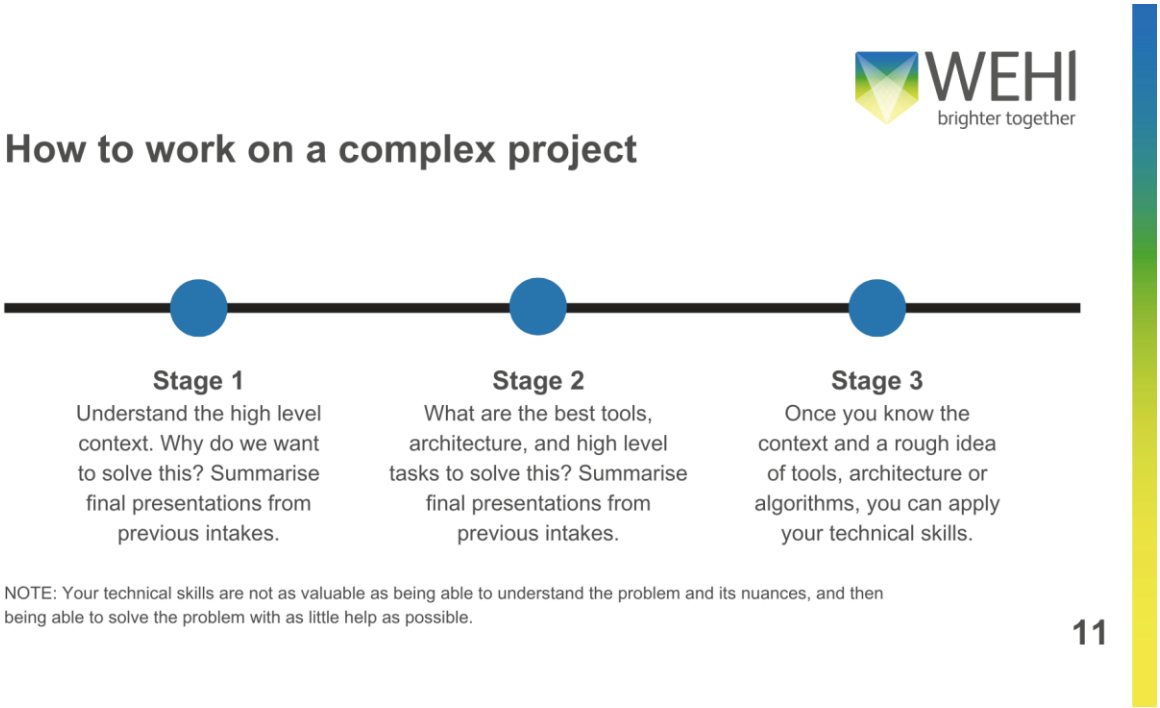
Student Projects Outline - Summer 24/25

Introduction

Please remember these are long-term goals and this internship is a best effort internship. This means that I am not expecting any of these goals to be completed. I am simply expecting that we're going to do the best that we can, collaboratively.

What I want from you is to be able to take these cryptic notes, watch the final presentations, and use them to understand what you might want to do as part of this internship (see below).



Key Links

Please bookmark these key links:

- the Welcome Slides for more general information,
- FAQs to help you troubleshoot problems,
- the Student Internship Handbook,
- Student Projects Outline to help explain high level context and tasks,

- the previous summary report from semester 2 2024,
- the Sharepoint from previous intakes that is tied to WEHI-wide student intern group,
- you can connect to your team mates, sister projects and skills register.

High level context to help with Stage 1

To find out more about the projects please look at the information below:

1. Introduction to REDMANE (slides by Rowland)*
    a. Create synthetic multi-omics data to match synthetic clinical and other metadata
    b. Extend functionality of the Data Registry application in ReactJS and FASTAPI
    c. Ingestion of data and metadata into the Data Registry using authentication
    d. Setup authentication for multiple Data Portals using OIDC, AAF, and KeyCloak
    e. Setup cBioPortal as a Data Portal on the Nectar Research Cloud
    f. Setup generic secure Shiny/R App as a Data Portal on the Nectar Research Cloud
    g. Setup Omero as a Data Portal on the Nectar Research Cloud
    h. Setup Storage Calculator as a Data Portal on the Nectar Research Cloud
    i. Identify other Data Portals for key data types
2. Clinical Dashboards
    a. Create synthetic multi-omics data to match synthetic clinical and other metadata
    b. Setup generic secure Shiny/R App as a Data Portal on the Nectar Research Cloud
3. Student Organiser
    a. Create a way to review resumes quickly
    b. How can we search through all help documents using RAG LLM?
    c. Create new views for "Allocation" and "Review resume" versus "interview"

* for more information on REDMANE please access the General REDMANE FAQ.

High level tasks to help build context for Stage 2

Please review these and ask lots of questions in weeks 1 and 2. We want you to understand why and make suggestions to improve these potential tasks. Asking questions is really important to ensure you understand why you are doing things!

Project/Subproject	**Bluesky Tasks (not feasible)**

REDMANE Demo and Quality team	• Setup a permanent demo
	  ○ Data registry with SSL
	  ○ Using synthetic datasets that are stored in data portals:
	  ▪ Clinical data (WEHI REDCap)
	  ▪ WGS data (cBioPortal)
	  ▪ Imaging (Omero)
	  ▪ Spatial Omics (?)
	  ○ Two organisations (VM) to “store” large raw synthetic data linked to the other data portals
	• Provide Quality Control and first level support to other groups, including how do you get your code into the demo environment.
	• Discuss and draft what is needed for a Data Portal to be accepted into the “App store”:
	  ○ Has to be dockerized
	  ○ OICD authentication
	  ○ Per dataset authorisation for users and groups
	  ○ Open Source
	  ○ Ability to link back to Data Registry
	  ○ Create empty templates for Django, Flask, Ruby on Rails etc that are ready to be part of the REDMANE ecosystem and can be built on top of easily

REDMANE Data Portals	• Setup Omero as a Data Portal • Create synthetic data for imaging or special omics that is tied to clinical data and put it into Omero • Document setup of the system (if needed) • Work with the Demo and Quality team to classify if Omero is ready to be a data portal • Setup authentication to be in sync with data registry and cBioPortal
REDMANE Web Dev	• Work on adding more functionality to the data registry based on the wireframes • Add in authentication and authorization into the data registry • Help the demo and quality team identify what is the definition of a data portal that can work within this ecosystem • Work with the data integration team to provide an API with authentication • Review the wireframes in the presentation and identify new pages if necessary • Set up unit tests as well as functional tests to ensure functionality is not lost

	• Work with the demo and quality team to provide pull requests so that new functionality can be added to the demo environment • Work with the data integration team to demo how and update from an organization who is just received new files would look like • Discuss with the PDF coding team how you would integrate the functionality into the data registry • UI to add in metadata at scale like Stemformatics and tie in ontologies to ensure metadata is consistent
REDMANE Workflows	• Set up one or two basic workflows using synthetic or public data such as whole genome sequencing or single cell rnaseq. • Tie the synthetic data to clinical data along with the clinical dashboards team. • Run this workflow on Milton hpc, possibly using next flow. • Work with the data registry team to see if they can kickstart a workflow on Milton using information within the data registry for things like input data and where should the output data should go • Work with the clinical dashboards team to create tutorials of less than 20 slides to explain to a new audience the nuances of the data set types
Clinical Dashboards Synthetic Data	• Create synthetic clinical data within red cap that ties in with data that might already be available in a data portal for example cBioportal • Create or find public data that could be used to mimic multi-omics data that is tired to the clinical data for example they would have the same sample IDs in the file name of the omics data • Work with the data ingestion team to ensure that they understand the nuances of the data sets • Simile work with the workflow team to ensure everyone understands the nuances of the data sets being created • Work with the workflow team to create tutorials of less than 20 slides to explain to a new audience the nuances of the data set types
REDMANE Data Ingestion	• Create scripts for different operating systems to ingest data into the data registry • These scripts should create intermediate files to help someone verify the quality and accuracy of the information that is going to be sent to the data registry. This would include a machine readable and a human readable version of the information such as ro crate

	• These scripts need to have some level of authentication to ensure that we keep the information private and secure • This team should be working with the clinical dashboards team and the workflows team to understand the nuances of the files within each data type
REDMANE Capacity Planning	• Storage Calculator to calculate future storage needs based on previous numbers • Work with the REDMANE Data Ingestion team to get numbers and possibly will need longitudinal (time series numbers) • Look at time series databases • Look at the previous Storage Calculator and use as a basis • Select two points on a time graph and calculate the extended linear regression of those two points • Calculate and update storage numbers to get better potential costs of storage in the future.
Student Organiser Data Viz	• I need to filter specific views of data across the application. I need a per intake view that shows either all students who applied or only students who finished. I would also like to see a list of students who worked on a project in past intakes or for a specific intake • In a similar vein I would like to calculate some of the numbers for each of these specific filtered students such as how many total hours would the student organizer from intake 6 have and what are the names of the students in that intake for that project • I would also like to visualize the numbers and other dashboard views in a more graphical format and some over time as well • I would also like to build more information into the allocation screen where I'm assigning students into projects so that I get a better understanding of the skill set within a team and the hours within a team • I would also like to improve on the synthetic data set so that it can be used to test all the scenarios that the web application can do
Student Organiser PDF coding	• I want to be able to streamline reviewing a resume and cover letter that is in a PDF form • I want to be able to highlight a sentence or a paragraph and add a comment. This is the functionality that you can get inside Google Drive when you have a PDF.

	• I want to extend this functionality so that not only can I highlight text create a comment but I also can add one or more tags predefined tags about that comment. For example I might highlight a paragraph and in my comment I would say this shows excellent communication skills and then I would add a tag to that comment called communication • I couldn Sport the highlighted text the comments amate and any text it at side of that application and stored in the student application as a summary • I could then access this summary of a person through the allocations page when I'm trying to assign students to projects
Student Organiser RAG LLM / Onboarding	• I want students to be able to easily search through all the documentation I have provided on the website and on documents I store in fig share. • I especially want them to be able to do a search using an LLM and return accurate data in the form of a link and associated information • I want to review how we setup the documentation so that it is easy to add documentation into the LLM and it is easy to link to specific questions if needed. • I also want to review how we onboard students to get them up to speed for the high-level context (stage 1) and the architectural and algorithmic limitations and suggestions (stage 2) so we can move the “whiteboard presentation” from week 4 to week 2 or 3.
Quantum Computing	• Reflect on how you like to learn, based on the idea of andragogy, the idea of self-directed learning as an adult • Look at the wiki and the roadmap and decide how you want to learn Quantum Computing and what you would like to get out of it • Create an individual plan for yourself and share with your team mates • Ensure you document your individual plan along the way • Update the wiki to share an updated version of your plan to help others in the future

Intake 10 Summary Report

RCP#0032 Intake 10 Student Internship Summary reports.docx