

Agenda for the week

	PulseOx Dataset
	Your own Project

Day	Goals	Materials
Monday	<ul style="list-style-type: none">- How to navigate MIMIC and BigQuery- Explore the Pulse Oximetry dataset	Pulling Data Workshop RMarkdown: <i>Introduction</i>
Tuesday	<ul style="list-style-type: none">- Build a study cohort for Pulse Oximetry- Explore the data distributions	RMarkdown: <i>Cohort</i> RMarkdown: <i>Tableone</i>
Wednesday	<ul style="list-style-type: none">- Dive into your own project!- Define the question, dataset, cohort	Build you own RMarkdown!
Thursday	<ul style="list-style-type: none">- Explore data distributions- Prepare the final presentation	Build you own RMarkdown!
Friday	<ul style="list-style-type: none">- Final Presentations	Slide Deck

Please address these questions in your presentation, 1 minute each

- **Problem** definition: Specifically, how would the model improve patient outcomes?
- What **database** are you going to use? Please identify potential issues with the database. How did patients get into the database? Who got excluded?
- **Inclusion** and **exclusion** criteria: Consider sampling selection bias that can seep into the study design as a result of these criteria.
- List of **features** and their definition: Consider bias with the use of these features and their definition.
- **Outcome** or event of interest: How is this defined? Consider bias in how the outcome or the event of interest is defined.
- **Evaluation** of in silico model performance How will the model be incorporated in practice? Any unintended consequence of the algorithm when employed.

pulling **data**,
not **teeth**

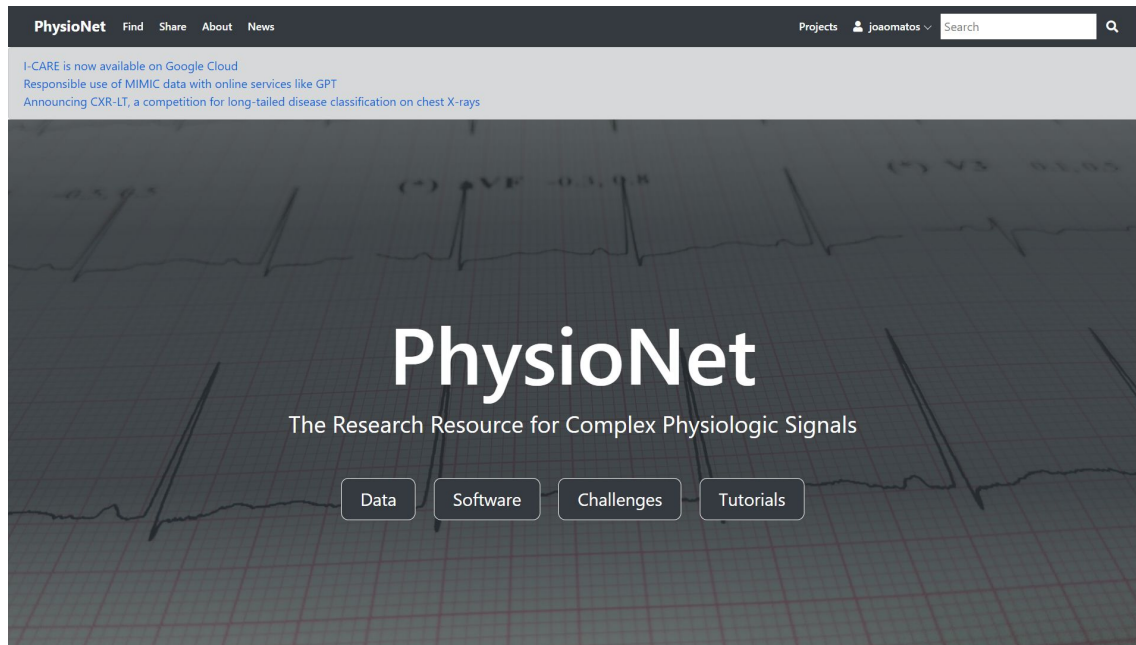
João Matos, Aug 14th 2023

How to navigate MIMIC and BigQuery – 101

Why a workshop on something upstream to Data Science?

- MIMIC is really well constructed
- Several groups across the world are following its schema / rationale
- You may come across different databases similar to MIMIC
- Data Science starts before a ***data <- read.csv("path")***
- You will need to navigate through complex databases to answer your research questions in the future
- Understanding the data is 95% of a ML task, and most biases are probably encoded in the Data Engineering step of the process

MIMIC-IV is shared on PhysioNet



<https://physionet.org/>

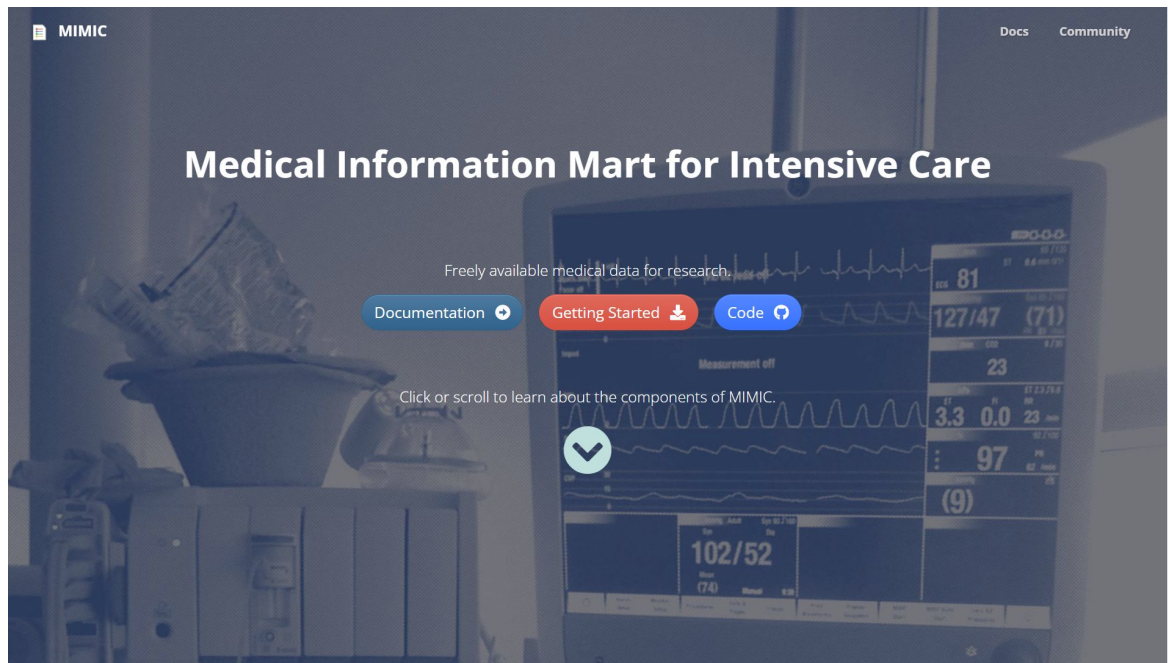
The databases in PhysioNet contain real patient data

- Data has been de-identified
- Users must be credentialed and have a reference
- CITI training must be taken
- A data use agreement must be signed for each dataset

Checkpoint:

has everyone completed
credentialing and the CITI training?

MIMIC-IV has a web page with detailed information



<https://mimic.mit.edu/>

How to get PhysioNet data on BigQuery

- A) Add a GMAIL email to PhysioNet in “Emails” and select it in “Cloud Details”
- B) “Request access using Google BigQuery” in the dataset page
- C) Go to your GMAIL inbox and follow instructions:
 - a) Navigate to: <https://console.cloud.google.com/bigquery>
 - b) Click the "+ADD DATA" button
 - c) Select "Star a project by name", then enter "physionet-data".

BigQuery is a convenient way of navigating MIMIC

The screenshot displays the Google Cloud BigQuery interface. At the top, the Google Cloud logo and 'My First Project' are visible. A search bar is present with the text 'Search (/) for resources, docs, pro...'. Below this, the left sidebar shows a list of workspace resources. The 'physionet-data' folder is expanded, revealing several datasets, with 'mimiciv_icu' selected and highlighted. The main panel on the right displays the 'Data set info' for 'mimiciv_icu'. This panel includes a table with various metadata fields and their values, such as 'Data set ID', 'Created', 'Default table expiry', 'Last modified', 'Data location', 'Description', 'Default collation', 'Default rounding mode', 'Case insensitive', 'Labels', and 'Tags'. At the bottom of the main panel, there are buttons for 'PERSONAL HISTORY' and 'PROJECT HISTORY', along with a 'REFRESH' button.

Google Cloud My First Project Search (/) for resources, docs, pro... Search

Viewing workspace resources. SHOW STARRED ONLY

- protean-chassis-368116
- physionet-data
 - External connections
 - eicu_crd
 - eicu_crd_demo
 - eicu_crd_derived
 - mimiciii_clinical
 - mimiciii_demo
 - mimiciii_derived
 - mimiciii_notes
 - mimiciv_derived
 - mimiciv_hosp
 - mimiciv_icu**
 - project-racial-bias-abgs

Data set info

EDIT DETAILS

Data set ID	physionet-data.mimiciv_icu
Created	24 Jun 2022, 22:25:56 UTC-4
Default table expiry	Never
Last modified	16 Mar 2023, 18:39:18 UTC-4
Data location	US
Description	
Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Case insensitive	false
Labels	
Tags	

PERSONAL HISTORY PROJECT HISTORY REFRESH

Checkpoint:
has everyone set BigQuery up?

The MIMIC Code Repository has open-source code to crowdsource the process of curating the data

JOURNAL ARTICLE

The MIMIC Code Repository: enabling reproducibility in critical care research

Alistair E W Johnson ✉, David J Stone, Leo A Celi, Tom J Pollard

Journal of the American Medical Informatics Association, Volume 25, Issue 1, January 2018, Pages 32–39, <https://doi.org/10.1093/jamia/ocx084>

Published: 27 September 2017 **Article history** ▼

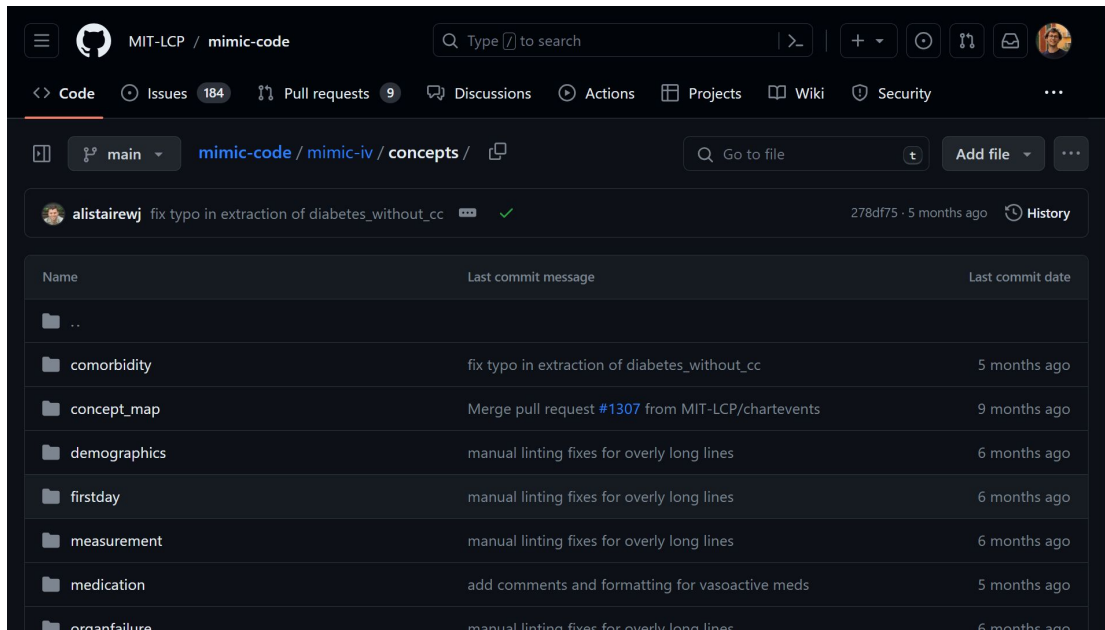
<https://academic.oup.com/jamia/article/25/1/32/4259424>

No one starts from scratch as there are derived tables

Findable on BigQuery

▼ mimivc_derived	☆	⋮
age	☆	⋮
antibiotic	☆	⋮
apsiii	☆	⋮
bg	☆	⋮
blood_differential	☆	⋮
cardiac_marker	☆	⋮
charlson	☆	⋮
chemistry	☆	⋮
coagulation	☆	⋮
complete_blood_count	☆	⋮
creatinine_baseline	☆	⋮
crrt	☆	⋮

With the SQL code that generated them on GitHub

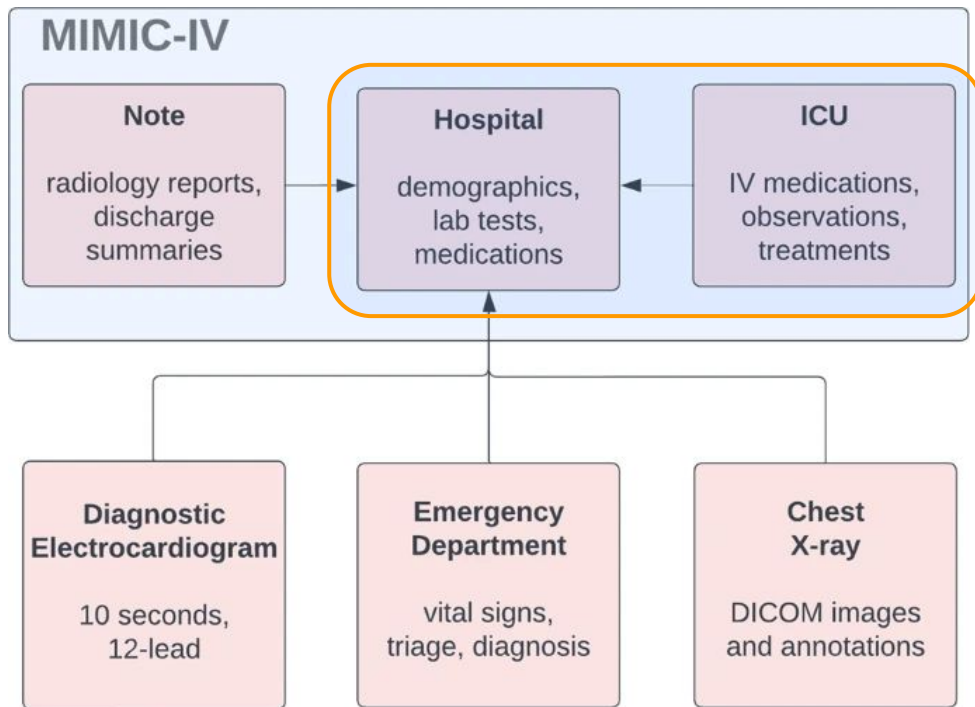


The screenshot shows the GitHub repository MIT-LCP / mimic-code. The file structure is as follows:

Name	Last commit message	Last commit date
..		
comorbidity	fix typo in extraction of diabetes_without_cc	5 months ago
concept_map	Merge pull request #1307 from MIT-LCP/chartevents	9 months ago
demographics	manual linting fixes for overly long lines	6 months ago
firstday	manual linting fixes for overly long lines	6 months ago
measurement	manual linting fixes for overly long lines	6 months ago
medication	add comments and formatting for vasoactive meds	5 months ago
organfailure	manual linting fixes for overly long lines	6 months ago

<https://github.com/MIT-LCP/mimic-code/tree/main/mimic-iv/concepts>

MIMIC-IV has several modules, we'll be looking at two



<https://www.nature.com/articles/s41597-022-01899-x>

Live Coding:

let's extract ICU stays and patient
information on race and gender

Possible solution

SELECT

```
    icu.subject_id  
, icu.hadm_id  
, icu.intime  
, icu.outtime  
, adm.race  
, pat.gender
```

FROM `physionet-data.mimiciv_icu.icustays` AS icu

LEFT JOIN `physionet-data.mimiciv_hosp.admissions` AS adm
ON icu.hadm_id = adm.hadm_id

LEFT JOIN `physionet-data.mimiciv_hosp.patients` AS pat
ON icu.subject_id = pat.subject_id



1 – 50 of 73181



Live Coding:

let's extract patients who are, e.g,
Female and Native American

Possible solution

```
SELECT *  
FROM `physionet-data.mimiciv_derived.icustay_detail`  
WHERE gender = "F"  
AND LOWER(race) LIKE "%native%"
```

Results per page: 50 ▼ 1 – 50 of 111

Live Coding:

let's extract transfusion data

Possible solution

```
SELECT label, itemid
FROM `physionet-data.mimiciv_icu.d_items`
WHERE LOWER(label) LIKE "%red blood cell%"
```

Query results

JOB INFORMATION		RESULTS	JSON	EXEC
Row	label	itemid		
1	Packed Red Blood Cells	225168		

```
SELECT
    stay_id
    , SUM(amount) AS total_amount
FROM `physionet-data.mimiciv_icu.inpatevents`
WHERE itemid = 225168
GROUP BY stay_id
```

Query results








JOB INFORMATION		RESULTS	JS
Row	stay_id	total_amount	
1	38337741	1000.000026...	
2	39791680	1725.0	
3	31946853	374.9999866...	
4	33410078	2475.000010...	
5	37771327	6524.000065...	
6	32944882	425.0	

This week, we will be working with a derived dataset

Challenge

Credentialed Access

MIT Critical Datathon 2023: a MIMIC-IV Derived Dataset for Pulse Oximetry Correction Models

João Matos , Tristan Struja , David S Restrepo , Luis Filipe Nakayama , Jack Gallifant , Luca Weishaupt , Nikita Mullangi , Maria Loureiro , Skyler Shapiro , Adrien Carrel , Leo Anthony Celi 

Published: May 8, 2023. Version: 1.0.0

When using this resource, please cite: [\(show more options\)](#)

Matos, J., Struja, T., Restrepo, D. S., Nakayama, L. F., Gallifant, J., Weishaupt, L., Mullangi, N., Loureiro, M., Shapiro, S., Carrel, A., & Celi, L. A. (2023). MIT Critical Datathon 2023: a MIMIC-IV Derived Dataset for Pulse Oximetry Correction Models (version 1.0.0). *PhysioNet*.
<https://doi.org/10.13026/jfpc-pz79>.

Please include the standard citation for PhysioNet: [\(show more options\)](#)

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* [Online]. 101 (23), pp. e215–e220.

Contents ▾

Parent Projects

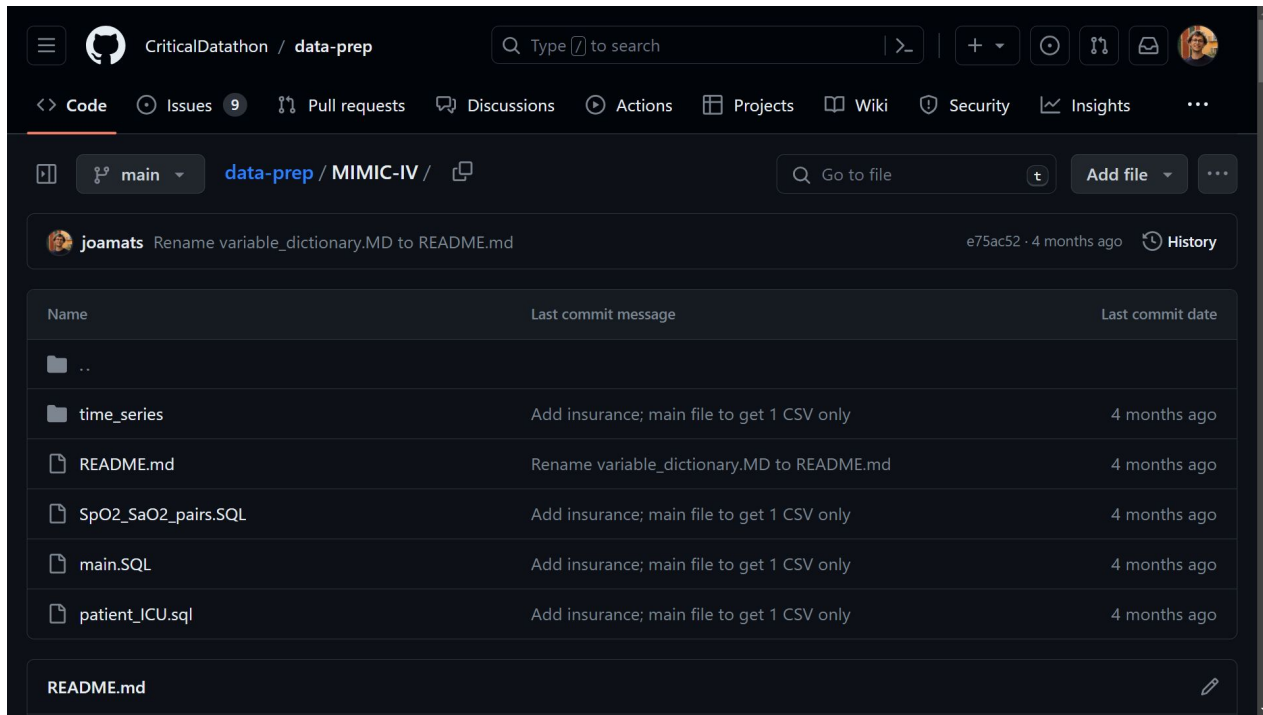
MIT Critical Datathon 2023: a MIMIC-IV Derived Dataset for Pulse Oximetry Correction Models was derived from:

- [MIMIC-IV v2.2](#)

Please cite them when using this project.

<https://physionet.org/content/mit-critical-datathon-2023/1.0.0/>

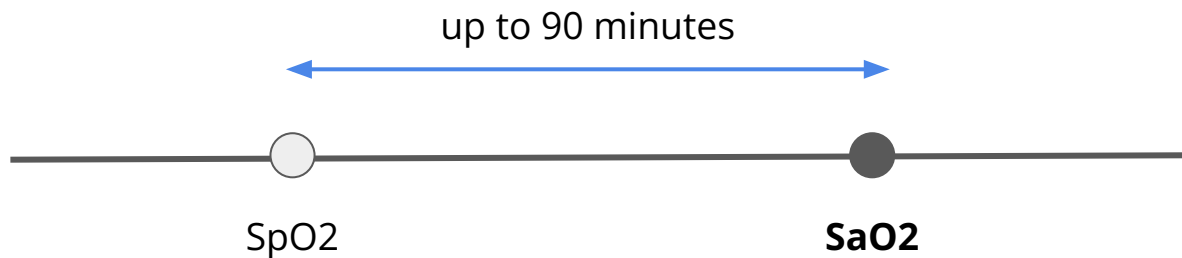
All the code behind is open-source on GitHub



<https://github.com/CriticalDatathon/data-prep>

The dataset contains pairs of SpO₂ and SaO₂ values

- 81,797 SaO₂ – SpO₂ pairs from the ICU, derived from MIMIC-IV
- Aligns the pairs with patient information, closest vital signs, laboratory values, SOFA scores, and treatment information



Explore!
a preview of the dataset

<https://physionet.org/content/mit-critical-datathon-2023/1.0.0/>

Download the dataset from PhysioNet

Files









Total uncompressed size: 50.1 MB.

Access the files

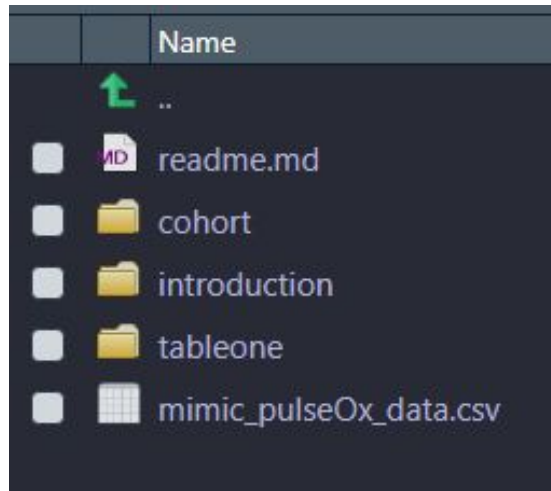
- [Download the ZIP file](#) (12.8 MB)
- [Request access](#) using Google BigQuery.
- Download the files using your terminal:

```
wget -r -N -c -np --user joaomatos --ask-password https://physionet.org/files/mit-critical-datathon-2023/1.0.0/
```

Folder Navigation: <base>

Name		Size	Modified
 LICENSE.txt		2.5 KB	2023-05-07
 SHA256SUMS.txt		259 B	2023-05-08
 mimic_pulseOx_data.csv		50.0 MB	2023-05-04
 mimic_pulseOx_dictionary.csv		9.1 KB	2023-05-04

Make sure to place the CSV in your *datathon* folder



Explore!
the RMarkdown called “introduction”

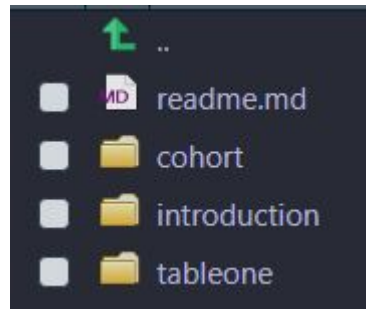
The goal for today is to set up and explore the data

Workshop0

2023-07-27

Workshop **0** Preparatory Materials

Please go through these materials before attending further Workshops.



What is a Datathon?

A datathon is a collaborative event designed to bring clinicians and data scientists together in the development of data-driven models. This process involves the use of de-identified datasets from different sources, such as electronic health records. The main objective is to analyze these datasets using both data science and medical knowledge.

A datathon combines medical science and data science to solve real-world problems with existing datasets. It encourages participants from diverse