1. Sign up for comprehensive exam Oral Exam 20 mins <https://www.coursera.org/learn/siads591592/irt/c2XQ4/read-exam-guidelines-sign-up-for-an-exam-slot-review-exam-rubric/step-1/submit/>
2. Sign up team for project
3. Set up cadence on checking in with instruction teams
4. Talk about Power BI for presentation – which visuals to elevate (and why some visuals can stay in the notebook)
5. Discuss South Kora dataset
6. Preview Peer proposals: <https://docs.google.com/spreadsheets/d/1Vp-c13FWAZVXaDXqAdYkYrlFy8HCmuMeNzWj6skI5XU/edit#gid=1064547803>
7. Link to our team’s document: https://docs.google.com/document/d/11EjOYxspkVUcdoIH0aKk9cCP1iNXOpleZbi64PTHEZk/edit#

Your reviews should be several sentences long and should take into consideration the following points:

* **professional:** what would a co-worker think about your review?
* **pleasant:** courtesy goes a long way
* **helpful:** what sort of advice would you want?
* **scientific:** focus on facts, not opinions
* **realistic**: keep scope in mind
* **empathetic**: how would you feel if you received the review you wrote?

1. **organized:** make it easy for the recipient to follow your train of thought

#####################################################################################

**Submission my Aug 2nd**

**By Monday August 2, 2021:** Please review the project guidelines, build your team, and describe your proposed project via the document links below. These Documents are also located in Resources.

Read: [Project guidelines](https://docs.google.com/document/d/1tEyXZcpVUJhy2Zsakxhuyp5BoE-VWqz4RPKhV4b9Gnc/edit?usp=sharing)

Sign up: [Project proposal and team tracker](https://docs.google.com/spreadsheets/d/1Vp-c13FWAZVXaDXqAdYkYrlFy8HCmuMeNzWj6skI5XU/edit?usp=sharing)

Copy and complete: [Project proposal template](https://docs.google.com/document/d/1i_1DE_u3cjCqJ4nYWHNmdY5S1qDDrmZsdLA1ai7NH08/edit?usp=sharing)

**NOTE:** You will need to be logged into your UMich Google Account in order to access these documents.

Once you have done all 3 items listed, click submit.

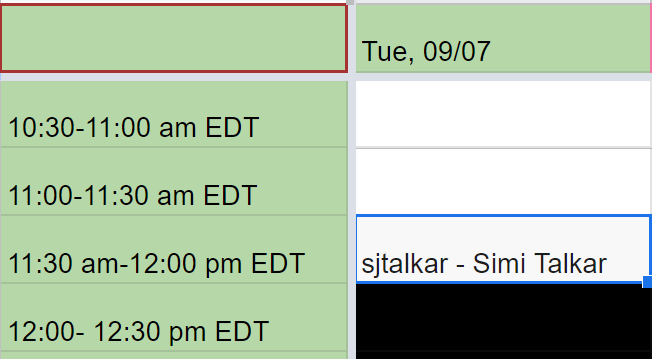
#####################################################################################

[Comprehensive Exam Oral Guideline](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit)

[Comprehensive exam Rubric](https://docs.google.com/document/d/1Kt4kQogM8qHmKPmVon2ljTMc4GWxjY6WCmqgH1E-eng/edit)

Went above and beyond to research information; solicited material in addition to what was provided; brought in personal ideas and information to enhance oral exam

Comprehensive Exam Sign up



[1.0 Introduction](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.5gz3twh312cy)

[2.0 Exam sign up](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.ropbf7hlh87f)

[3.0 Scenario (dataset)](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.85dbmqiqbk8n)

[4.0 Topics](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.qlb9byue57hs)

[5.0 Grading Rubric](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.d7e1w4357a7b)

[6.0 Suggestions for Preparing for the Exam](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.xdedcqicenqc)

[Changelog](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.hg9yk5wxcr1)

#####################################################################################

[Project Guidelines](https://docs.google.com/document/d/1tEyXZcpVUJhy2Zsakxhuyp5BoE-VWqz4RPKhV4b9Gnc/edit)

Link to Project Document to sign up as a team by Aug 2nd

<https://docs.google.com/spreadsheets/d/1Vp-c13FWAZVXaDXqAdYkYrlFy8HCmuMeNzWj6skI5XU/edit#gid=1064547803>

#####################################################################################

[TBD] One of Five Topics - which are they?

Focus on one.

Provided with dataset on South Korea – where is it?

Concrete example for one of the five topics.

[TBD] : Creation of a Power BI/ PDF for final presentation

Avoid Code (Mention and Link to the Code – File name and Function )

SOUTH KOREA DATASET:

<https://towardsdatascience.com/coronavirus-a-big-data-lesson-from-south-korea-5bb703b8b0ae>

<https://www.kaggle.com/kimjihoo/coronavirusdataset>

https://www.kensci.com/blog/covid-19-and-a-data-driven-story

PHASE 1  
1) Exploratory Data Analysis  
2) Surface Insights

[TBD] : Project title and finish up proposal

**COMPREHENSIVE ORAL EXAM**

1. **Context : Kaggle dataset (Covid and South Korea)**
2. **Concrete examples**
3. **Of the five topics we will discuss one**
4. **Three phases: Spend 5 mins into introducing the topic**
5. **Follow up questions**
6. **Final thoughts – Summary – further aspects**
7. **Screen cast – visualizations – not necessary.**

**COMPREHENSIVE EXAM GUIDELINES**

SIADS 591/592

Milestone I Comprehensive

Oral Exam Guidelines

version 2021.07.24.1.AW

[1.0 Introduction](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.5gz3twh312cy)

[2.0 Exam sign up](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.ropbf7hlh87f)

[3.0 Scenario (dataset)](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.85dbmqiqbk8n)

[4.0 Topics](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.qlb9byue57hs)

[5.0 Grading Rubric](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.d7e1w4357a7b)

[6.0 Suggestions for Preparing for the Exam](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.xdedcqicenqc)

[Changelog](https://docs.google.com/document/d/1J22R1sG6LEmgyqLWlNDZA3w534u2qapGP9wnSh0oU1g/edit#heading=h.hg9yk5wxcr1)

## 1.0 Introduction

The purpose of this milestone course is to provide you with an opportunity to demonstrate your mastery of the material from the courses you've taken so far in the MADS program. One instrument that we will use to assess your knowledge is a comprehensive exam.  Comprehensive exams can take many forms: a timed written exam or a collection of papers that you write are commonly used formats.  In this course the comprehensive exam will take the form of a structured 20-minute live conversation (via Zoom) with the course instructor.  Alternative arrangements can be made if this proves impossible for technical reasons, but in all cases a "live" conversation is required.  Here's how it will work:

You will be given a list of topics as well as one "scenario" (i.e. dataset).  We suggest that you prepare responses for each of the topics as applied to the scenario.  At the commencement of the comprehensive oral examination you will be told which topic (selected at random) will form the basis of the conversation.

The 20-minute session will consist of a 4-5 minute presentation of your answer followed by approximately 10-12 minutes of in-depth discussion of the topic.  You will then have approximately 4-5 minutes to provide closing remarks or reflections.

## 2.0 Exam sign up

You must select your exam date and time via the Comprehensive Exam Sign Up Google sheet (see Coursera week 1 readings for the link to the document).

“Sign up” tab: select your date/time slot, add your full name and uniqname to the white cell.

* Do not add your name to any gray cells.
* Select **only** **one** date/time slot.
* You are encouraged to use the siads591\_[semester]\_001\_comp\_exam Slack channel to trade date/time slots with your peers.
* If there are no date/time slots that work for you, please reach out to Emily Sartorius <[sartorie@umich.edu](mailto:sartorie@umich.edu)> via Slack or email as soon as possible who will work with you to find a date and time that works for you and the course instructor.
* Ahead of your exam date and time, you will receive a Google calendar meeting invitation with the Zoom conference link.
* If you are suddenly unable to make your exam date and time, reach out to Emily Sartorius <[sartorie@umich.edu](mailto:sartorie@umich.edu)> with a **minimum** of 24 hours notice.

## 3.0 Scenario (dataset)

Data Science for COVID-19 (DS4C)

DS4C: Data Science for COVID-19 in South Korea

<https://www.kaggle.com/kimjihoo/coronavirusdataset>

This dataset covers the initial impact of the coronavirus in South Korea. You may find the [detailed description](https://www.kaggle.com/kimjihoo/ds4c-what-is-this-dataset-detailed-description) particularly helpful.

## 4.0 Topics

These are the topics for which you should prepare answers:

1. Imagine a client who has asked you to do some analysis of the South Korean COVID-19 dataset. Who is the client and what have they asked you to do? Describe one maxim, question, or ethical commitment from the problem formulation phase of a data science project and apply it to this project. What would you do differently as a data scientist given your chosen maxim, question, or ethical commitment? Be prepared to illustrate your approach with examples drawn from the data.

1. Discuss the role of data visualization in data exploration and analysis. Be prepared to describe one basic and one advanced visual exploration technique that could be applied to the South Korean COVID-19 dataset. Explain how each visualization is encoded, how it is to be interpreted, and evaluate each in terms of its expressiveness and effectiveness if it was to be shared with a general audience. You should plan to experiment with different visualizations as part of your preparation for this question.

1. We are all familiar with the maxim “correlation does not imply causation”. But what does it mean when we assert that two or more variables are statistically related? Describe the concept as well explain how to calculate correlation coefficients. Utilize the South Korean COVID-19 dataset to illustrate your points. Do associations exist (either positive or negative) between variables derived from the data? Is there a causal story to be told about the transmission, incidence, or persistence of the coronavirus in South Korea? (Chis Brooks and Resnick)

1. The split-apply-combine data analysis pattern, identified by Hadley Wickham in a [2011 paper published in the Journal of Statistical Software](https://www.jstatsoft.org/article/view/v040i01), describes a strategy for analyzing data. Describe the pattern and how it has been implemented in the Pandas library. Discuss how the split-apply-combine pattern can be utilized to glean insights from the South Korean COVID-19 dataset.  Provide at least two examples of its use from your exploratory analysis of the data.

1. It's often said that data cleaning and manipulation takes up 80% of a data scientist's time, although that claim has recently been [disputed](https://blog.ldodds.com/2020/01/31/do-data-scientists-spend-80-of-their-time-cleaning-data-turns-out-no/).  No matter what the actual number is, describe the tools and/or techniques you utilize to make the data cleaning and manipulation effort more manageable, efficient, and scalable.  Use explicit examples from your analysis of the South Korean COVID-19 dataset.

## 5.0 Grading Rubric

The comprehensive exam will be adjudicated according to a standard rubric (see Coursera week 1 readings for the link to the document).

## 6.0 Suggestions for Preparing for the Exam

We suggest you work together to prepare for your exam.  Feel free to leverage Slack, email, and any other communication means to support your work.  The course instructors will be available during regularly scheduled office hours to help you work through your ideas.

**WEBINAR:  
Covid 19 dataset in South Korea**

[**https://pages.databricks.com/202009-APJ-WB-DS4C-COVID19-SouthKorea-Tech-Talk\_01.OnDemand-Thank-You-webinar-OD.html**](https://pages.databricks.com/202009-APJ-WB-DS4C-COVID19-SouthKorea-Tech-Talk_01.OnDemand-Thank-You-webinar-OD.html)

**Extra data :** <https://kosis.kr/eng/>

**Detailed discussion :** [**https://www.kaggle.com/kimjihoo/ds4c-what-is-this-dataset-detailed-description**](https://www.kaggle.com/kimjihoo/ds4c-what-is-this-dataset-detailed-description)

**The data includes :**

1. **Infected people travel routes**
2. **Public transport they took**
3. **Medical institutions that are treating them**
4. **High degree of details**

**Age, sex, city,   
Infection\_case – avenue of infection (overseas inflow, contact with patient)  
Infected by (Id of who infected the patient)  
(Contact Number) The number of contacts with patient)**

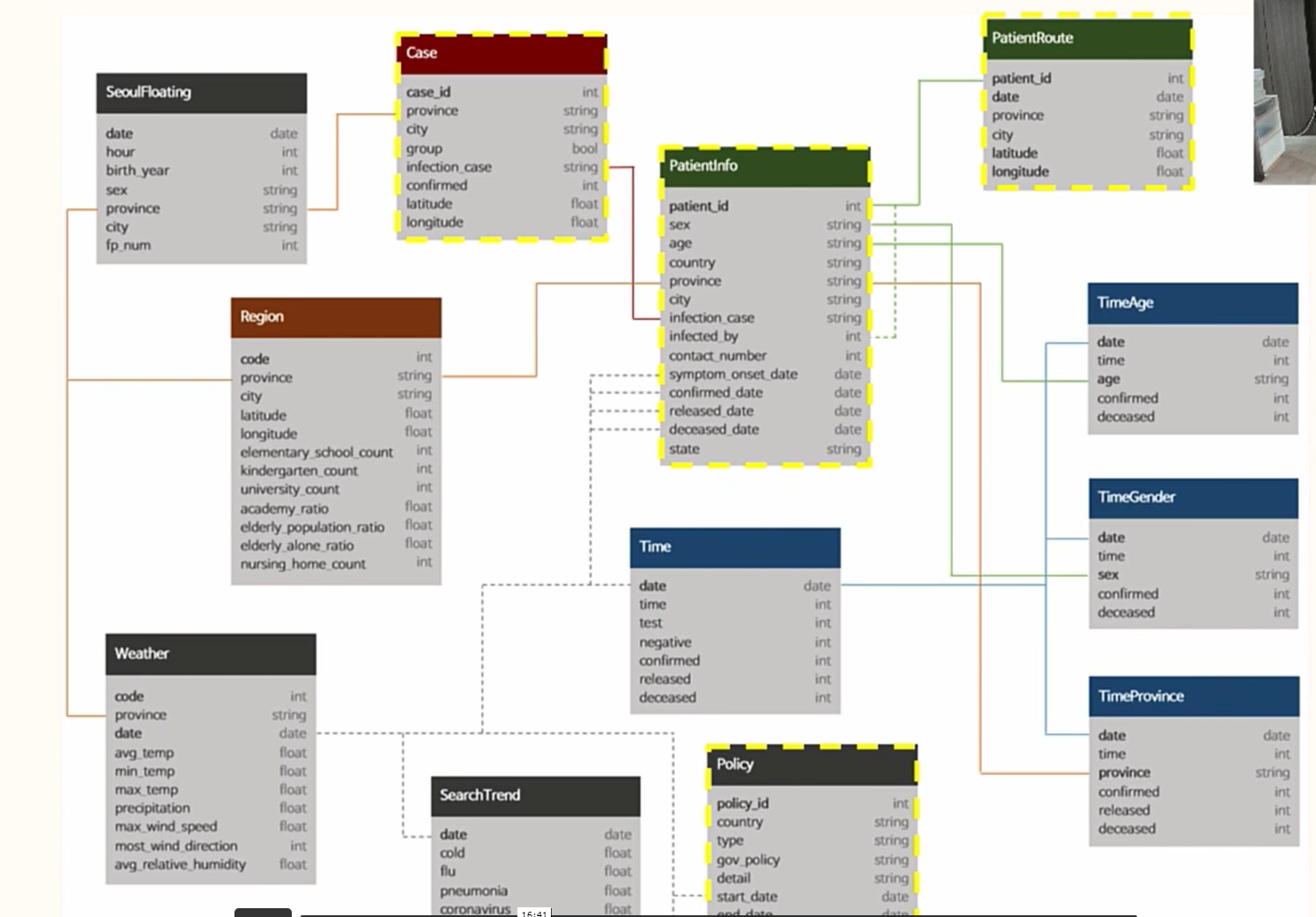
**Transmission network.**

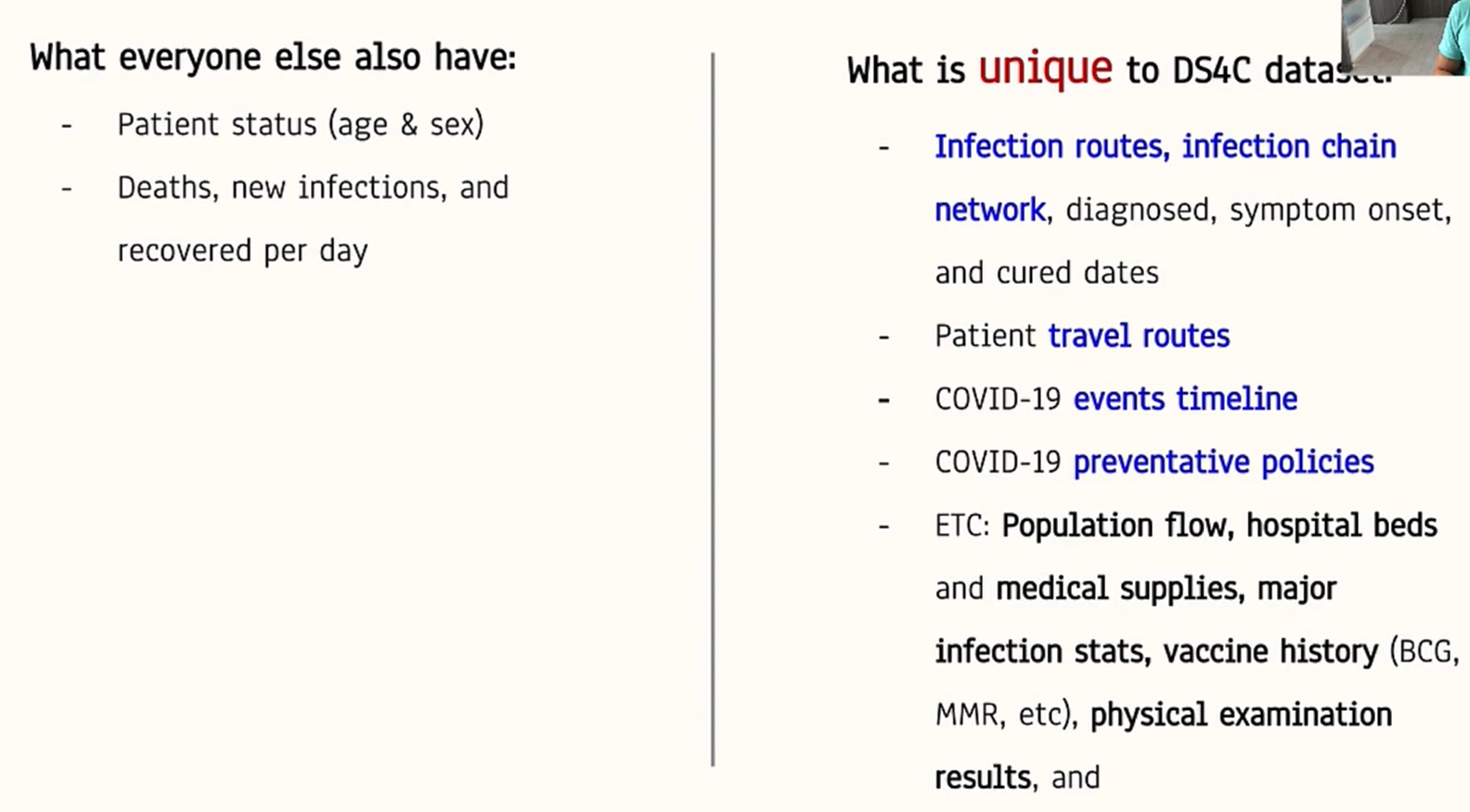
****

**MASS OUTBREAK CASES DATA**

**Map data with latitude and longitude**

**Infection\_case**





**https://www.kaggle.com/kimjihoo/ds4c-korea-wonderland-fight-against-covid-19**

**Data problems we do not have: Decentralized publication, Absence of unified formatting, Data embedded in natural language**

<https://vega.github.io/vega-datasets/data/world-110m.json>

Mapping with Altair UW

<https://github.com/uwdata/visualization-curriculum/blob/master/altair_cartographic.ipynb>

## world-110m.json

## countries.json

## https://www.districtdatalabs.com/altair-choropleth-viz

Geo json files of South Korea

<https://github.com/southkorea/southkorea-maps/tree/master/gadm/json>

https://stackoverflow.com/questions/64234865/world-map-slider-in-altair-python

[**https://colab.research.google.com/github/altair-viz/altair-tutorial/blob/master/notebooks/09-Geographic-plots.ipynb**](https://colab.research.google.com/github/altair-viz/altair-tutorial/blob/master/notebooks/09-Geographic-plots.ipynb)

[**https://colab.research.google.com/github/uwdata/visualization-curriculum/blob/master/altair\_cartographic.ipynb**](https://colab.research.google.com/github/uwdata/visualization-curriculum/blob/master/altair_cartographic.ipynb)

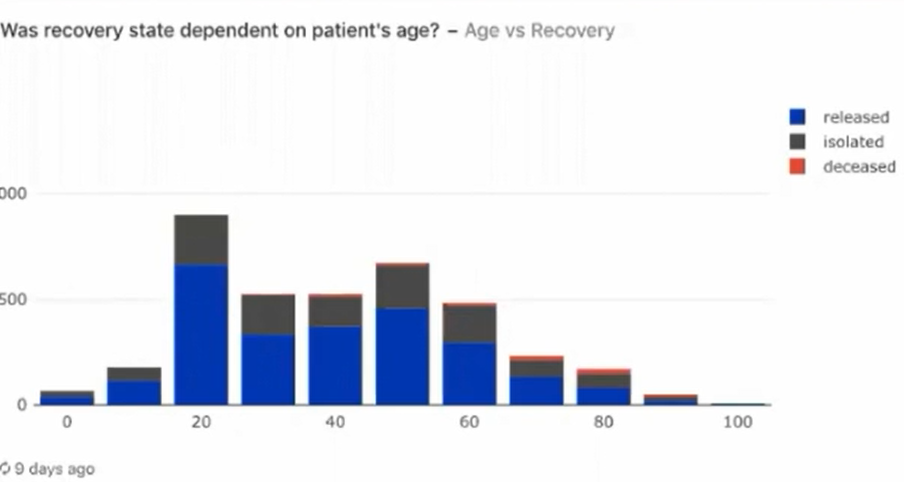
**QUESTIONS: Study difference based on vaccination rates**

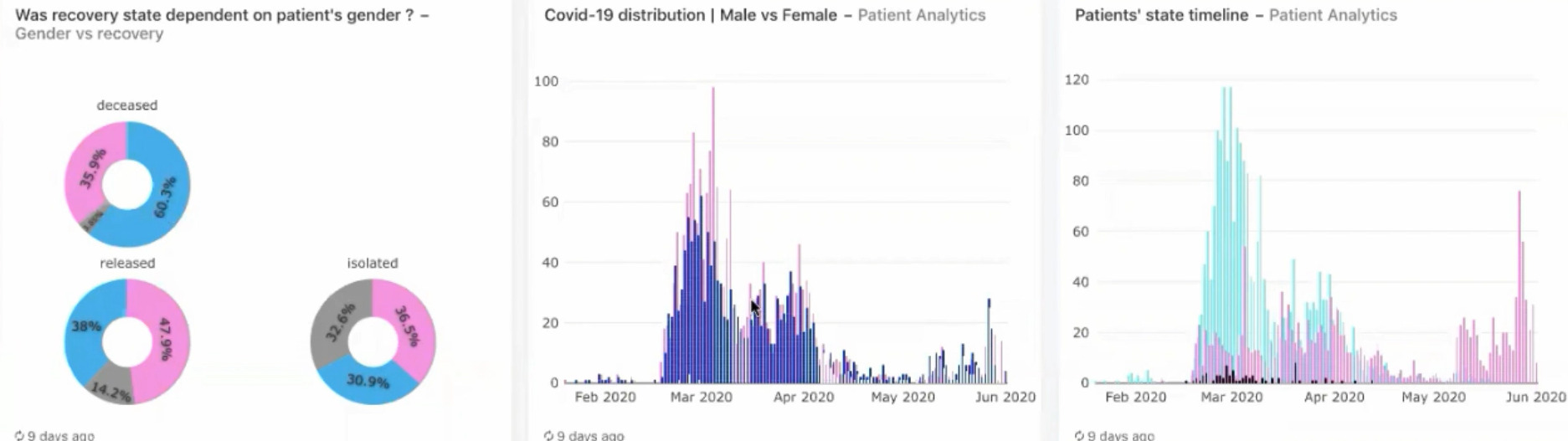
**When did the first case originate**

**Segment by Govt Policy**

**Waning and webing of cases**

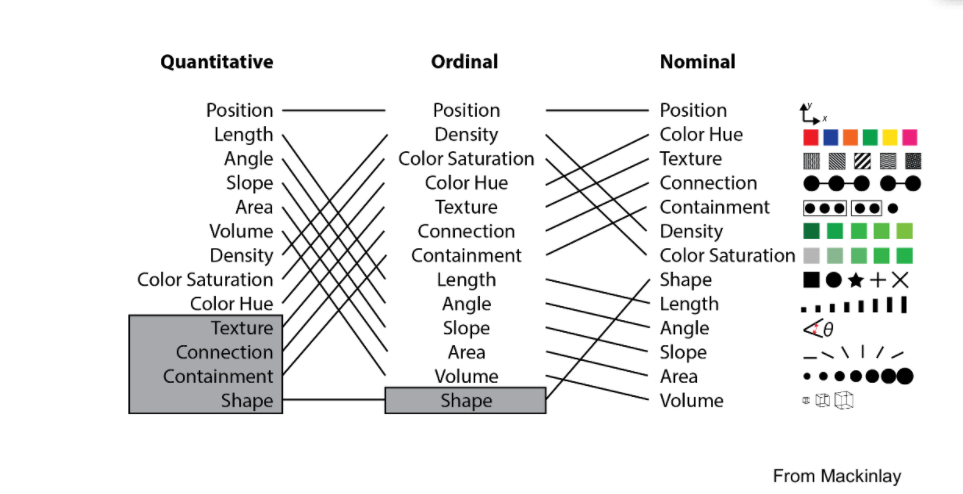
**Charts:**





[**https://www.kaggle.com/subinium/how-to-use-folium-geospatial-data**](https://www.kaggle.com/subinium/how-to-use-folium-geospatial-data)

**https://github.com/sjtalkar/SeriallyBuildDashboard/blob/main/callbacks\_spatial.py**

**Anscombe’s Quartet**

In the end, many hundreds of predictive tools were developed. None of them made a real difference, and some were potentially harmful. She and her colleagues have looked at 232 algorithms for diagnosing patients or predicting how sick those with the disease might get. They found that none of them were fit for clinical use. Just two have been singled out as being promising enough for future testing.

Both teams found that researchers repeated the same basic errors in the way they trained or tested their tools. Incorrect assumptions about the data often meant that the trained models did not work as claimed.

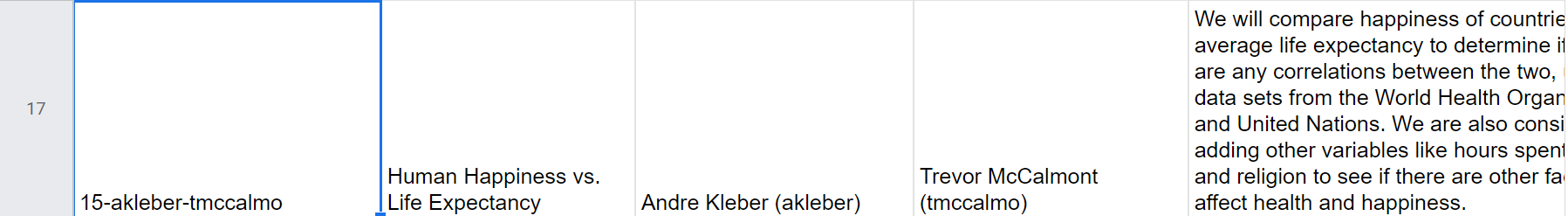
But this meant that many tools were built using mislabeled data or data from unknown sources.

Driggs highlights the problem of what he calls Frankenstein data sets, which are spliced together from multiple sources and can contain duplicates. This means that some tools end up being tested on the same data they were trained on, making them appear more accurate than they are.

They can also be fixed by adjusting the models, if researchers are aware of them. It is possible to acknowledge the shortcomings and release a less accurate, but less misleading model. But many tools were developed either by AI researchers who lacked the medical expertise to spot flaws in the data or by medical researchers who lacked the mathematical skills to compensate for those flaws.

The simplest move would be for AI teams to collaborate more with clinicians, says Driggs. Researchers also need to share their models and disclose how they were trained so that others can test them and build on them. “Those are two things we could do today,” he says. “And they would solve maybe 50% of the issues that we identified.”

**April 16th**





**Review 1**

**https://docs.google.com/document/d/1Ebo0R514593twL--Y-v-YgU5W\_7UQAy3NpBWtmCSOfc/edit**

**Review 2**

[**https://docs.google.com/document/d/136wzkKDK72QB8vNWw3ERFBj99jtb4Z5W4SHp-YgyCJc/edit#heading=h.awft0jfnr3so**](https://docs.google.com/document/d/136wzkKDK72QB8vNWw3ERFBj99jtb4Z5W4SHp-YgyCJc/edit#heading=h.awft0jfnr3so)

**Comprehensive exam**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **7Tue** | Milestone 1 Comprehensive Exam  [View on Google Calendar](https://calendar.google.com/calendar/r?eid=Z2d1YWNsZHNxbDFua2xlazBqb2hsbnVlZjAgc2p0YWxrYXJAdW1pY2guZWR1&ctok=c2p0YWxrYXJAdW1pY2guZWR1)   |  |  | | --- | --- | | When | Tue Sep 7, 2021 8:30am – 9am (PDT) | | Where | https://umich.zoom.us/j/6584136217 | | Who | wusean@umich.edu, agiove@umich.edu, Milestone 1 Exam Calendar\* |   Yes  Maybe  No  More options | Agenda  Tue Sep 7, 2021   |  |  | | --- | --- | | *No earlier events* | | | 8:30am | Milestone 1 Comprehensive Exam | | *No later events* | | |
| **You have been invited to the following event.**Milestone 1 Comprehensive Exam  |  |  | | --- | --- | | When | Tue Sep 7, 2021 8:30am – 9am Pacific Time - Los Angeles | | Where | <https://umich.zoom.us/j/6584136217> ([map](https://www.google.com/url?q=https%3A%2F%2Fumich.zoom.us%2Fj%2F6584136217&sa=D&ust=1628978421132000&usg=AOvVaw1Q73RyGfhhXnOBwE8ay3eA)) | | Calendar | [sjtalkar@umich.edu](mailto:sjtalkar@umich.edu) | | Who | |  |  | | --- | --- | | • | [sartorie@umich.edu](mailto:sartorie@umich.edu) - creator | | • | [wusean@umich.edu](mailto:wusean@umich.edu) | | • | [agiove@umich.edu](mailto:agiove@umich.edu) | | • | [sjtalkar@umich.edu](mailto:sjtalkar@umich.edu) | |   [**more details »**](https://calendar.google.com/calendar/event?action=VIEW&eid=Z2d1YWNsZHNxbDFua2xlazBqb2hsbnVlZjAgc2p0YWxrYXJAdW1pY2guZWR1&tok=NTQjY18xcWsxcGI5b21wM245aWpkMDRuNWFyb3Zvb0Bncm91cC5jYWxlbmRhci5nb29nbGUuY29tMzE3OTY5YTFkNjYwNTE3ZTRkY2U3ZTlkNTRiMmM0YzE1MjE5MTQ5Ng&ctz=America%2FLos_Angeles&hl=en&es=1)  Emily Joan Sartorius is inviting you to a scheduled Zoom meeting.  Join Zoom Meeting [https://umich.zoom.us/j/6584136217](https://www.google.com/url?q=https%3A%2F%2Fumich.zoom.us%2Fj%2F6584136217&sa=D&ust=1628978421129000&usg=AOvVaw1bFqjh_eZt2v8cVXzO9j-8)  Meeting ID: 658 413 6217 One tap mobile +16468769923,,6584136217# US (New York) +13017158592,,6584136217# US (Washington DC)  Dial by your location +1 646 876 9923 US (New York) +1 301 715 8592 US (Washington DC) | | |

**From other projects:**

The *hotels.csv* dataset is retrieved by using Tin Duong’s [Booking Scrapper](https://apify.com/dtrungtin/booking-scraper) API which is available on apify.com. This API is free to use and allows one to scrape hotel metadata from Booking.com.