Supplementary Appendix A

LEADS Leadership and Excellence in Analytics and Data Science

LEADS Data & Visualization Assignment

2017-2018 Cohort

TECHNOLOGY INNOVATION CENTER   JOHNS HOPKINS MEDICINE

Introduction

For the 2017-2018 LEADS year there is a single assignment – to be completed in groups.

Each team is being asked to build a data product performed from a social network analysis of the LEADS faculty members of their professional networks inside the Data Trust.   The best data product created will be shared with the students in LEADS as a personal guide to help build their profession network within the data trust.

This data product project will have the following steps.
1. Develop an information model from the flat file to be able to answer a series of questions
2. Work on cleaning the data and building the model in a SQL Server environment
3. Explore data visualization with Tableau
4. Creation of the data product

TECHNOLOGY INNOVATION CENTER   JOHNS HOPKINS MEDICINE

Exercise 1: Team Building

Meet and familiarize yourself with your group members, figure out who has what skills, and how you're going to teach each other from your strengths. Open the data set, and consider how you can turn it into a format to be read into Tableau or other analytics tools.

No deliverable.

Exercise 2: Normalized Data Model

Create a normalized data model. See below for an example of a Normalization Process.

Normalization Process
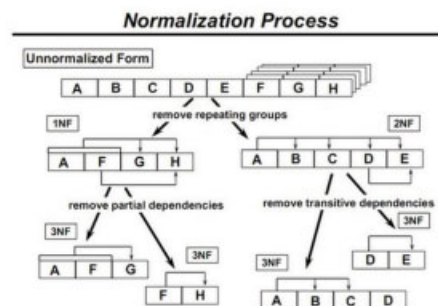
Submit your data model by **11/30/2017**. The output can be in any tool of your choosing.

Exercise 3: ETL Process

Describe each step needed to clean the data and transform it into your SQL server environment.

Submit your data transformation checklist by **12/31/2017**.

Please describe any challenges fitting the data to your data model and lessons learned.

Exercise 4: Tableau

Using Tableau, create innovative visualizations to answer the questions below:

1. Who are the top 5 most connected individuals in descending order?
2. Which data trust members are not connected to any LEADS faculty?
3. Find the most connected individuals for each data trust.
4. Find the most connected individuals for each of the 10 data trust teams
5. What is the diameter of the network? (The longest path between two individuals)?
6. What is the shortest introduction path between Diana Gumas and Desmond Jackson?
7. Who on the Data Trust isn't within 1 degrees of separation of Jason Miller?
8. Who on the Data Trust isn't within 2 degrees of separation of Paul Nagy?

Present to class on **1/10/2018**

Exercise 5: Data Product

A general discovery tool for students to learn how to build their professional network, such as The Six Degrees of Separation

Data product general discovery tool pitches to class on **3/14/2018**

Evaluation Criteria

There are no grades assigned to these assignments – the goal is to provide an authentic and collaborative data manipulation and visualization exercise.

However there will be an award for the most insightful, visually appealing and parse-able final product – as judged by an expert panel.

Winning team presents at graduation on **3/21/2018**

Supplementary Appendix B

LEADS - Capability Assessment

The goal of this self-assessment is to determine where your skills are today and where you hope your skills will be in the future. Your completion of the survey will serve as your consent to be in this research study.

Name *

First Last

1. Use Excel

	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

2. Use and create Databases

	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

3. Use a data visualization tool such as Tableau

	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

4. Design and manage an analytics project

	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

5. Identify data stewards for data needs outside your subject area

	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

6. Procificiency in data trust policies

	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

7. Use the Epic Data Warehouse as a data source

	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

8. Ensure data governance rules are followed in an analytics project

	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

9. Understand predictive analytics techniques

	Novice	Beginner	Intermediate	Expert
	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

10. Understand IRB rules and policies for clinical research

	Novice	Beginner	Intermediate	Expert
Currently	○	○	○	○
Goal	○	○	○	○

11. What other tools would you like to learn how to use?

LEADS - Class Satisfaction Assessment

Please rate the presenter and the content of today's lecture. Your completion of the survey will serve as your consent to be in this research study.

Course Name

[▼]

Are you a Faculty Member or Student?

[▼]

1. How interesting was the content of the presentation?

- ○ Not At All
- ○ Very
- ○ Somewhat
- ○ Extremely

2. How engaging was the instructor in presenting the content?

- ○ Not At All
- ○ Very
- ○ Somewhat
- ○ Extremely

3. How applicable to my daily work was the content of this lecture?

- ○ Not At All
- ○ Very
- ○ Somewhat
- ○ Extremely

4. How important to my professional growth as a data analyst was the content of this lecture?

- ○ Not At All
- ○ Very
- ○ Somewhat
- ○ Extremely

Please write any suggestions you have for how to improve this lecture:

[   ]

LEADS - Knowledge Assessment

Your completion of the survey will serve as your consent to be in this research study

Name

[   ] [   ]
First   Last

Department

[   ]

1. The chair of pulmonology comes to you with a request to help her to obtain all available clinical data for a cohort of older adult patients with restrictive lung disease so that these patients can be followed for a departmental quality improvement project with plans to publish a paper in a quality journal based on the findings. Which level of IRB-review would be most appropriate prior to executing the query?

○ No need for IRB application

○ Not Human Subject Research / Quality Improvement

○ Exempt

○ Expedited

○ Convened

2. A friend of yours who is a nurse lead in the division of gastroenterology comes to you to ask your advice. She needs help getting data to figure out the number of Outpatient cancelations this year from outside of the State of Maryland. To which data trust analytics team would you refer her for help?

○ Ambulatory Operations

○ Ambulatory Quality

○ Hospital Operations

○ Research/CCDA

○ Hospital Utilization Management

3. Your Data Trust Analytic Team leader asks you to query the Epic Data Warehouse to pull a list of ambulatory providers who practice in Howard County who are between 22-42 years old and whose last names start with L. Which is the most efficient language to use to execute this query?

○ SQL

○ R

○ Python

○ SAS

○ Jupyter

4. A doctor in the division of pediatric gastroenterology overhears you talking about your analytic work in line at Balducci's. He interrupts your conversation to ask for your advice as to the best way to carry out a preliminary data analysis to figure out how many patients are seen each year at Hopkins who are under 21 years old and carry a diagnosis of inflammatory bowel disease. He says that he "doesn't want to involve the IRB just yet." You tell him that the best program to use to perform his initial analysis is:

○ Slicer/Dicer

○ Clarity

○ Tableau

○ SQL

○ Caboodle

5. A group of clinical researchers asks for your help with the creation of an internal research registry that contains the lab data, anthropometrics, prescribed drugs, visit dates, visit diagnoses, and hospital length of stay for every patient with cystic fibrosis seen at Hopkins. They don't know where to store the data. You tell them that they should store the patient data in:

○ Google Drive/Amazon Cloud

○ A password protected Johns Hopkins clinical desktop computer

○ A password protected folder on a Johns Hopkins server

○ The ICTR SAFE Virtual Desktop

○ Home desktop computer given that no protected health information is present

6. The head of your Data Trust Analytic Team comes to you requesting that you create an analysis showing length of stay of every patient admitted to a Hopkins hospital with a heart attack over the last 4 years. She wants visuals to help providers and administrators understand the information and also wants the ability to filter data by entity, facility, and date. She wants the information to be easily accessible to a broad range of providers. To carry this out, you decide to use the following:

○ Excel

○ Slicer/Dicer

○ The ICTR SAFE Virtual Desktop

○ Tableau

○ Protenus

7. You are double checking the code of a more junior analyst on your analytics team when you notice that he hasn't included a statement specifying a table to query. You remind him that the SQL statement to specify a table to query is:

○ Select

○ From

○ Where

○ Group By

○ Order By

8. You are trying to figure out how to merge a table of billing records with a table of patient records in SQL using medical record number as the foreign key. You want all data from both tables to be incorporated in the merge. You plan to use the following to execute the query:

○ Inner Join

○ Primary Key

○ Right Outer Join

○ Left Outer Join

○ Full Outer Join

9. Your department requires you to query data out of the Epic Data Warehouse on a monthly basis. Which of the following are requirements to get individual account access to the Epic Data Warehouse? Select all that apply.

☐ Clarity Certification

☐ Membership on a Data Trust Analytic Team

☐ Epic Data Warehouse Certification

☐ Epic Data Warehouse Training

☐ Data Trust Data Stewardship Subcouncil Approval

A great package for data manipulation in R is?

○ DPLYR

○ SCIKIT

○ PANDAS

○ MATPLOTLIB

○ DMANIP

A great package for data manipulation in Python is?

○ DPLYR

○ CARET

○ PANDAS

○ GGPLOT

○ DMANIP

LEADS - Practice Assessment

The goal of this self-assessment is to determine how often you employ each one of these skills. Your completion of the survey will serve as your consent to be in this research study.

Name

[   ] [   ]
First   Last

1. Use Excel
○ Rarely   ○ Sometimes
○ Often   ○ Always

2. Use and create databases
○ Rarely   ○ Sometimes
○ Often   ○ Always

3. Use a data visualization tool such as Tableau
○ Rarely   ○ Sometimes
○ Often   ○ Always

4. Design and manage an analytics project
○ Rarely   ○ Sometimes
○ Often   ○ Always

5. Identify data stewards for data needs outside your subject area
○ Rarely   ○ Sometimes
○ Often   ○ Always

6. Follow data trust policies
○ Rarely   ○ Sometimes
○ Often   ○ Always

7. Use the Epic Data Warehouse as a data source
○ Rarely   ○ Sometimes
○ Often   ○ Always

8. Ensure data governance rules are followed in an analytics project
○ Rarely   ○ Sometimes
○ Often   ○ Always

9. Employ predictive analytics techniques
○ Rarely   ○ Sometimes
○ Often   ○ Always

10. Follow IRB rules and policies for clinical research
○ Rarely   ○ Sometimes
○ Often   ○ Always

11. What other tools do you currently use in your work?

Supplementary Appendix C Final course evaluation results

Question	Not At All	Somewhat	Very	Extremely
1. How interesting was the content of the course?	0	1	15	9
2. How engaging were the instructors in presenting the content?	0	4	13	6
3. How applicable or relevant to my daily work was the content of this course?	0	7	13	4
4. How important to my professional growth as a data analyst was the content of this course?	0	1	14	9
5. How organized were the course lectures, materials and assignments?	0	5	14	5
6. How well did the facilities, food, and scheduling communications meet your needs?	0	1	17	6
7. How likely are you recommend this course to a colleague or friend?	0	1	11	13
8. Please provide any additional feedback you would like to provide about the LEADS course.	10	13	1	0
1. Skills learned in the course were not applicable to my job tasks	16	9	0	0
2. Supervisor wants me to use other techniques	11	12	2	0
3. I did not learn enough in the course to be apply to apply the techniques	15	6	2	0