**Progress report – June 2019**

**Updated title:** Analysis of hospital based ayurvedic clinical practice to gain real world data knowledge

**Old Title:** Observational analysis of ayurvedic principles, ayurvedic hospital data, and patient outcomes

**Change of guide:**

1. Dr. Girish Tillu has taken up a full time position in Savitribai Phule Pune University. Hence he cannot continue being a guide.
2. Dr. Ashwini Godbole who has been on the DAC so far has agreed upon taking the role of a Primary Guide.
3. Dr. Girish Tillu will continue in the role of a “Co-guide” as well as will be a part of DAC going forward. Dr. Ashwini Mathur will continue in the role of a “Co-guide” and will be a part of DAC as well.

By Vinay Mahajan, Girish Tillu, Ashwini Mathur, Ashwini Godbole

Summary: The following progress has been made so far

**Course work [Online courses on Coursera.org]:**

1. Health Informatics on FHIR, Georgia Institute of Technology, by Mark L. Braunstein, MD
2. eHealth: More than just an electronic record, The University of Sydney, many instructors
3. Data science, Johns Hopkins University, by Jeff Leek, Roger D. Peng, Brian Caffo
   1. The Data Scientist’s Toolbox
   2. R Programming
   3. Getting and Cleaning Data
   4. Exploratory Data Analysis
   5. Reproducible Research
   6. Statistical Inference
   7. Regression Models
   8. Practical Machine Learning
   9. Developing Data Products
   10. Data Science Capstone

The completion certificates have been added at the end of the document.

1. Online course on National library of Ayurveda medicine (NLAM) website for Overview of Ayurveda course on <http://www.nlam.in/>
   1. Ayurveda Introduction
   2. Ayurveda Avtaran
   3. Ayurveda Literature
   4. Srushti Utpati [Cosmology]
   5. Purusha – Prakriti
   6. Tridosha
   7. Dhatu
   8. Mala
   9. Prakriti
   10. Dhatu Sara [Prominent tissue]

**Literature review (completed):**

1. Cochrane library review for clinical trials, assessment of quality by using Jadad score and CONSORT score
2. Cochrane library review for Hospital data analysis methods

**IAIM hospital data review (completed – moved to eRX database):**

1. Understanding of existing data [till 31st July 2016] – Diagnosis data, Vital sign data, Lab data
2. Data visualization using TABLEAU software and R programming language
3. Understanding of existing data in PDF files – conversion into structured CSV files

**Course work: Trans Disciplinary Knowledge framework course from 6th to 10th March 2017 @ TDU campus, Bengaluru.**

**Course work:**

As a part of Course work, I have enrolled for the following online courses offered by Coursera, <https://www.coursera.org/>. Coursera provides universal access to the world’s best education, partnering with top universities and organizations to offer courses online. These courses include recorded video lectures, auto-graded and peer-reviewed assignments, and community discussion forums. When a course is completed, a sharable electronic Course Certificate is issued.

The following sections contain a brief overview of the courses.

**Course 1: Health Informatics on FHIR, Georgia Institute of Technology, by Mark L. Braunstein, MD. [duration: 4 weeks]**

Through this course non-technical students can gain basic proficiency in health informatics: the application of computing to healthcare delivery, public health and community-based clinical research.

The overall course paradigm is the Institute of Medicine’s vision of a “Learning Health System” that uses data from actual patient care to gain new knowledge and feeds that knowledge back as care is delivered to achieve a safer, higher quality and more cost effective health delivery system.

1. Module 1 “WELCOME TO HEALTH INFORMATICS ON FHIR”: covers the US healthcare delivery system’s unique structural, economic and policy issues and the strategic role for health informatics. It also looks at the federal programs to encourage adoption of electronic record systems.
2. Module 2 “Data and Interoperability Standards”: gives a high level overview of some key health standards with a particular emphasis on the new Fast Healthcare Interoperability Resource (FHIR) standard.
3. Module 3 “Real World Applications & Challenges”: explores how these technologies are being deployed and some of their current limitations using specific commercial and open source systems as examples. It features an interview with the developers of an innovative, new EHR.
4. Module 4 “Big Data and Analytics”: presents examples of cutting edge research using “big data” and shows how analytic based tools are helping overcome some of the challenges posed in the prior module. It features an interview with developers of an innovative cloud-based service to bring together datasets and analytic tools from diverse sources.

**Course 2: eHealth: More than just an electronic record, The University of Sydney [duration: 5 weeks]**

The course “eHealth: More than just an electronic record!” is multidisciplinary in nature, and aims to equip the global audience of health clinicians, students, managers, administrators, and researchers to reflect on the overall impact of eHealth on the integration of care. It explores the breadth of technology application, current and emerging trends, and showcases both local (Australian) and international eHealth practice and research.

The entire eHealth Course consists of 5 modules and takes about 5 weeks to complete. Completion certificates are issued on the basis of participation in all 5 modules. Completing the health practice assignment in Module 5 entitles to advanced standing in some of the eHealth courses run by the Faculty of Health Sciences, University of Sydney.

Course coverage:

1. The fundamentals of eHealth and where it is heading
2. What kind of health data we are currently collecting and how it will transform healthcare in the future
3. How new technologies are helping health consumers participate in their own healthcare
4. How eHealth can improve the coordination and efficiency of healthcare and what the barriers might be

**Course 3 Data science, Johns Hopkins University [Every course is of 4 weeks, last course is of 7 weeks]**

1. The Data Scientist’s Toolbox: This is an introductory course to the main tools and ideas in the data scientist's toolbox. The course gives an overview of the data, questions, and tools that data analysts and data scientists work with. There are two components to this course. The first is a conceptual introduction to the ideas behind turning data into actionable knowledge. The second is a practical introduction to the tools that is used in the program like version control, markdown, git, GitHub, R, and RStudio.
2. R Programming: Learn how to program in R and how to use R for effective data analysis. How to install and configure software necessary for a statistical programming environment and describe generic programming language concepts as they are implemented in a high-level statistical language. The course covers practical issues in statistical computing which includes programming in R, reading data into R, accessing R packages, writing R functions, debugging, profiling R code, and organizing and commenting R code.
3. Getting and Cleaning Data: Before anyone can work with data one has to get it. This course covers the basic ways that data can be obtained. The course covers obtaining data from the web, from APIs, from databases and from colleagues in various formats. It also covers the basics of data cleaning and how to make data “tidy”. Tidy data dramatically speed downstream data analysis tasks. The course covers the components of a complete data set including raw data, processing instructions, codebooks, and processed data.
4. Exploratory Data Analysis: This course covers the essential exploratory techniques for summarizing data. These techniques are typically applied before formal modeling commences and can help inform the development of more complex statistical models. Exploratory techniques are also important for eliminating or sharpening potential hypotheses about the world that can be addressed by the data.
5. Reproducible Research: Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them. The need for reproducibility is increasing dramatically as data analyses become more complex, involving larger datasets and more sophisticated computations. Reproducibility allows for people to focus on the actual content of a data analysis, rather than on superficial details reported in a written summary. In addition, reproducibility makes an analysis more useful to others because the data and code that actually conducted the analysis are available.
6. Statistical Inference: Statistical inference is the process of drawing conclusions about populations or scientific truths from data. There are many modes of performing inference including statistical modeling, data oriented strategies and explicit use of designs and randomization in analyses. Furthermore, there are broad theories (frequentists, Bayesian, likelihood, design based, …) and numerous complexities (missing data, observed and unobserved confounding, biases) for performing inference. A practitioner can often be left in a debilitating maze of techniques, philosophies and nuance. This course presents the fundamentals of inference in a practical approach for getting things done.
7. Regression Models: Linear models, as their name implies, relates an outcome to a set of predictors of interest using linear assumptions. Regression models, a subset of linear models, are the most important statistical analysis tool in a data scientist’s toolkit. This course covers regression analysis, least squares and inference using regression models. Special cases of the regression model, ANOVA and ANCOVA are covered as well. Analysis of residuals and variability are investigated. The course covers modern thinking on model selection and novel uses of regression models including scatterplot smoothing.
8. Practical Machine Learning: One of the most common tasks performed by data scientists and data analysts are prediction and machine learning. This course covers the basic components of building and applying prediction functions with an emphasis on practical applications. The course provides basic grounding in concepts such as training and tests sets, overfitting, and error rates. The course introduces a range of model based and algorithmic machine learning methods including regression, classification trees, Naive Bayes, and random forests. The course covers the complete process of building prediction functions including data collection, feature creation, algorithms, and evaluation.
9. Developing Data Products: A data product is the production output from a statistical analysis. Data products automate complex analysis tasks or use technology to expand the utility of a data informed model, algorithm or inference. This course covers the basics of creating data products using Shiny, R packages, and interactive graphics.
10. Data Science Capstone: The capstone project class allows students to create a usable/public data product that can be used to show your skills to potential employers. Projects are drawn from real-world problems and are conducted with industry, government, and academic partners.

**Literature review (completed) – paper submitted to 1 journal but has been rejected:**

1. Cochrane library review for randomized ayurvedic clinical trials:
   1. The freely available clinical trial articles in Cochrane database are reviewed to understand the current status and quality of published ayurvedic clinical trials, using Jadad score and CONSORT score. These 2 scores provide the objective assessment on a scale of 5 and 25 points respectively.
   2. Additional sections of the comprehensive analysis: (1) the type of study designs, (2) treatments, (3) number of patients, (4) gender distribution, (5) diseases covered, (6) duration of clinical trials, (7) study end points, (8) study results, (9) comparison of treatments, etc.
2. Cochrane library review for Hospital data analysis methods:

This review provides information about the methods used in western hospital data setting. Analysis methods provided will be studied and recommended for usage on ayurvedic data analysis. Relevant methods will be applied on the IAIM hospital data going forward.

**IAIM hospital data review (completed – moved to eRX database):**

1. Study of existing patient level data [till 31st July 2016] – Diagnosis data, Vital sign data, Lab data
   1. INSTA database is reviewed for the above mentioned 3 domains. Basic descriptive statistics and statistical graphs are created to understand data, clinical anomalies, and some trends.
   2. Current database contains data for almost 40,000 distinct patients
   3. 826 distinct diseases have been coded using ACD dictionary
2. Data visualization using Tableau software[https://public.tableau.com/] and R programming language
   1. Tableau Public is free software that can allow anyone to connect to a spreadsheet or file and create interactive data visualizations for the web.
   2. Quite a few data visualizations are created to review the current data
   3. List of analyses: (1) Distinct number of patients, (2) Patients coming from different countries, (3) Number of visits, (4) Number of diseases per patient, (5) Number of patients visiting the hospital on different days of a week, (6) Boxplots for age distribution for patients with number of diseases, (7) Patient profile kind of a view of patient visits, etc.
3. Understanding of existing data in PDF files – conversion into structured CSV files
   1. Patient treatment data and inpatient discharge data are available in forms of a PDF files, a lot of important data is stored as a part of textual data. The efforts should be put to convert these into structured database

INSTA to DATABASE algorithm steps followed to get the data from the INSTA database:

1. Login onto the INSTA website
2. Go to the Patient Search tab
3. Identify unique patients
4. Use the unique URL for each patient MRNO and MR\_NO to get the source data for each patient
5. Achieve #4 by creating a VBA macro
   1. Access the source code for each patient
   2. Filter the source code and get the displayURL parameter from the Java code
   3. This parameter contains links to all the data for individual patients
   4. Save TXT files for each patient
6. Create a R program to read all the TXT files
   1. Process this data to get the patient number, visit combination
   2. Create a 1-1 match for Patient + Visit + visit date + unique number of visit to hospital
   3. This creates 4.25 lakh URLs for almost 50,000 patients in August 2017
7. Data coming from URLs, there are 20 different types of programs proving information

|  |  |  |  |
| --- | --- | --- | --- |
|  | Name of the program | | Number of URLs |
| 1 | /vitalForm/genericVitalForm.do?method | generateReport&visited | 72026 |
| 2 | /pages/GenericDocuments/GenericDocumentsPrint.do?\_method | print&forcePdf | 7238 |
| 3 | /emr/print.do?method | printConsultation&consultation\_id | 117066 |
| 4 | /emr/print.do?method | printTriageSummary&consultation\_id | 19753 |
| 5 | /dischargesummary/dischargesummaryPrint.do?\_method | print&patient\_id | 6428 |
| 6 | /Service/ServicesConductionPrint.do?\_method | print&prescription\_id | 165746 |
| 7 | /pages/DiagnosticModule/DiagReportPrint.do?\_method | printReport&reported | 29170 |
| 8 | /progress/PatientProgress.do?\_method | getPrint&mr\_no | 2623 |
| 9 | /wardactivities/DoctorOrderPrint.do?\_method | print&patientId | 2434 |
| 10 | /wardactivities/VisitSummaryRecord.do?\_method | generateReport&patient\_id | 2607 |
| 11 | /IntakeOutput/genericIntakeOutputForm.do?method | generateReport&visited | 135 |
| 12 | /dietary/DietaryMasterPrint.do?method | printPrescription&patient\_id | 7 |
| 13 | /Service/ServiceReportsPrint.do?\_method | print&forcePdf | 4 |
| 14 | /Radiology/TestDocumentsPrint.do?\_method | print&doc\_id | 84 |
| 15 | /Dietary/DietaryGenericDocumentsPrint.do?\_method | print&doc\_id | 1 |
| 16 | /MLCDocuments/MLCDocumentPrint.do?\_method | print&doc\_id | 2 |
| 17 | /wardactivities/DoctorsNotes.do?\_method | generateReport&patient\_id | 82 |
| 18 | /InitialAssessment/InitialAssessmentPrint.do?\_method | printInitialAssessment&consultation\_id | 2 |
| 19 | Discharge | *NA* | 1 |
| 20 | /GenericForms/GenericFormPrintAction.do?\_method | print&printerId | 1 |

1. The cells marked in yellow in the table above represent the pdf files not downloaded from the INSTA database.
   1. Vital form data is available in a CSV file download
   2. Lab data is available in a CSV file download
2. Use CURL command in Cygwin to extract the individual PDF files. This function works when a user is logged into INSTA system E.g.

curl 'https://182.71.223.195/instahms/Service/ServicesConductionPrint.do?\_method=print&prescription\_id=175665&printerId=2&doc\_id=175665' -H 'Accept-Encoding: gzip, deflate, br' -H 'Accept-Language: en-US,en;q=0.8' -H 'Upgrade-Insecure-Requests: 1' -H 'User-Agent: Mozilla/5.0 (Windows NT 6.3; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/60.0.3112.90 Safari/537.36' -H 'Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,\*/\*;q=0.8' -H 'Referer: https://182.71.223.195/instahms/loginForm.do' -H 'Cookie: JSESSIONID=D2D79C9FEC4DC1AB75706C93C80AE523; user\_schema@/instahms=iaim; firstMenuIndex=0; lastUser=VmluYXltYWhhamFu' -H 'Connection: keep-alive' -H 'Cache-Control: max-age=0' --compressed –insecure > FILENAME.pdf

1. Some of the filenames have “?” at the end of the file, so use DOS command to change the filename
   1. i.e. remove “?”At the end of the filename, this is done by using “rename” command as follows: rename \*.pdf? \*.pdf.
   2. The ? at the end of the files make these files unavailable for additional text processing
2. Use pdftotext command in Cygwin to convert the PDF files into txt files.
   1. Options used pdftotext –layout –nopgbrk provides the same data in text format and
   2. It reduces the file size significantly, e.g. 600 KB pdf files becomes 10 KB file,100 KB pdf file becomes 2-3 KB file.
   3. The following programming statements created TXT files for each patient in pat\_txts folder

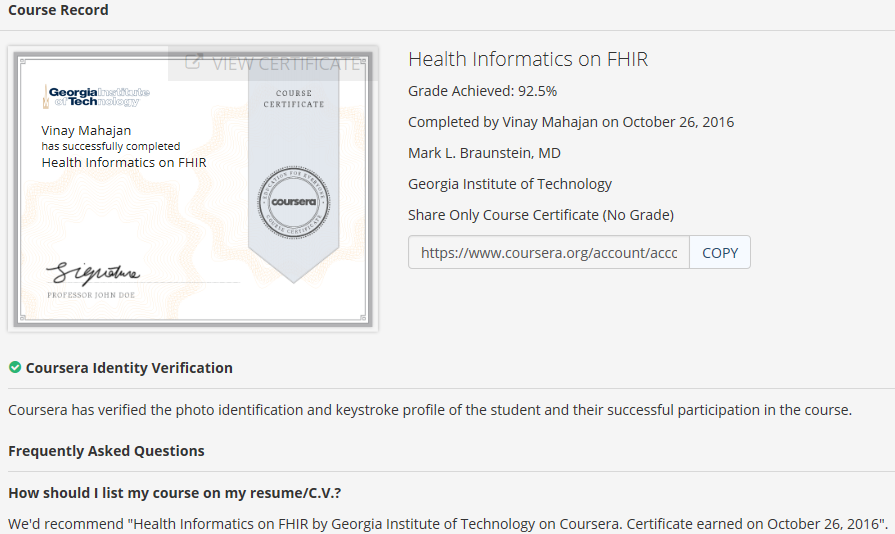
vinay@viany /cygdrive/c/Users/Lucky/Documents/Hospital\_data/04\_2017\_DOWNLOAD/pat\_pdfs

$ find . -name "\*pdf"|sed 's/.pdf//g'|head -n 10|awk '{print "pdftotext -layout -nopgbrk", $1 ".pdf", "../pat\_txts/" $1 ".txt"}'|sh

1. Use these text files as input to various R programs and create source data.
   1. Various data tables to be created
2. SQLite freeware database will be used to store the data.

Certificates:

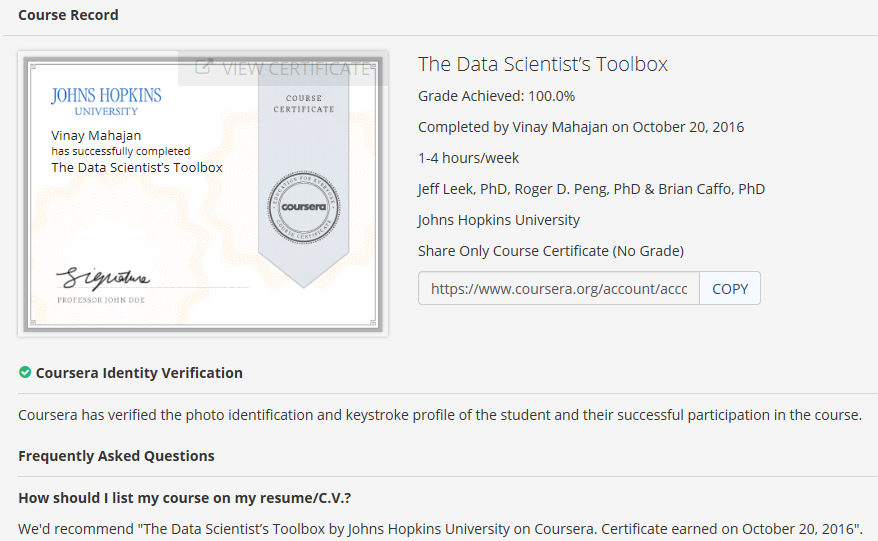
1. Health Informatics on FHIR by Georgia Institute of Technology on Coursera. Certificate earned on October 26, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/X3N9H4LEUKDU)



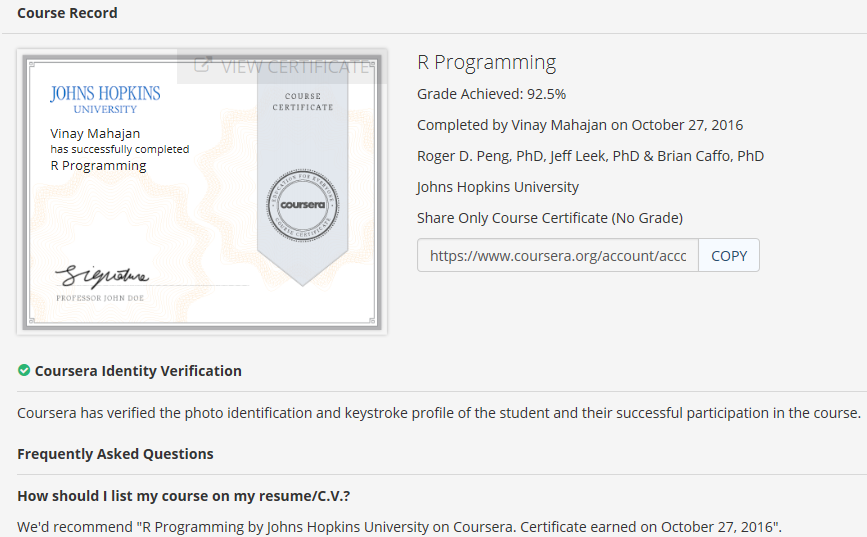
1. eHealth: More than just an electronic record by The University of Sydney on Coursera. Certificate earned on November 20, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/6F39DUQYQGSB)



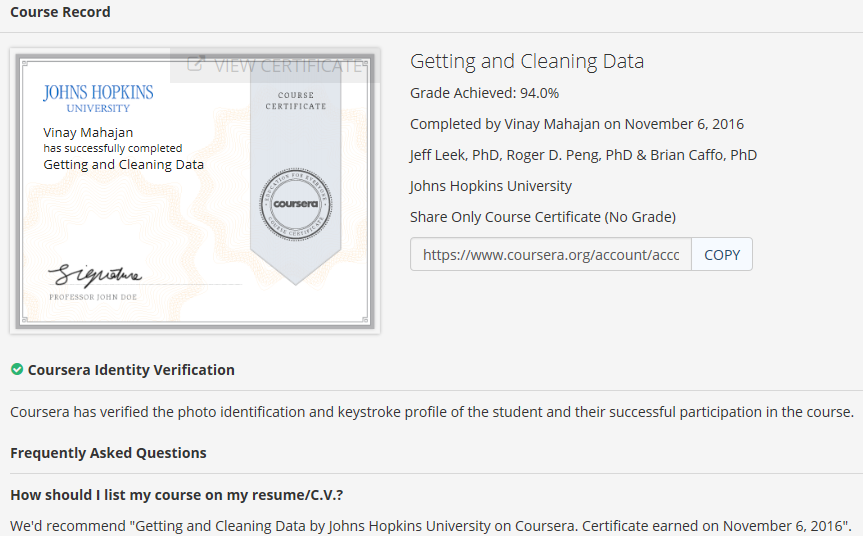
1. The Data Scientist’s Toolbox by Johns Hopkins University on Coursera. Certificate earned on October 20, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/BXXL8J883L4J)



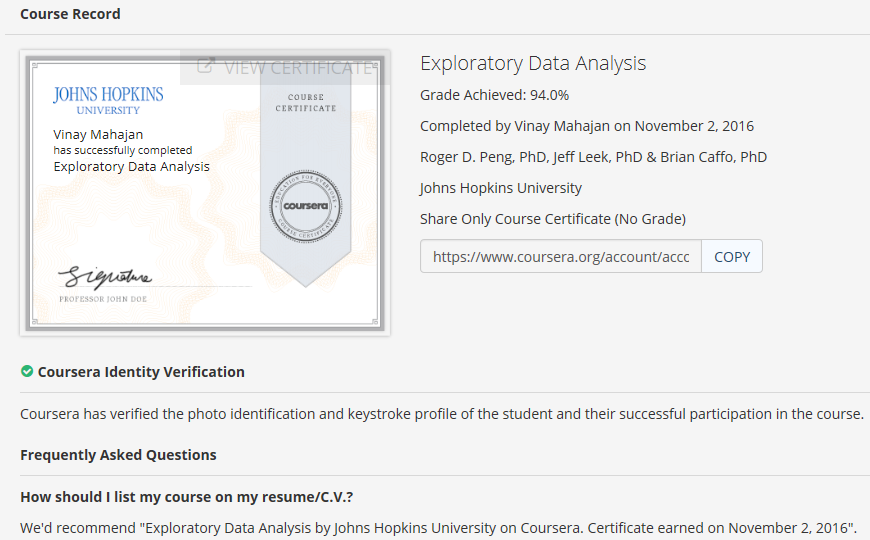
1. R Programming by Johns Hopkins University on Coursera. Certificate earned on October 27, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/L2AWM6WTLSPE)



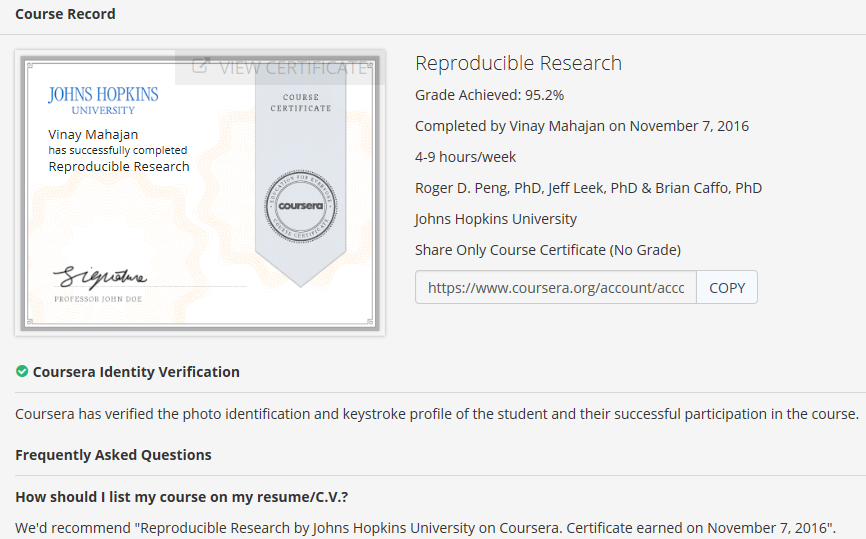
1. Getting and Cleaning Data by Johns Hopkins University on Coursera. Certificate earned on November 6, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/85SY8LSFT6MB)



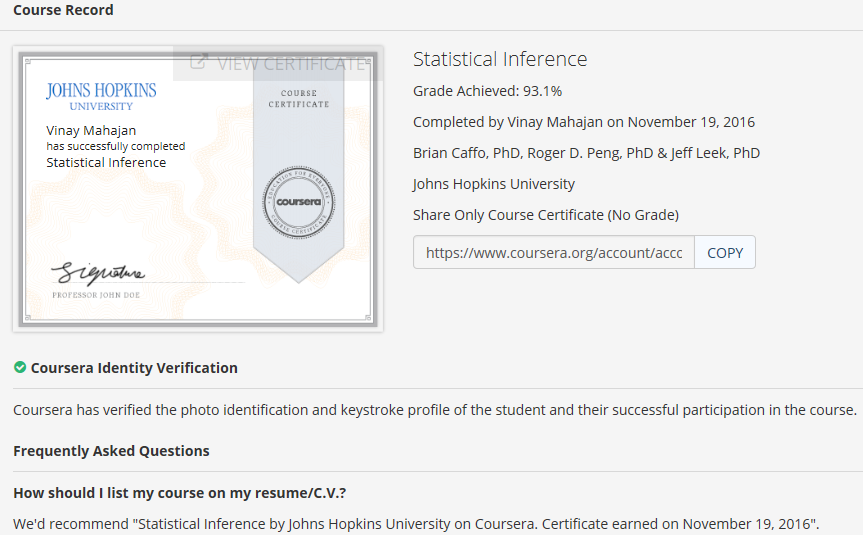
1. Exploratory Data Analysis by Johns Hopkins University on Coursera. Certificate earned on November 2, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/PBGB3MJXFABP)



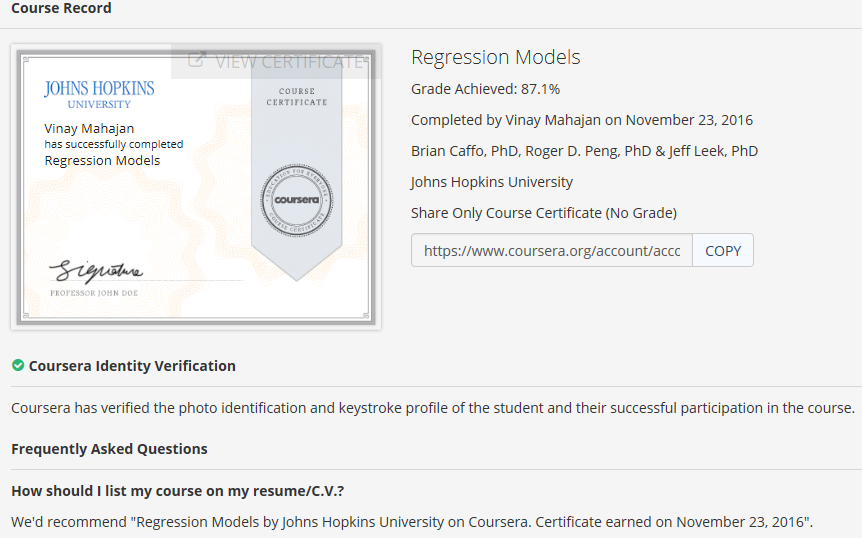
1. Reproducible Research by Johns Hopkins University on Coursera. Certificate earned on November 7, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/W5TUYXVH3HQ8)



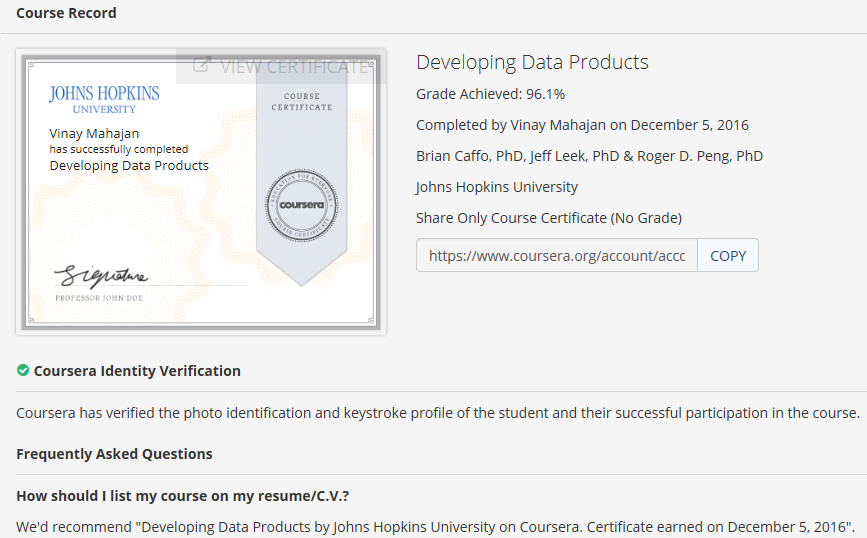
1. Statistical Inference by Johns Hopkins University on Coursera. Certificate earned on November 19, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/5BL5LQKK3ZHM)



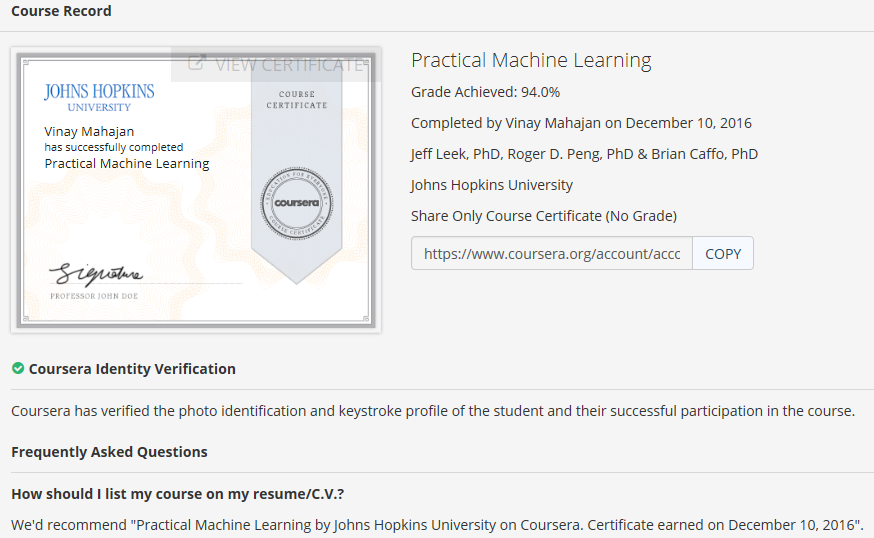
1. Regression Models by Johns Hopkins University on Coursera. Certificate earned on November 23, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/A7N4G8CNR8G5)



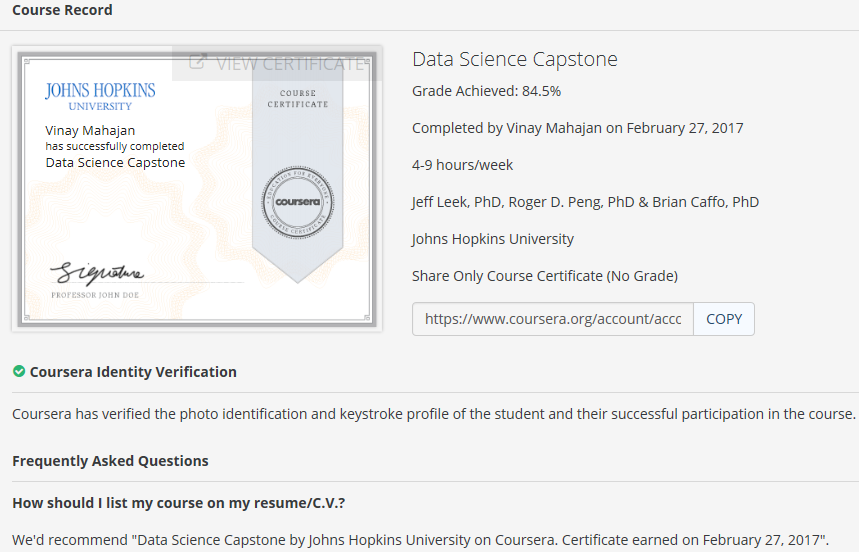
1. Developing Data Products by Johns Hopkins University on Coursera. Certificate earned on December 5, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/2K6B6ATU5RM2)



1. Practical Machine Learning by Johns Hopkins University on Coursera. Certificate earned on December 10, 2016 [Link](https://www.coursera.org/account/accomplishments/certificate/Y96XABGLXWGB)



1. Data science capstone Learning by Johns Hopkins University on Coursera. Certificate earned on February 27, 2017 [Link](https://www.coursera.org/account/accomplishments/certificate/XUU65A8BUKSB)



Time frame: April 2018 onwards.

Alternative approach using the SQL database built by the eRX team

Login on the SQL database and access the hospital database. There are approximately 190 tables holding the different components of hospital data. There is a need to understand the data structure, data contents and quantity at disposal.

General flowchart for data processing:

1. Login on the SQL server using the credentials
2. Use the IAIM schema and access the following tables to generate base table with Demographics information, Patient Visit information and prescribed treatments
   1. STATE\_MASTER
   2. COUNTRY\_MASTER
   3. CITY
   4. STATE
   5. PATIENT\_DETAILS
   6. PATIENT\_REGISTRATION
   7. MRD\_DIAGNOSIS
   8. PATIENT\_PRESCRIPTION
   9. PATIENT\_MEDICINE\_PRESCRIPTIONS
   10. IP\_PRESCRIPTION
   11. SERVICES\_PRESCRIBED
   12. SERVICES
   13. MEDICINE\_SALES\_VIEW
3. There are many CRF pages built to collect relevant Ayurvedic data, measurement data, Hospital visit data, food / exercise advice, etc. This data is present in the following tables:
   1. PATIENT\_SECTION\_DETAILS
   2. PATIENT\_SECTION\_VALUES
   3. SECTION\_MASTER
   4. SECTION\_FIELD\_OPTIONS
   5. SECTION\_FIELD\_DESC
   6. PATIENT\_CONSULTATION\_FIELD\_VALUES
4. The datasets created in steps 2 and 3 are further processed using R programming language and the analysis ready datasets are created
5. Patients suffering with metabolic diseases and Rheumatoid / Musculoskeletal diseases are sub-setted for analysis. There are 10 metabolic diseases and 106 Rheumatoid / Musculoskeletal diseases defined by ayurvedic vaidyas, based on ACD classification
6. The following github repository stores the analysis carried out so far: <https://github.com/Coursephd/PostgreSQL>

Presentation at 8th World Ayurveda conference, Ahmedabad Dec 2018: I was invited to present the real world data science view and patterns in data in transdisciplinary session for ayurveda. The presentation was based on the data analysis carried out and additional scientific material from Dr. Girish Tillu and Dr. Ashwini Mathur.

The chapter wise split of the thesis document will be as follows (subject to change in future based on discussions and ongoing research):

Chapter 1: Retrospective analysis of published literature, this chapter will provide empirical evidence about the status of Ayurvedic clinical research.

Chapter 2: Parallels between the ICH guidelines and Classical Samhitas to form the underlying fundamental basis of scientific principles. Perspectives on data science relevant to the research question will be demonstrated.

Chapter 3: Description about the hospital data, 5 W’s of data. Who, Why, What, Where, When and How.

Chapter 4: Description of software and analytical tools used, statistical analysis methods employed, data analysis and visualization method framework

Chapter 5: 2 to 3 specific disease areas to be used for detailed analysis.

Time frame: Jan 2019 onwards.

Discovering unexpected and useful patterns in databases is a fundamental data mining task. One of the most popular data mining tasks on sequences is sequential pattern mining. It consists of discovering interesting subsequences in a set of sequences, where the interestingness of a subsequence can be measured in terms of various criteria such as its occurrence frequency, length, and profit. Sequential pattern mining has many real-life applications since data is encoded as sequences in many fields such as bioinformatics, e-learning, market basket analysis, text analysis, and webpage click-stream analysis. The patient level data generated for diseases as well as prescribed treatments provide a sizeable data for data mining task.

In the talk "P for Patterns" presented at the World Ayurvedic Conference, we had initiated some data mining activities. From Jan 2019 onward, the research work on the Pattern finding algorithms using an Open Source Java based Library SPMF (Sequential Pattern Mining Library) was initiated. The details for the algorithms could be found at the following link:

<http://www.philippe-fournier-viger.com/spmf/index.php>

[Philippe Fournier-Viger](http://www.philippe-fournier-viger.com/), Ph.D. is a data mining researcher and professor. He is the founder and main author of the SPMF data mining software, an open-source software offering more than 150 algorithms for discovering itemsets, association rules, sequential patterns and rules in sequences and transactions. The SPMF software has been cited in more than 640 papers and was visited by more than 700,000 visitors since 2010. He has also written or participated in more than 200 research papers which have received more than 3000 citations. He is one of the two editor-in-chief of the Data Science and Pattern Recognition journal. He edited the book “High-Utility Pattern Mining” (Springer).

The SPMF utility offers implementations of 171 data mining algorithms as of Feb 2019:

1. association rule mining,
2. itemset mining,
3. sequential pattern
4. sequential rule mining,
5. sequence prediction,
6. periodic pattern mining,
7. episode mining
8. high-utility pattern mining,
9. time-series mining.
10. clustering and classification,

Medicine and disease data are classified before and after for each reference disease (as explained in “15.15 Dataset with each disease considered as a reference disease having day 1”). Some of the data mining algorithms available on the SPMF library are executed on the underlying data.

The disease data is split into unique disease trajectories before the onset of reference disease, after disease and across all times. Same is done for the medicine data as well.

The data mining utility accepts the data in numerical values hence the data is converted from textual data to numerals. The text file is passed to individual algorithms to generate the support, confidence and lift.

Work done on CKD, Cancer diseases, Tableau displays