

Hands-On Full Life Cycle Data Science Workshop

WS-04

Steve Johnson PhD, University of Minnesota
Tamara Winden PhD, University of Kansas
Lisiane Pruinelli PhD, University of Minnesota



Preparation for this workshop

As you arrived here today, instructions were sent out to you with a link to download the workshop Jupyter Notebook to your PC.

Please do the following:

1. Download that file (via the link sent today) to your PC
2. Launch your Jupyter Notebook
3. Within Jupyter, click the “Upload” button to get it into Jupyter
4. Select the file from the list of files to run it

Lisiane Pruinelli: I have no relevant relationships with commercial interests to disclose.

Steve Johnson: I and my spouse/partner have no relevant relationships with commercial interests to disclose.

Tamara Winden: I have no relevant relationships with commercial interests to disclose.

Learning Objectives

After participating in this session the learner should be better able to:

- Understand and be able to apply data science techniques to health care data
- Understand the challenges of working with EHR data
- Demonstrate modeling techniques for health care data

Introductions

- Speakers
 - Steve Johnson, VP Technology and Clinical Informatics, Wolters Kluwer (Provation)
 - Tamara Winden, Chief Research Informatics Officer, Assistant Professor, University of Kansas
 - Lisiane Pruinelli, Assistant Professor, School of Nursing, University of Minnesota
- Attendees

Housekeeping

- Break at 11:30
- This is an interactive session and a large group so we encourage asking for help and helping each other where possible
- Were here to learn together

Agenda

1. Background
2. Research question
3. Jupyter Notebook
4. Data Preparation
5. Exploratory Data Analysis
6. Modeling and Analytics
7. Model Evaluation
8. Deployment

- EHR = large electronic data sets that will help answer operational and clinical questions.
- Combining data sets from multiple organizations results in truly big data of tens of millions of patients to address population health and inform clinical research.
- Challenge of health care data
 - Data quality
 - Data privacy and security
 - Lack of standards
 - Extremely complex

- Solving real-world problems using data science
- Data science
 - Data scientist is a data analyst who lives in California
 - “Data science field with a broad scope, encompassing approaches for generation, characterization, management, storage, analysis, visualization, integration and use of large, heterogeneous data sets that have relevance to population health.¹”
- This workshop will use a hands-on approach to demonstrate big data science to illustrate these issues and discuss approaches to dealing and analyzing them for better health care initiatives.

1. NOT-LM-17-006: Request for Information (RFI): Next-Generation Data Science Challenges in Health and Biomedicine.

The Research Question

FINER Criteria for Good Research Questions

F	Feasible	<ul style="list-style-type: none">•Adequate number of subjects•Adequate technical expertise•Affordable in time and money•Manageable in scope
I	Interesting	<ul style="list-style-type: none">•Getting the answer intrigues investigator, peers and community
N	Novel	<ul style="list-style-type: none">•Confirms, refutes or extends previous findings
E	Ethical	<ul style="list-style-type: none">•Amenable to a study that institutional review board will approve
R	Relevant	<ul style="list-style-type: none">•To scientific knowledge•To clinical and health policy•To future research

Farrugia P, Petrisor BA, Farrokhyar F, Bhandari M. Research questions, hypotheses and objectives. *Canadian Journal of Surgery*. 2010;53(4):278-281.

Understanding the Research Question

Workshop research question:

“Can we predict which patients might overdose on opioids?”

Institutional review board and data use

- IRB approval
- Data sharing agreements

Data requests, extraction:

- What does it take to get data at a health system
- Data extract process

For this workshop:

- We'll use synthetic EHR data for this workshop. A data dictionary will be provided.

Agenda

1. Background
2. Research question
3. Jupyter Notebook
4. Data Preparation
5. Exploratory Data Analysis
6. Modeling and Analytics
7. Model Evaluation
8. Deployment

Agenda

1. Background
2. Research question
3. Jupyter Notebook
4. Data Preparation
5. Exploratory Data Analysis
6. Modeling and Analytics
7. Model Evaluation
- 8. Deployment**

Deployment: To the bedside...

- Clinical Decision Support (CDS) Tools

Clinical decision support (CDS) provides clinicians, staff, patients or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care. CDS encompasses a variety of tools to enhance decision-making in the clinical workflow.

- Pop-up message
 - Passive messages
 - Snapshot views
 - Reporting tools (reports, dashboards)
 - Orders/order sets
- Outcomes analysis post go-live
- Continuous improvement

Getting it done...

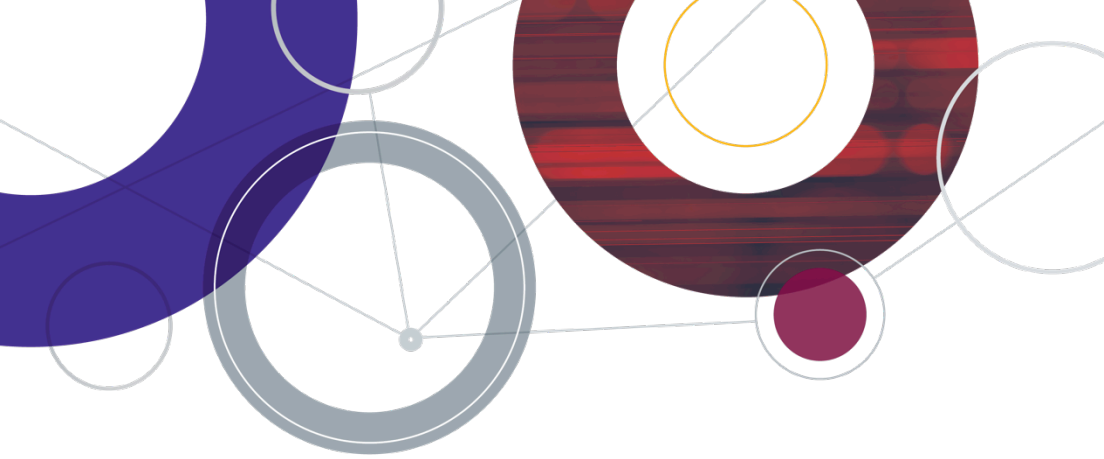
- **Clinician Support**
 - Must have a clinician champion
 - Support from impacted user group
 - Who is going to do data entry, act on CDS tool?
- **IT Change Management**
 - Involve IT early and often (clinical informatics or optimization analysts)
 - Governance and approvals
 - Testing
 - Workflow documentation
 - Training and materials
 - Implementation

Thank you!

Email us at:

- Steve Johnson joh06288@umn.edu
- Tamara Winden, twinden@kumc.edu
- Lisiane Pruinelli pruin001@umn.edu

This workshop is partially funded by the University of Minnesota, School of Nursing Foundation grant



AMIA is the professional home for more than 5,400 informatics professionals, representing frontline clinicians, researchers, public health experts and educators who bring meaning to data, manage information and generate new knowledge across the research and healthcare enterprise.



f @AMIAInformatics

🐦 @AMIAinformatics

in Official Group of AMIA

YouTube @AMIAInformatics

#WhyInformatics

Question 1

Question: When using EHR data in an observational study, which of the following is correct? IRB review...

- A. Does not apply since this is not a clinical trial.
- B. Is not needed since the data is being collected in the EHR to document care
- C. Is not needed since we are only accessing the EHR and not interacting directly with patients
- D. Is required since we are using patient-specific health data

Question 2

Question: How do data quality issues impact data science results?

- A. Because EHR data is used to document healthcare, there are no data quality issues
- B. Data quality can significantly impact results, making them biased or even invalid
- C. Data quality issues can be ignored because machine learning and AI will take care of it
- D. Poor data quality doesn't impact results, because the data quality issues disappear with enough data

Question 3

Question: A good practice for exploratory data analysis is to use a number of techniques to visualize your dataset. Which of the following is NOT a good reason to visualize your data?

- A. Data visualization allows you to better understand and "get to know" your data
- B. Humans can spot visual patterns much better than looking at descriptive statistics
- C. Visualization allows you to see large amounts of data summarized and simplified
- D. Visualization looks nicer than just having tables of numbers

Question 4

Question: After receiving a dataset for a data science project, a researcher will spend a significant amount of the total project time performing data preparation and exploratory data analysis. Which of the following is NOT a data preparation task?

- A. Assessing missing data and imputing appropriate values or removing records
- B. Adding variables, transforming data and joining data tables
- C. Ensuring all data is of the appropriate type and is consistent with the data mode
- D. Obtaining the appropriate level of IRB approval for the project

Question 5

Question: Many datasets have imbalanced data where there is a much smaller amount of data for positive classes than negative ones. For example, the prevalence of a number of diseases is low in the overall population (i.e. Ebola). Which is the best method to evaluate the performance of a classification model for imbalanced data?

- A. Accuracy
- B. F1 Score
- C. Precision
- D. Recall

Question 6

Question: Once data are properly prepared and cleaned, researchers should select the most appropriate modeling algorithm depending on the data and the research question. Which modeling technique does NOT work well on categorical variables?

- A. Classification
- B. Linear Regression
- C. Neural Network
- D. Random Forest

Question 7

Question: Once you have a good performing model and you have tested it sufficiently on your training data, it is time to implement it in the real world. Which tasks are essential for successful model deployment?

- A. Ensure the model is activated at the right time in the workflow, appropriately alerts the clinician and allows specific action
- B. It is not necessary to get buy-in from the clinicians, since the model embedded in the EHR will tell them what to do
- C. Such models should only be created for research purposes and should not be used to support direct patient care
- D. The model should only be implemented in the latest version of the EHR software

Question 8

Question Data science is fast becoming very useful in healthcare, but success requires close collaboration between people with clinical knowledge and people with data and computer expertise. Which of the following is true about data science in healthcare? Data science...

- A. Can be understood sufficiently by everyone in the healthcare organization
- B. Is easy since we give raw EHR data to machine learning algorithms to produce highly accurate results
- C. Is somewhat of a fad and it won't have a significant impact in improving healthcare
- D. Only needs to be understood by a few people in the healthcare organization

Question 1

Question: When using EHR data in an observational study, which of the following is correct? IRB review...

- A. Does not apply since this is not a clinical trial.
- B. Is not needed since the data is being collected in the EHR to document care
- C. Is not needed since we are only accessing the EHR and not interacting directly with patients
- D. Is required since we are using patient-specific health data

Answer: d. is required since we are using patient-specific health data

Explanation: IRB review is always required when identifiable patient information is used in a study, even if the patient isn't directly involved.

Reference: <https://aspe.hhs.gov/report/feasibility-using-electronic-health-data-research-small-populations/privacy-and-security-conditions-required-research-using-ehr-and-other-electronic-health-data>

Question 2

Question: How do data quality issues impact data science results?

- A. Because EHR data is used to document healthcare, there are no data quality issues
- B. Data quality can significantly impact results, making them biased or even invalid
- C. Data quality issues can be ignored because machine learning and AI will take care of it
- D. Poor data quality doesn't impact results, because the data quality issues disappear with enough data

Answer: b. Data quality can significantly impact results making them biased or even invalid

Explanation: It is critical to understand how complete, consistent and correct your data is for the purpose that you have for the data. Having more data may just give you more errors and using any data science technique on incorrect data will likely produce incorrect results.

Reference: Brown J, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Medical care. 2013;51(8 0 3):S22-S29.
doi:10.1097/MLR.0b013e31829b1e2c

Question 3

Question: A good practice for exploratory data analysis is to use a number of techniques to visualize your dataset. Which of the following is NOT a good reason to visualize your data?

- A. Data visualization allows you to better understand and "get to know" your data
- B. Humans can spot visual patterns much better than looking at descriptive statistics
- C. Visualization allows you to see large amounts of data summarized and simplified
- D. Visualization looks nicer than just having tables of numbers

Answer: d. Visualization looks nicer than just having tables of numbers

Explanation: The reason visualization is so important is that it engages the visual pattern matching center of our brain to allow you to see relationships and trends in data that are hard to discern from tables of numbers.

Reference: Tufte, E., & Graves-Morris, P. (2014). The visual display of quantitative information.; 1983.

Question 4

Question: After receiving a dataset for a data science project, a researcher will spend a significant amount of the total project time performing data preparation and exploratory data analysis. Which of the following is NOT a data preparation task?

- A. Assessing missing data and imputing appropriate values or removing records
- B. Adding variables, transforming data and joining data tables
- C. Ensuring all data is of the appropriate type and is consistent with the data mode
- D. Obtaining the appropriate level of IRB approval for the project

Answer: d. Obtaining the appropriate level of IRB approval for the project

Explanation: Obtaining the appropriate level of IRB approval for the project should be done before you request and receive your data. There will be significant time spent in preparing your data for analytics including dealing with missing information, inconsistent formats and creating additional variables from the underlying data. Data preparation, visualization and exploration can consume 80% of the time spent on a data science project.

Reference: Pyle, D. (1999). Data preparation for data mining (Vol. 1). Morgan Kaufmann.

Question 5

Question: Many datasets have imbalanced data where there is a much smaller amount of data for positive classes than negative ones. For example, the prevalence of a number of diseases is low in the overall population (i.e. Ebola). Which is the best method to evaluate the performance of a classification model for imbalanced data?

- A. Accuracy
- B. F1 Score
- C. Precision
- D. Recall

Answer: b. F1 Score

Explanation: Imbalanced data has a smaller amount of positive class data. Precision, which is $TP / (TP + FP)$ doesn't take into account FN and TN and Recall, which is $TP / (TP + FN)$ doesn't take into account FP and TN which can lead to biased results. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution.

References:

<https://www.kdnuggets.com/2016/12/best-metric-measure-accuracy-classification-models.html>

<http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/>

Question 6

Question: Once data are properly prepared and cleaned, researchers should select the most appropriate modeling algorithm depending on the data and the research question. Which modeling technique does NOT work well on categorical variables?

- A. Classification
- B. Linear Regression
- C. Neural Network
- D. Random Forest

Answer: b. Linear Regression

Explanation: Linear regression is appropriate for continuous variables. Categorical variables can be directly used in random forest models. Techniques, such as “one hot encoding,” can be used to incorporate categorical variables into neural networks and other classifiers.

Reference: Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Question 7

Question: Once you have a good performing model and you have tested it sufficiently on your training data, it is time to implement it in the real world. Which tasks are essential for successful model deployment?

- A. Ensure the model is activated at the right time in the workflow, appropriately alerts the clinician and allows specific action
- B. It is not necessary to get buy-in from the clinicians, since the model embedded in the EHR will tell them what to do
- C. Such models should only be created for research purposes and should not be used to support direct patient care
- D. The model should only be implemented in the latest version of the EHR software

Answer: a. Ensure the model is activated at the right time in the workflow, appropriately alerts the clinician and allows specific action

Explanation: Models not only have to be predictive, but the prediction needs to be presented to a clinician at a time in a clinical workflow when they can use the information to perform an action or intervention to potentially improve an outcome. Clinicians must be intimately involved in developing and deploying the model for it to be successful.

Reference: Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., ... & Middleton, B. (2003). Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association*, 10(6), 523-530.

Question 8

Question Data science is fast becoming very useful in healthcare, but success requires close collaboration between people with clinical knowledge and people with data and computer expertise. Which of the following is true about data science in healthcare? Data science...

- A. Can be understood sufficiently by everyone in the healthcare organization
- B. Is easy since we give raw EHR data to machine learning algorithms to produce highly accurate results
- C. Is somewhat of a fad and it won't have a significant impact in improving healthcare
- D. Only needs to be understood by a few people in the healthcare organization

Answer: a. Can be understood sufficiently by everyone in the healthcare organization

Explanation: If everyone involved in a data science initiative understands the basic concepts of how data science works, the project will have a higher probability of success. Some of the team members will have more clinical knowledge and some will have more technical knowledge, but everyone can speak the same “data science” language and can help reduce the knowledge gap. Having many knowledgeable people participate in data science initiatives improves the chances for success.

Reference: <https://www.cio.com/article/3234353/analytics/the-secrets-of-highly-successful-data-analytics-teams.html>