February 19, 2024

STEPHEN OMONDI

946934043

soomondi@syr.edu

ms-ads-portfolio project

graduation presentation

# IST 659 Data Admin Concepts & Database Management

## **Lessons Learned**

1. Understanding the fundamental data and database concepts
2. Understanding the database development lifecycle
3. How to create databases and database objects using popular database management system products
4. How to solve problems using data modeling and data normalization techniques
5. Developing insights into future data management tool and technique trends.
6. How to recommend and justify strategies for managing data security, privacy, audit control, fraud detection, backup, and recovery.

## **Course Concepts**

Week 1 | The Relational Data Model

Week 2 | Conceptual Modeling

Week 3 | Logical Modeling

Week 4 | Normalization

Week 5 | Physical Database Design

Week 6 | Querying, Inserting, Updating, and Deleting Data

Week 7 | Advanced Querying

Week 8 | Database Programming

Week 9 | Database Administration

Week 10 | User Interface Design

**Course Outcomes**

96.67%

## **Final Project:**

I built a complete database for a fictional logistic company with entities such as employees, drivers, routes and packages. I included entire CRUD operations to build and manage the database.

## **Tools and techniques used:**

SQL-server, Visio, draw.io, Access, data modeling, SQL, data modeling, relational data models, normalization, ERD, Transaction Management, Concurrency, distributed databases, multitier client/server architectures, web-based database applications, data warehousing, NoSQL

[Click here to view Project in GitHub](https://github.com/soomondi/syracuse/tree/main/IST%20659%20-%20Data%20Admin%20Concepts%20%26%20Database%20Management/Final%20Project)

# IST 687 Applied Data Science

**Lessons Learned**

1. Essential concepts and characteristics of data.
2. Scripting/code development for data management using R and R-Studio.
3. Principles and practices in data screening, cleaning, and linking.
4. Communication of results to decision makers.
5. How to identify a problem and the data needed for addressing the problem.
6. How to perform basic computational scripting using R and other optional tools.
7. How to transform data through processing, linking, aggregation, summarization, and searching.
8. How to organize and manage data at various stages of a project lifecycle.
9. How to determine appropriate techniques for analyzing data.

## **Course Concepts**

Week 1 | Introduction

Week 2 | Rows and Columns

Week 3 | Writing Functions and Descriptive Stats

Week 4 | Sampling

Week 5 | Accessing Data

Week 6 | Pictures Versus Numbers

Week 7 | Map Mashup

Week 8 | Linear Modeling

Week 9 | Data Mining

Week 10 | Text Mining

## **Course Outcomes**

90.89%

## **Homework/Assignments:**

A series of incremental learning how to use R for statistical analysis via homework assignments of incremental complexity.

## **Tools and techniques used:**

R and R-studio set up, Sampling, Linear Modelling, Data Mining, Text Mining

[Click here to view course portfolio](https://github.com/soomondi/syracuse/tree/main/IST%20687%20-%20Applied%20Data%20Science/Final%20Project)

# IST 707 Data Analytics

## **Lessons Learned**

1. How to document, analyze, and translate data mining needs into technical designs and solutions.
2. How to apply data mining concepts, algorithms, and evaluation methods to real-world problems.
3. How to employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy.

## **Course Concepts**

Week 1 | What Is Data Mining?

Week 2 | Data Preparation

Week 3 | Association Rule Mining

Week 4 | Clustering Techniques

Week 5 | Decision Tree Theory

Week 6 | Model Evaluation

Week 7 | Naive Bayes

Week 8 | SVM and Others

Week 9 | Text Mining

Week 10 | Literature Review / Student Project Presentation

## **Course Outcomes**

94.90%

## **Final Project**

Worked on a Pre-diabetes prediction models with machine learning using Kaggle data taken from hospitals in Frankfurt, Germany. Conducted training and splitting of the dataset and employed different machine learning algorithms to model and analyze the data such as Naive Bayes, Decision Trees, SVM polynomial, SVM Linear, SVM radial Kernel, and Random Forest. I provided discussion on accuracy of each approach.

## **Tools and techniques used:**

Data mining, data preparation, concept description, association rule mining, classification, clustering, evaluation, and analysis. Waka, Rapid Miner, SVM, Random Forest, Decision Trees, Naive Bayes.

[Click here to view course portfolio](https://github.com/soomondi/syracuse/tree/main/IST%20707%20-%20Data%20Analytics/Final%20Project)

# IST 772 Quantitative Reasoning in Data Science

## **Lessons Learned**

1. Gained knowledge of contemporary inferential statistical concepts (from the perspective of two contemporary philosophies) and data analysis strategies by making sensible choices about:

* How data collection, the data themselves, and the analysis processes relate to the kinds of inferences that can be drawn
* What kinds of analysis will be feasible and developing the skill of planning data collection and measurement to facilitate appropriate analysis.

1. How to prepare data for analysis, including screening data, dealing with missing data, doing data transformations
2. How to test assumptions that data must meet for analyses and inferences to be reasonable.
3. How to interpret data analysis results and outputs and communicating them to others using language that accurately describes uncertainty.

## **Course Concepts**

Week 1 | Essential Statistical Concepts and Vocabulary

Week 2 | Events and Probabilities

Week 3 | Sampling and Long-Run Probabilities

Week 4 | The Logic of Inference with Confidence Intervals

Week 5 | Bayesian and Traditional Hypothesis Testing

Week 6 | Comparing Groups and Analyzing Experiments

Week 7 | Associations Between Variables

Week 8 | Linear Multiple Regression

Week 9 | Logistic Regression

Week 10 | Analyzing Change Over Time

## **Course Outcomes**

81%

## **Final Project**

I conducted a statistical analysis of vaccine reporting data from kindergartens in California - analyzing proportions of private and public participation. Investigated conditional, medical, and religious factors in predicting whether a school is private or public. Investigated conditional percentages based on percentage of specific vaccines that are missing. Used both Frequentist and Bayesian approaches to all predictions.

[Click here to view the Project](https://github.com/soomondi/syracuse/tree/main/IST%20772%20-%20Quantitative%20Reasoning%20in%20Data%20Science/Final%20Project)

# IST 722 Data Warehouse

## **Lessons Learned**

1. Expert work with SQL
2. Understanding Data warehousing architectures
3. Building a modern data warehouse
4. Dimensional Modeling and development
5. ETL Development
6. Business Intelligence
7. Big Data and the Data Warehouse

## **Course Concepts**

Week 1 | Introduction to Data Warehousing

Week 2 | Data Warehouse Architectures

Week 3 | Building the Data Warehouse

Week 4 | Dimensional Modeling

Week 5 | Dimensional Model Development

Week 6 | Introduction to ETL

Week 7 | ETL Development

Week 8 | Managing the Data Warehouse

Week 9 | Business Intelligence

Week 10 | Big Data and the Data Warehouse

## **Course Outcomes**

96.70%

## **Final Project**

I created a data warehouse that facilitated corporate analysis across businesses and conducted a BI analysis of those businesses.

## **Tools and techniques used:**

Data profiling, High level Dimensional Modeling, Detailed dimensional modeling, ETL (Extract, Transform, Load), Business Intelligence

[**Click here to view the course portfolio**](https://github.com/soomondi/syracuse/tree/main/IST%20722%20-%20Data%20Warehouse/Final%20Project/STEPHEN_OMONDI_FINAL_PROJECT_SUBMISSION)

# IST 736 Text Mining

## **Lessons Learned**

1. Computational methods to find patters in large text corpora,
2. How to use Machine learning and Natural Language processing to process text corpora
3. How to use statistical design in analyzing text tokens

## **Course Content**

Week 1 | Introduction to Text Mining

Week 2 | Document Vectorization

Week 3 | Corpus Analysis

Week 4 | Naive Bayes for Text Categorization

Week 5 | Text Classifier Evaluation and Human Annotation

Week 6 | scikit-learn

Week 7 | SVMs for Text Categorization

Week 8 | Document Clustering and Topic Modeling

Week 9 | Literature Review on Text Mining Applications/Project Idea Presentation

Week 10 | New Topics in Text Mining/Project Clinic and Presentation

## **Course Outcomes**

84.60%

## **Final Project**

I analyzed public tweets on 2019 democratic presidential debates immediately after the debates. I leverage Twitter posts with public sentiment on policy issues, options, support and criticism of the positions taken by the candidates.

## **Tools and techniques used:**

Machine Learning, Natural Language Processing, Statistics, Business Intelligence, information organization and access, social behavior analysis, digital humanities.

[Click here to view course portfolio](https://github.com/soomondi/syracuse/tree/main/IST%20736%20-%20Text%20Mining/Final%20Project)

# IST 769 Advanced Database Management

## **Lessons Learned**

1. How to work with relational, document, key-value, columnar, and streaming database systems through the lens of CAP theorem.
2. Working with relational, Hadoop and NoSQL databases.

## **Course content**

Week 1 | Course Introduction: Setting the Stage

Week 2 | Relational Databases: Programming

Week 3 | Relational Databases: Transactions, Concurrency, and Triggers

Week 4 | Relational Databases: Performance, Security, and noSQL

Week 5 | Hadoop: HDFS, MapReduce and YARN

Week 6 | Hadoop: Pig, Hive and Hcatalog

Week 7 | Hadoop: Hbase, Impala, and Other Tools

Week 8 | NoSQL: MongoDB and Redis

Week 9 | NoSQL: Cassandra

Week 10 | NoSQL: Kafka and KSQL

## **Course Outcomes**

87.50%

## **Final Project**

Hands-on work writing SQL and NoSQL variants for Streaming (Kafka), document (Mongo), columnar (Cassandra), relational (SQL) database systems.

## **Tools and techniques used:**

Kafka, Cassandra, Hadoop, Mongo, SQL, CAP Theorem

[Click here to view course portfolio](https://github.com/soomondi/syracuse/tree/main/IST%20769%20-%20Advanced%20Database%20Management/Final%20Project)**IST 652 Scripting for Data Analysis**

## **Lessons Learned**

1. How to work with Structured, Semi-structured and unstructured data
2. How to handle and process structured numeric text data from spreadsheets and databases
3. How to use data obtained through standard data exchange formats such as HTML or XML from web pages or JSON from web-based APIs
4. How to use data obtained by pattern matching from text or log files.

## **Course content**

Week 1 | Data Pipeline and Python Language Basics

Week 2 | Booleans and Dictionaries

Week 3 | Exploring and Transforming Data for Structured Data

Week 4 | Arrays, Functions, and Categorical Summarization

Week 5 | Stacking and Unstacking Data

Week 6 | Semi-structured Data

Week 7 | Mongo Database, JSON From RSS

Week 8 | Processing Twitter and Facebook

Week 9 | Unstructured Data

Week 10 | Network Structures

## **Course Outcomes**

94.80%

## **Final Project: Analysis of Tweeter Data:**

I analyzed tweets based on the "#KarenStrikesAgain" hashtag. I logged into the twitter API and searched for tweets based on the "karen" meme as a search term. I collected raw tweets and placed them into a neat pandas dataframe, then summarized the top 20 tweets based on location and date of tweets. I further performed analysis of the tweets over time, breaking down the tweets by years, months, weeks and hours. I answered questions such as when and where the tweets surged or waned and provided visualizations using pandas plotly to demonstrate this. Also provided sentiment analysis of the tweets which turned out to be mostly negative - only 5% of the KarenStrikesAgain had a positive connotation.

## **Tools and techniques used:**

Python programming, pandas, plotly, matplotlib, seabon,sentiment analysis, text mining with NLTK, parts of speech tagging, working with csv, data prep before analysis.

[Click here to view course portfolio](https://github.com/soomondi/syracuse/tree/main/IST%20652%20-%20Scripting%20for%20Data%20Analysis/Final%20Project)

**IST 664 Natural Language Processing**

## **Lessons Learned**

1. How to use the NLTK library for semantic and linguistic computational interpretation of text
2. Understanding and working with text corpus
3. Preprocessing of large text corpus
4. Tokenization using different approaches such as the Sklearn Count vectorizer, NLTK Word Tokenizer, Wordpunct Tokenizer,
5. Using the Regular Expression library of Python
6. Stemming and Lemmatization
7. Feature selection
8. Sentiment/Subjectivity analysis
9. Parts of Speech Tagging
10. Classification with Naive Bayes

## **Course content**

Week 1 | Introduction to NLP and Processing Text Words

Week 2 | Corpus Statistics and Language Modeling

Week 3 | Regular Expressions, Morphology, and Processing Text Files

Week 4 | POS Tagging and Introduction to Machine Learning

Week 5 | Context-Free Grammars (CFG) and Parsing

Week 6 | Semantics

Week 7 | Discourse and Dialogue

Week 8 | Sentiment Analysis

Week 9 | NLP Applications: Information Extraction, Machine Translation, and Summarization

Week 10 | NLP Applications: Information Retrieval, Question Answering, and Conversational Agents

## **Final Project: Kaggle Movie Review:**

I conducted an analysis of Kaggle Movie Review Data based on reviews from rotten tomatoes website. Deployed a training data consisting of 156K phrases for analysis.

## **Tools and techniques used:**

NLTK, ScikitLearn, Classification, Model training, feature engineering, visualization, Sentiment/Subjectivity Analysis, Tokenization, Lemmatization, Stemming, Regular Expression, POS tagging.

[Click here to view the course portfolio](https://github.com/soomondi/syracuse/tree/main/IST%20664%20-%20Natural%20Language%20Processing/Final%20Project)

**CONTACTS:**

**LinkedIn:** <https://www.linkedin.com/in/obongoomondi/>

**Instagram**: <https://www.instagram.com/obongo.omondi/>

**Phone:** +1 (251) 463-0351

**Email:** [obongoomondi@outlook.com](mailto:obongoomondi@outlook.com)