Introducing Data Science Topics to Non-computing Majors

Xumin Liu and Erik Golen
Rochester Institute of Technology

This work is supported by the National Science Foundation under Award 2021287.
Agenda

- Presenter and participant introductions
- About our NSF funded project
- Introduction to data science
- Get started with the data
- Data exploration and visualization
- Data Analytics models
- Teaching DS to non-computing majors
Introductions

• Your name, position, and affiliation
• Why did you decide to attend this workshop?
Agenda

- Presenter and participant introductions
- About our NSF funded project
- Introduction to data science
- Get started with the data
- Data exploration and visualization
- Data analytics models
- Teaching DS to non-computing majors
About our NSF funded project

- DUE # 2021287: Developing a Hands-on Data Science Curriculum for Non-Computing Majors (2020-2023)

- Project website: [https://cs.rit.edu/~xl/IUSE.html](https://cs.rit.edu/~xl/IUSE.html)
Our teaching experience:

• ISCH-110 (Principles of Computing)
  • Fall, Spring 2021-2022
  • Duration for the course module: 1 week (3 hours)

• ISCH-370 (Principles of Data Science)
  • Fall, Spring 2021-2022
  • Duration: 14 weeks

• Both ISCH-110 and ISCH-370 are core course of Principles of Computing Immersion curriculum at RIT, offered to non-computing majors
What we will share:

• Teaching materials:
  • Lecture slides
  • In-class demos (Python coding and DSLP exercises)
  • Python coding assignments (with keys)
  • DSLP assignments (with keys)
  • Course project description
  • Quizzes/Exams (with keys)

• Access to DSLP:
  • Register accounts (students and instructors) on the current server
  • Host and manage the platform at your institution
Agenda

▪ Presenter and participant introductions
▪ About our NSF funded project
▪ Introduction to data science
▪ Get started with the data
▪ Data exploration and visualization
▪ Data analytics models
▪ Teaching DS to non-computing majors
What is data science?

- General definition
  - **Data science** is an interdisciplinary field that uses **scientific** methods, processes, algorithms and systems to extract **knowledge and insights** from structured and unstructured **data**, and apply knowledge and actionable insights from **data** across a broad range of application domains.
Data Science Venn Diagram
Data to Knowledge Pipeline

Purpose: We want to **automatically** analyze and **extract knowledge** from **large volumes of data** to create **data products**.

Applied to a specific *domain*. To perform data mining, domain expertise is not necessarily required by the analyst.
Data Product Example

- User profile, View history, Movie info. --> Netflix Movie Recommendation
  - Netflix offers over 7K TV shows and movies
  - Personalized recommendations produce $1 billion a year in value from customer retention
  - Majority of Netflix users choose 80% of the recommendations made by the system
Agenda

- Presenter and participant introductions
- About our NSF funded project
- Introduction to data science
- Get started with the data
- Data exploration and visualization
- Data analytics models
- Teaching DS to non-computing majors
Getting Started with the Data

- Load the data
- Review the data
- Clean the data
Load the data

- How the data is stored
  - Files (locally or online)
    - Text files (.txt)
    - Formatted files (.csv, .xlsx)
    - PDF (.pdf)
    - ...
  - Databases
    - Sql (mySQL, Oracle, …)
    - noSQL (MongoDB, Hbase)
  - Web
    - html, xml, json,…
Loading existing datasets in the DSLP

- You can also upload your own data to the DSLP
## Titanic dataset

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Definition</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PassengerID</strong></td>
<td>Row number of passenger</td>
<td>Numbered from 1 to 891</td>
</tr>
</tbody>
</table>
| **Survived** | Did the passenger survive the disaster or not? | 0 = No  
1 = Yes |
| **Pclass** | Class of passenger ticket. Indicator of the passenger’s socio-economic status | 1 = Upper  
2 = Middle  
3 = Lower |
| **Name** | Name of the passenger! | |
| **Sex** | Gender of the passenger | “male” or “female” |
| **Age** | Passenger’s age (in years) at the time of the disaster | Age is fractional if less than 1. If estimated, in the form xx.5. |
| **SibSp** | Total siblings and spouses aboard the ship | Sibling = brother, sister, stepbrother, stepsister  
Spouse = husband, wife (mistresses and fiancés were ignored) |
| **Parch** | Total parents and children aboard the ship | Parent = mother, father  
Child = daughter, son, stepdaughter, stepson  
Some children travelled only with a nanny; Parch=0 for those cases |
| **Ticket** | Ticket number | |
| **Fare** | Amount of money spent on the passenger ticket | Unit is likely English pounds, but this is unclear |
| **Cabin** | Cabin number | Listed alphanumerically. The format of this data is inconsistent. Many missing values. |
| **Embarked** | Indicates the port where the passenger embarked on the ship | C = Cherbourg  
Q = Queenstown  
S = Southampton |
View the data via DSLP

• The query module - Demo
Clean the data

• Detect and correct corrupt of inaccurate data
  • Missing data
  • Duplicated data
  • Irregular data: outliers

• Important step of a data science pipeline
  • GIGO: Garbage In, Garbage Out.
Missing data

- A common data quality issue,
- Can be caused by collection error or the information is not available.
- Most machine learning models do not accept missing data
- Can be easily detected and corrected
- How to handle?
  - Eliminate null values (remove the records/columns)
  - Fill in the blank
Dealing with missing data
Removing outliers

- Define outliers:
  - Specify the ranges using simple univariate statistics such as standard deviation and interquartile range

- Remove outliers:
Activity 1

• Exercise questions:
  1. How many passengers are missing values for Age?
  2. How many passengers remain in the cleaned dataset after removing N/A rows?
  3. Remove outliers where a data record is considered as an outlier if the age value is outside of the 10th to 90th percentile range. How many passengers remain in the dataset?
Activity 1 - Answers

- Exercise questions:
  1. How many passengers are missing values for Age? 177
  2. How many passengers remain in the cleaned dataset after removing N/A rows? 181
  3. Remove outliers where a data record is considered as an outlier if the age value is outside of the 10th to 90th percentile range. How many passengers remain in the dataset? 145
Agenda

- Presenter and participant introductions
- About our NSF funded project
- Introduction to data science
- Get started with the data
- Data exploration and visualization
- Data analytics models
- Teaching DS to non-computing majors
Data Exploration and Visualization

- Preliminary investigation of the data to better understand its characteristics
- Helps select the proper data preprocessing and analysis techniques
- Use human’s abilities to recognize patterns in data
- Techniques:
  - Summary Statistics
  - Visualization
Summary statistics

• Frequency and Mode - categorical data
• Percentiles - Continuous data
• Measure of Location – continuous data
  • Mean, Median, Trimmed mean
• Measure of Spread – continuous data
  • Range, Variance, Standard Deviation, Interquartile Range (IQR)
Classic Visualization Example

- Sea Surface Temperature for July 1982
  - Tens of thousands of data points are summarized in a single image
Common Visualizations

- **Line Graph**: Displays the evolution of one or more numeric attribute values
- **Scatter Plot**: Displays the correlation relationship between a pair of attributes
- **Histogram**: Shows the distribution of values of a single variable
- **Box Plot**: Shows the distribution of values of a single numerical attribute
Line Graph

- Displays the evolution of one or more numeric attribute values
- Need to determine what x-axis and y-axis represents
  - By default, x-axis is the index
  - You can have multiple columns for y-axis
Scatter Plot

- Used for continuous values
- Attribute values determine the position (x,y)
- Shows the correlation
Histogram

- Shows the distribution of values of a single variable
- Divide the possible values into bins and shows the number of objects in each bin
  - For categorical data, each value is a bin
  - For continuous data, the range of values is evenly distributed into bins
Box Plots

- Shows the distribution of the values of a single numerical attribute
- Provides a simple graphical depiction of interquartile range and outliers
Activity 2

• Exercise questions:
  1. How many columns are numeric and how many are categorical?
  2. How many passengers embarked the Titanic at Southampton? (hint: click on “Embarked” tag to check)
  3. In the 5-bin histogram, which bin contains the most people?
  4. What is the approximate range of ages in the bin with the most people in the 5-bin histogram?
  5. Which group (survived or perished) of the passengers have averagely higher fare?
  6. Explain the relationship between fare and passenger class and why this makes sense or not.
Activity 2 - Answers

• Exercise questions:
  1. How many columns are numerical and how many are categorical? 7 are numerical 5 are categorical.
  2. How many passengers embarked the Titanic at Southampton? (hint: click on “Embarked” tag to check) 644
  3. In the 5-bin histogram, which bin contains the most people? 2
  4. What is the approximate range of ages in the bin with the most people in the 5-bin histogram? 18-32
  5. Which group (survived or perished) of the passengers have averagely higher fare? Survived
  6. Explain the relationship between fare and passenger class and why this makes sense or not.
Agenda

- Presenter and participant introductions
- About our NSF funded project
- Introduction to data science
- Get started with the data
- Data exploration and visualization
- Data analytics models
- Teaching DS to non-computing majors
Data Analytics Models

- Overview of Machine Learning Models
- Data Classification
Overview of Machine Learning Models

• Suppose we observe $Y_i$ and $X_i = (X_{i1}, ..., X_{ip})$ for $i = 1, ..., n$
  • $X$: a set of attributes/features,
  • $Y$: a targeted variable (to be predicted)
  • $n$: the number of observations
• There may be a relationship between $Y$ and at least one of the $X$’s.
• We can model the relationship as

$$Y_i = f(X_i) + \varepsilon_i$$

  - an unknown function
  - is a random error with mean zero

• Later on, we can use $X$ to predict $Y$ for an unseen records
  • E.g., if a passenger can survive, if an email is a spam, what is the customer’s rating…
Supervised vs. Unsupervised Learning

- **Supervised Learning:**
  - Both the predictors, $X_i$, and the response, $Y_i$, are observed
    - Example: predict if an email is spam or not
  - Models: classification, regression

- **Unsupervised Learning:**
  - Only the $X_i$’s are observed.
    - Example: market segmentation where we try to divide potential customers into groups based on their characteristics.
  - Models: clustering, association rule discovery
Supervised learning

**Induction**: go from specific instances to general principles (model).
**Deduction**: go from general principles (model) to a specific conclusion.
## Hypothetical Training Data for Disease Diagnosis

<table>
<thead>
<tr>
<th>Patient ID#</th>
<th>Sore Throat</th>
<th>Fever</th>
<th>Swollen Glands</th>
<th>Congestion</th>
<th>Headache</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Strep throat</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Allergy</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Cold</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Strep throat</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Cold</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Allergy</td>
</tr>
<tr>
<td>7</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Strep throat</td>
</tr>
<tr>
<td>8</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Allergy</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Cold</td>
</tr>
<tr>
<td>10</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Cold</td>
</tr>
</tbody>
</table>
Possible Data Model

```
Swollen Glands
  No
  Yes
    Diagnosis = Strep Throat

Fever
  No
  Yes
    Diagnosis = Allergy
    Diagnosis = Cold
```
# Classification – Confusion Matrix & Accuracy

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class=Yes</td>
<td>Class=Yes</td>
</tr>
<tr>
<td>Class=Yes</td>
<td>a (True Positive)</td>
</tr>
<tr>
<td>Class=No</td>
<td>c (False Positive)</td>
</tr>
</tbody>
</table>

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$

e.g., The percentage of diagnoses that are predicted correctly
Activity 3

• Exercise questions:
  1. Using the generated model, what is the predicted Survived value for a passenger record where Age=40, Sex=0, Fare=100.
  2. How many times did the classifier model correctly predict that a passenger would survive?
  3. What is the overall accuracy of the model?
Activity 3 - Answers

- Exercise questions:
  1. Using the generated model, what is the predicted Survived value for a passenger record where Age=40, Sex=0, Fare=100. 0
  2. How many times did the classifier model correctly predict that a passenger would survive? 24
  3. What is the overall accuracy of the model? \(\frac{7 + 24}{37} = 81\%\).
Agenda

- Presenter and participant introductions
- About our NSF funded project
- Introduction to data science
- Get started with the data
- Data exploration and visualization
- Data analytics models
- Teaching DS to non-computing majors
## Student Background - Major

### Majors

<table>
<thead>
<tr>
<th>Applied mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>University Exploration</td>
</tr>
<tr>
<td>Statistics &amp; Actuarial Science</td>
</tr>
<tr>
<td>Electrical Engineering</td>
</tr>
<tr>
<td>Computing and Information Technology</td>
</tr>
<tr>
<td>Applied Stats</td>
</tr>
<tr>
<td>Applied Statistics and actuarial Sciences</td>
</tr>
<tr>
<td>Applied Statistics and Data Analytics</td>
</tr>
<tr>
<td>3D Digital Design</td>
</tr>
</tbody>
</table>

ISCH-110 (Intro to Computing Course)

### Majors

<table>
<thead>
<tr>
<th>School of Individualized Study (SOIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film Production</td>
</tr>
<tr>
<td>Communication</td>
</tr>
<tr>
<td>Management Information Systems</td>
</tr>
<tr>
<td>Accounting</td>
</tr>
<tr>
<td>Management Information Systems</td>
</tr>
<tr>
<td>History (required course for these students)</td>
</tr>
</tbody>
</table>

ISCH-370 (Principles of Data Science course)
Student Background - Programming

What are your past experiences with programming? Mark all that apply.

ISCH-110

ISCH-370
Survey result – Data Science Module - Level I

The Data Science Module improved my understanding of data science.
Survey result – Data Science Module - Level I

Using the DSLP improved my confidence in conducting data science inquiries and analytical tasks
Survey result – Data Science Module Level II

1. I felt the lectures were easy to follow.
2. The lectures improved my understanding of data science.
3. The lectures increased my interest in Data Science.
4. I believe I could perform similar data science tasks to those in the lectures.

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree
Survey result – Python coding vs DSLP

Google Colab assignments

DSLP assignments
Key Takeaways

- Students from various disciplines are interested in DS topics.
- Students are enthusiastic about choosing their own datasets for their projects but could be frustrated with their coding capabilities.
- Non-computing major students appear to learn best from hands-on exercises with minimal or no coding components.
Thank you!

- Please contact Xumin Liu (xmlics@rit.edu) if you would like to have access to the teaching materials and/or the DSLP.