Programming for Data Science

CSC310
Course Details

- **Dr Lutz Hamel**
- Best way to get in touch - email:  
  - lutzhamel@uri.edu
- Everything is online  
  - Assignments & Gradebook & Syllabus  
    - BrightSpace
  - Lecture Notes  
    - https://lutzhamel.github.io/CSC310/
  - Book  
    - Python Data Science Handbook
What is Data Science?

Data science is the discipline of the extraction of knowledge from data.

It relies on

- computer science
  - for AI, data structures, algorithms, visualization, big data support, and general programming
- statistics/mathematics
  - for data models and inference
- domain expertise
  - for asking questions and interpreting results

What is Data Science?

- Data science is the discipline of the extraction of knowledge from data.

How do we do that?

- We build MODELS of data!
Models: Play Tennis

Lots of data - very little information!

Build a model - a decision tree!
Models: Play Tennis

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>

ID3 Decision Tree

This model summarizes the whole table correctly!
What is Data Science?

- Data science is the discipline of the extraction of knowledge from data.

Where does the data come from?

- The data pipeline!

The Data Pipeline
What is Data Science?

* Data science is the discipline of the extraction of knowledge from data.

How do we preprocess our data for model building?

- Statistics!
  - Descriptive Statistics
  - Missing Value Processing
  - Normalization

Descriptive vs. Inferential Statistics

**Purpose**: Descriptive statistics aim to summarize data, while inferential statistics aim to make predictions or generalizations about a population from a sample.

**Data coverage**: Descriptive statistics deal with the entire dataset, whereas inferential statistics focus on samples from which to generalize about a population.

**Analysis outcome**: The outcome of descriptive statistics is a summary of data, while the outcome of inferential statistics is predictions, decisions, or inferences about population parameters.

**In summary**, descriptive statistics help describe, show, or summarize data in a meaningful way, allowing the data to be visualized easily, whereas inferential statistics take data from a sample and make inferences or predictions about a population.
What is Data Science?

Data science is the discipline of the extraction of knowledge from data.

How do we ask the right questions?

Domain Expertise!

Knowledge cannot be generated in a vacuum. You need the context of a domain in order to generate new insights. E.g. bioinformatics, climate modeling, sales forecasting, etc.