# Data Science Short-Course – Fact Sheet

**Number of participants:** Any

**Daily schedule:** Typically begins in the 8:30-9:30 AM range and ends in the 5:00-5:30 PM range, with a lunch hour and short morning and afternoon breaks

**Classroom setup:** Desks or theater style

**A/V requirements:** Projection for Prof. Widom's Macbook laptop; screen visible by all students; audio capability helpful for module V1 but not required

**Internet:** Required for all modules except {O1, M1, C1, L2, L3}

**Students background:** Students should be generally comfortable with logic and basic mathematics; for modules {D1, D2, P1, R1, M2, M3, P2, N1, U1} some computer programming background is expected (equivalent to one secondary school or college introductory course in any programming language).

> ***Important note:*** *Some modules may be too basic for advanced undergraduate or post-graduate computer science or other technical students who have already taken courses in data management, data mining, or machine learning. All modules are suitable for younger technical students, or for students in non-technical disciplines with background as specified above. The material is also suitable for lecturers and faculty in both technical and non-technical disciplines. We recommend using the course as a vehicle for bringing together students and others from a variety of disciplines with a shared interest in data.*

**Student hardware:** For all modules except {O1, M1, C1, L2, L3}, one desktop or laptop per 1-3 students, must be connected to the internet

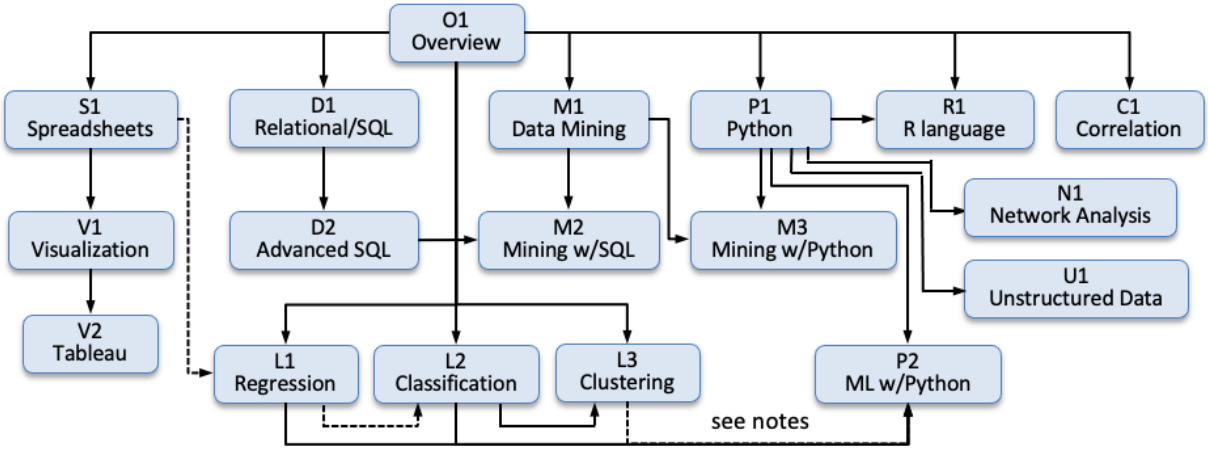
**Software preparation in advance of course (separate instructions are provided):**

- Google Drive (via Google account) required for modules {S1, V1, L1}
- Tableau Public required for module V2
- Google Colab (via Google account) required for modules {D1, D2, P1, R1, M2, M3, P2, N1, U1}
- For all modules except {O1, M1, C1, L2, L3}: comfortable downloading/uploading and manipulating files on local computer

**Course assistants:** For all modules involving assigned work (all modules except {O1, M1, C1, L2, L3}) it can be helpful to provide 1-2 course assistants to help students who are having difficulty. Assistants can be faculty, lecturers, post-graduate, or advanced undergraduate students in computer science or related fields – anyone strong in English and comfortable with the concepts and tools (or a quick learner).

*Course modules flowchart and content details on following pages*

## Course Modules




## Course Content

**Module O1: Overview of Data Science - Promises and Pitfalls, Tools and Techniques**

- *Prerequisites* - none
- *Topics* - Data-driven applications and services; brief introduction to data manipulation and analysis, data mining, machine learning, data visualization, data collection and preparation; pitfalls: correlation and causation, underfitting and overfitting, privacy, and others; brief introduction to languages, systems, and platforms for working with data
- *Length* – 1.5-2 hours
- *Style* - Prof. Widom presentation with audience Q&A

**Module S1: Data Analysis Using Spreadsheets**

- *Prerequisites* - O1
- *Topics* - Manipulating and analyzing data using spreadsheets including pivot tables
- *Length* – 2-2.5 hours
- *Style* - Students work along with Prof. Widom and work on assigned problems
- *Software* - Google Sheets

**Module V1: Data Visualization Using Spreadsheets**

- *Prerequisites* - S1
- *Topics* - Data visualization motivation; spreadsheet bar charts, pie charts, scatterplots, maps
- *Length* – 1.5-2 hours
- *Style* - Students work along with Prof. Widom and work on assigned problems
- *Software* - Google Sheets

**Module V2: Advanced Data Visualization Using Tableau**

- *Prerequisites* - V1
- *Topics* - Tableau bar charts, pie charts, scatterplots, packed bubbles, maps; Tableau dashboards; publishing interactive visualizations
- *Length* – 1-2 hours
- *Style* - Students work along with Prof. Widom and work on assigned problems
- *Software* - Tableau Public

**Module D1: Relational Databases and Basic SQL**

- *Prerequisites* - O1
- *Topics* - Introduction to relational database management systems (RDBMS); relational data model; creating and loading data; basics of SQL query language
- *Length* – 1.5-2 hours
- *Style* - Prof. Widom presentation interleaved with students working on assigned problems
- *Software* - SQLite relational database system via Google Colab

**Module D2: Advanced SQL**

- *Prerequisites* - D1
- *Topics* - More advanced SQL constructs (aggregation, subqueries, data modification, and others); coverage configurable to available time
- *Length* - 1-2 hours
- *Style* - Prof. Widom presentation interleaved with students working on assigned problems
- *Software* - SQLite relational database system via Google Colab

**Module P1: Python for Data Analysis and Visualization**

- *Prerequisites* - O1
- *Topics* - Introduction to Python; manipulating data in Python; plotting in Python; Pandas package
- *Length* - 3-4 hours
- *Style* - Prof. Widom presentation interleaved with students working on assigned problems
- *Software* - Python via Google Colab

**Module M1: Data Mining Algorithms**

- *Prerequisites* - O1
- *Topics* - History of data mining; market-basket data; frequent item-sets; association rules
- *Length* - 1 hour
- *Style* - Prof. Widom presentation with audience Q&A

**Module M2: Data Mining Using SQL**

- *Prerequisites* - M1, D2
- Topics - Computing frequent item-sets and association rules using relational databases and SQL
- *Length* - 1-2 hours
- *Style* - Prof. Widom presentation and students work on assigned problems
- *Software* - SQLite relational database system via Google Colab

**Module M3: Data Mining Using Python**

- *Prerequisites* - M1, P1
- *Topics* - Computing frequent item-sets and association rules using Python
- *Length* – 1.5-2 hours
- *Style* - Prof. Widom presentation and students work on assigned problems
- *Software* - Python via Google Colab

**Module L1: Machine Learning - Regression**

- *Prerequisites* - O1 required, S1 recommended but not required
- *Topics* - Regression introduction and applications; simple linear regression; regression and correlation; regression shortcomings and dangers; polynomial regression
- *Length* – 1.5-2 hours
- *Style* - Prof. Widom presentation and students work on assigned problems
- *Software* - Google Sheets

**Module L2: Machine Learning - Classification**

- *Prerequisites* - O1, L1 recommended but not required
- *Topics* - Introduction to classification; k-nearest-neighbors; decision trees; Naïve Bayes classifiers
- *Length* – 1 hour
- *Style* - Prof. Widom presentation with audience Q&A

**Module L3: Machine Learning - Clustering**

- *Prerequisites* - O1, L2
- *Topics* - Introduction to clustering; k-means
- *Length* - Less than 1 hour
- *Style* - Prof. Widom presentation with audience Q&A

**Module P2: Machine Learning Using Python**

- *Prerequisites* - P1, one or more of {L1, L2, L3}
- *Topics* - Python packages for regression, classification, and clustering
- *Length* - 1-2 hours depending on coverage
- *Style* - Prof. Widom presentation and students work on assigned problems
- *Software* - Python via Google Colab

**Module R1: The R Language**

- *Prerequisites* - O1, P1
- *Topics* - Manipulating data in R; plotting in R
- *Length* - 1-2 hours
- *Style* - Prof. Widom presentation interleaved with students working on assigned problems
- *Software* - R via Google Colab

**Module C1: Correlation and Causation**

- *Prerequisites* - O1
- *Topics* - correlation versus causation; determining correlation; determining causation
- *Length* - Less than 1 hour
- *Style* - Prof. Widom presentation with audience Q&A

**Module N1: Network Analysis**

- *Prerequisites* - P1
- *Topics* – Modeling networks as undirected and directed graphs; analyzing graph properties; programming using networkx package
- *Length* – 1.5-2 hours
- *Style* - Prof. Widom presentation and students work on assigned problems
- *Software* - Python via Google Colab

**Module U1: Unstructured Data**

- *Prerequisites* - P1
- *Topics* – Text analysis & natural-language processing; image analysis
- *Length* – 1.5-2 hours
- *Style* - Prof. Widom presentation and students work on assigned problems
- *Software* - Python via Google Colab