

Data 301/COSC 301/Data 501

Introduction to Data Analytics

Course Introduction

Dr. Irene Vrbik

University of British Columbia Okanagan
irene.vrbik@ubc.ca

Instructor

- ▶ **Instructor:** Irene Vrbik, Ph.D.
- ▶ **Office:** SCI 393
- ▶ **e-mail:** ivrbik@mail.ubc.ca or irene.vrbik@ubc.ca
- ▶ **Office Hours:** Friday 2:00 – 3:00 PM

Who am I?

- ▶ Academic Career:
 - ▶ Undergrad 2009 (McMaster University),
 - ▶ Masters 2010 (University of Guelph),
 - ▶ PhD 2014 (University of Guelph),
 - ▶ Postdoc 2014 (McGill University),
 - ▶ Natural Sciences and Engineering Research Council of Canada (NSERC) postdoctoral fellowship 2016 (UBCO),
 - ▶ Instructor, 2018 (UBCO)
- ▶ Have taught statistics/data science/probability at the University of Guelph, McGill, and UBCO ranging from undergrad to masters of data science.

The Essence of the Course

The overall goal of this course is for you to understand data analytics and be able to apply data analysis to data sets using a variety of software tools and techniques.



The Essence of the Course

- ▶ The most exciting aspect of data analytics is discovering and presenting useful data/information that can have an impact on business, society, etc.
- ▶ This course will provide the tools and skills for you to perform your own data analysis when encountering problems in the real-world.
- ▶ As an introductory course, the goal is to get exposure to the skills and techniques as there will not be time for mastery.

Official Calendar Description

Official Calendar: Techniques for computation, analysis, and visualization of data using software. Manipulation of small and large data sets. Automation using scripting. Real-world applications from life sciences, physical sciences, economics, engineering, or psychology. No prior computing background is required. Credit will be granted for only one of COSC 301, DATA 301 or DATA 501.[3-2-0]

Prerequisite: Either (a) third-year standing, or (b) one of COSC 111 (Computer Programming I) or COSC 122 (Computer Fluency).

Please familiarize yourself with the details outlined on the course syllabus (posted on Canvas).

Specific Description

Specific description: This course provides an introduction to data analytics to train students with practical industrial techniques for data manipulation, analysis, reporting, and visualization.

This is not an introduction to programming.

- ▶ Programming techniques will be taught to automate data analysis.
- ▶ Introduction to programming courses are COSC 111 or COSC 123.
- ▶ Prior computing experience is not required, but is helpful (for instance COSC 122 or COSC 111).

Course Objectives

1. Understand data representation formats and techniques and how to use them.
2. Experience a wide-range of data analytics tools include Excel, SQL databases, R, and visualization.
3. Develop a computational thinking approach to problem solving and use programs and scripting to solve data tasks.
4. Apply techniques to representative problems from the real word.

Toolkit

Our **toolkit** refers to the tools, software, and techniques that you can use today to solve your problems.

Some of the things you will be adding to your toolkit include:

- ▶ Excel and Excel VBA
- ▶ SQL/databases
- ▶ command line
- ▶ Python and Python libraries for data analysis/visualization,
- ▶ R and R libraries for data analysis/visualization
- ▶ Tableau

Skills and Capabilities

While these tools have many **capabilities** (of which we will only see a small subset) the **skills** we will cover in this course will be the building blocks of future learning.

Skills include:

- ▶ programming concepts (Python & R)
- ▶ data representation/metadata
- ▶ thinking algorithmically, designing, manipulating and cleaning data
- ▶ querying and filtering data
- ▶ statistical analysis
- ▶ visualizing information

My Course Goals

1. Provide the information in a simple, concise, and effective way for learning.
2. Strive for *all* students to understand the material and pass the course.
3. Be available for questions during class time, office hours, and at other times as needed.
4. Provide an introduction to data analytics tools and techniques so that students are able to apply data analysis to their own data sets.
5. Encourage students to continue with other data analytics/statistics/computer science courses.

Course Format

- ▶ Each lecture will be posted on [Canvas](#).
- ▶ I encourage you to print them out/follow along with them on your computer.
- ▶ If possible, it will be beneficial that you download the necessary programs prior to lecture so that you can follow along with examples on your laptop.

Email

I check email twice a day, once in the morning and once in the late afternoon. Emails will be answered in order of importance and when they were received. In order to ensure a reasonably prompt

response all Emails should use the following format:

- ▶ Subject line must begin with the course subject, eg. DATA 301 (or COSC 301 or DATA 501)
- ▶ The first line of the email you will state your name, student number and the course code as stated in the subject line.

Before e-mailing me, I encourage you to ask questions on Canvas' Discussion Board.

Labs

- ▶ Labs are held weekly starting next week.
- ▶ Please check your registration to determine your lab section and time.
- ▶ TAs will be in the lab each week to help with labs assignments and to provide help with general course material questions.
- ▶ While labs are not mandatory, you are highly encouraged to attend.
- ▶ You must be enrolled in a lab and you must only attend the lab you are enrolled in.
- ▶ While you will have access to University computers in the lab (all of the needed software should already be installed) feel free to bring in your laptops instead.

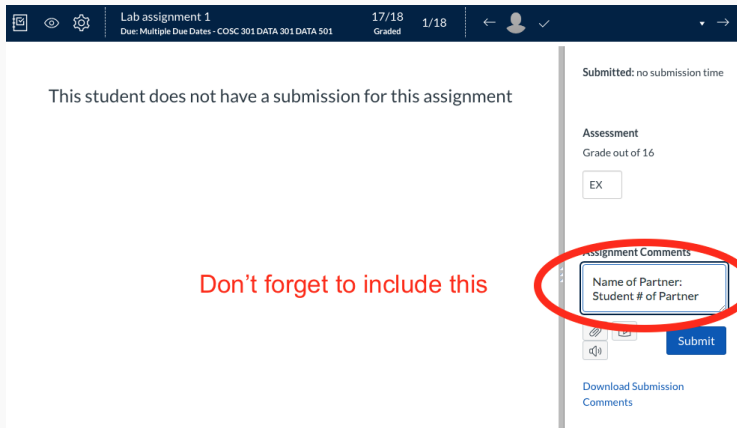
Assignments

- ▶ Lab assignments are worth 30%¹ of your overall grade.
- ▶ Lab assignments may take more than the scheduled lab time.
- ▶ Late assignments will have 10% deducted for each day (which includes weekends) beyond the due date. Assignments that are more than 5 days overdue will **not** be accepted.
- ▶ An assignment may be handed in any time before the due date.
- ▶ Assignments are critical to learning the material and are designed to prepare you for the exams and build up your skills!
- ▶ Please attend lab to go over any errors you made with your TA. Assignment solutions will *not* be posted.

¹if enrolled in Data 301/Cosc 301

Assignments

- ▶ Lab assignments are done individually or in groups of two depending on the assignment.
- ▶ You are not bound to your partner and may choose to go solo on some or all of the subsequent assignments.
- ▶ Your partner need not be in the same lab section as you
- ▶ If you are working with a partner, please decide on who will submit the documents to Canvas (only one person should upload these files).
- ▶ Regardless if you were the partner who uploaded the completed assignment, please include the name and student number of your partner in the **Assignment Comments**



The image shows a screenshot of the Canvas LMS interface for an assignment submission. The top navigation bar is dark blue with icons for a document, eye, and settings on the left. The center of the bar displays 'Lab assignment 1', 'Due: Multiple Due Dates - COSC 301 DATA 301 DATA 501', '17/18', '1/18', and 'Graded'. On the right are navigation arrows and a user profile icon. The main content area has a light gray background with the text 'This student does not have a submission for this assignment' in the center. To the right is a submission sidebar. It includes a 'Submitted' status of 'no submission time', an 'Assessment' section with 'Grade out of 16' and an 'EX' button, and an 'Assignment Comments' section. The comments section contains a text box with the placeholder text 'Name of Partner:' and 'Student # of Partner'. Below the text box are icons for attachments, a microphone, and a 'Submit' button. At the bottom of the sidebar are links for 'Download Submission' and 'Comments'. A red oval highlights the 'Assignment Comments' section, and a red text overlay 'Don't forget to include this' points to it.

Lab assignment 1
Due: Multiple Due Dates - COSC 301 DATA 301 DATA 501

17/18
1/18
Graded

This student does not have a submission for this assignment

Submitted: no submission time

Assessment
Grade out of 16

EX

Assignment Comments

Name of Partner:
Student # of Partner

Submit

Download Submission
Comments

Don't forget to include this

Figure: Each partner should submit the name and student number of their partner in the assignment comments on Canvas. **Only one** partner should upload the files of the complete assignment.

Evaluation

DATA 301/COSC 301

| | | |
|--------------|-----|-----------------------|
| Clickers | 5% | (in class) |
| Assignments | 30% | (weekly-ish) |
| Two Midterms | 30% | (in class) |
| Final Exam | 35% | (cumulative, 3 hours) |

DATA 501 Graduate Student Evaluation:

| | | |
|--------------|-----|----------------------------------|
| Project | 30% | (Details to be posted on Canvas) |
| Two Midterms | 30% | (in class) |
| Final Exam | 40% | (cumulative, 3 hours) |

Midterms

As stated on the TentativeSchedule (posted on [Canvas](#)), midterms are held in class on:

Midterm 1 Tuesday, Feb 25, 2020

Midterm 2 Thursday, March 26, 2020

Clause

A student must pass the final exam, or receive an average grade of at least 50% on the exams (midterms and final) to pass the course. Otherwise, the student will be assigned a maximum overall grade of 45. See table for examples:

| Midterm 1 | Midterm 2 | Final | Average Grade | Satisfy Clause |
|-----------|-----------|-------|---------------|----------------|
| 70% | 35% | 45% | 50% | Yes |
| 40% | 35% | 50% | 42% | Yes |
| 45% | 51% | 45% | 47% | No |

Grad students Final Project

- ▶ There is no clicker, lab or assignment component for students enrolled in Data 501 (although it is recommended that you work through these assignments for practice).
- ▶ Graduate students enrolled in DATA 501 will be expected to complete a Final Project
- ▶ This includes a hand-in written proposal, final paper, and in-class presentation (final week of class)
- ▶ Pay attention on Canvas for postings on checkpoint deadlines, rubrics, etc.

Academic Dishonesty

Cheating in all its forms is strictly prohibited and will be taken very seriously by the instructor.

Assignments

You are expected to submit original work done by you and acknowledge all sources of information or ideas while attributing them to others as required.

Exams

All exams are closed book, so no course materials should be present.

Academic Misconduct

Students are responsible for informing themselves of the guidelines of acceptable and unacceptable conduct for graded assignments established by their instructors for specific courses, and of the examples of academic misconduct set out [here](#). Some main concerns include:

- ▶ Plagiarism, which is intellectual theft, occurs when an individual submits or presents the oral or written work of another person as his or her own.
- ▶ Submitting the same, or substantially the same, essay, presentation, or assignment more than once (whether the earlier submission was at this or another institution) unless prior approval has been obtained from the instructor(s) to whom the assignment is to be submitted.

Disciplinary Measures

Academic misconduct that is subject to [disciplinary measures](#)

Academic misconduct often results in a one-year suspension from the University and a notation of academic discipline on the student's record. However, disciplinary measures which may be imposed, singly or in combination, for academic misconduct include, but are not limited to, the following:

- ▶ a failing grade or mark of zero on the assignment or in the course in which the academic misconduct occurred

Note that all incidents of suspected academic misconduct must be reported to the Dean's Office.

Academic Dishonesty

Don't cheat!

Do not cheat, copy, or mislead others about what is your work.

- ▶ Violations of academic integrity (i.e., misconduct) may result in a mark of zero on the assignment or exam and more serious consequences may apply if the matter is referred to the President's Advisory Committee on Student Discipline.
- ▶ Careful records are kept in order to monitor and prevent recurrences. **If you have any questions about how academic integrity applies to this course, please consult with your professor.**

How to Pass This Course

The most important things to do to pass this course:

Attend class and labs

- ▶ Follow along and annotate the lecture notes in class.
- ▶ Participate in class exercises and questions.
- ▶ Labs are for marks and are practice to learn the material for the exams.

To get an "A" in this course do all the above plus:

Practice on your own. Practice makes perfect.

Do more questions than in the labs. Try the techniques on your own data sets.

Disclaimer

There is a wide variety in previous experience.

- ▶ Some material you may already know. Help others!!
- ▶ Build up your computer experience in labs and outside of class.
- ▶ Third-year standing means that you know how to “figure things out.”

The course will be very straightforward: Do the work and practice the techniques to do well.

Systems and Tools

All required software is available on the lab computers in SCI 234. If you have your own personal laptop, it is advised that you download the required software prior to class so that you can follow along with examples in lecture.

- ▶ Our first unit uses Excel. A student download of Microsoft office is available [here](#)

Canvas will be used for accessing lectures, submitting assignments, posting marks, discussions, and more, . . .

- ▶ Support for students is available [here](#)

The In-Class Quizzes

- ▶ To encourage attendance and effort, 5% of your overall grade is allocated to answering in-class clicker questions.
- ▶ These questions are answered electronically using a clicker.
- ▶ The clicker can be purchased at the bookstore.
- ▶ Please refer to the **ClickerOrientation** file uploaded to the **Supplementary Material** Canvas on how to use them.

Clickers

- ▶ The clicker is personalized to you with your student number.
- ▶ At different times during all the lectures, questions reviewing material will be asked. Responses are given using clickers.
- ▶ You must answer 80% of the questions correctly to get the full 5%.
- ▶ There are some (not-for-grading) example clicker questions at the end of the lecture (I will take them up next class to ensure everyone has a chance to obtain a clicker first)

Don't forget to register your clickers on Canvas

UBC
Account
Dashboard
Courses
Calendar
Inbox
Help

COSC 301 DATA 301 DATA 501 > COSC 301 DATA 301 DATA 501 Introduction to Data Analytics

2019W2
Home
Announcements
Discussions
Grades
People
Syllabus
Modules
My Media
Media Gallery
Library Online
Course Reserves
Chat
iClicker
Course Evaluation

Register Your iClicker

Please use Chrome browser for the best registration experience.

Enter your 8-character remote ID and other information below...

Remote ID:

E-Mail:

Country:

Register

iClicker Student Registration FAQ

Where do I find my remote ID?

Your iClicker remote ID is printed on a sticker located on the back of your remote. The ID is the 8-character code below the barcode. Newer original iClicker remotes have a secondary ID location behind the battery compartment and iClicker 2 remotes display the ID upon power up. The remote ID will only contain letters A-F and numbers 0-9.

Behind Battery Sticker on Back Sticker on Back Power On Screen

What do I do if my registration fails in Safari?

What do I do if I cannot read the ID printed on my remote?

Reset Student Leave Student View

You are currently logged into Student View

Resetting the test student will clear all history for this student, allowing you to view the course as a brand new student.

An Introduction to Data Analytics

Data Analytics vs Data Analytics

Data analysis

Data analysis is the processing of data to yield useful insights or knowledge.

Data Analytics

Data Analytics is the science of examining raw data with the purpose of drawing conclusions about that information.

Data Analytics vs Data Analytics

The distinction between data analysis and analytics is blurry to say the least (even [Wikipedia](#) is confused).

[One source](#) might say that data analysis is a subcomponent of data analytics, while [another source](#) says data analytics is a subcomponent of data analysis.

I like to think of data analysis as the *method* (ie *action*) whereas data analytics are *tools* used to do so.

Analytics is supported by many tools such as Microsoft Excel, SQL, Python, R, all of which we will talk about in this course.



Other related terms

Other terms that may add confusion to these definitions are . . .

- ▶ *Data Science* is a field comprised of everything relating to data cleansing, preparation, and analysis. This umbrella term that encompasses data analytics, data mining, machine learning, and several other related disciplines.
- ▶ *Big Data* usually refers to immense volumes of data that inhibits the use of traditional data-processing applications.

See [this](#) article for more on the differences between Data Science, Big Data, and Data Analytics.

What is a Data Analyst?

Data analyst

A *data analyst* is a person who uses tools and applications to transform raw data into a form that will be useful.

- ▶ Vital in a “data-driven” world with larger and more critical data sets.
- ▶ The first step in data analysis is often *data collection/munging/processing* which involves finding, loading, cleaning, manipulating, transforming, and visualizing the data.
- ▶ The knowledge may be used for scientific discovery, business decision-making, or a variety of other applications.

Why is Data Analytics important?

Data analytics is important as society is collecting more and larger data sets all the time:

Web All web pages visited and links clicked, searches made, images and posts

Business Items purchased by date, supply chain/customers, industrial sensors

Science Massive data sets (biological/genomic, astronomy, physics, healthcare)

Environmental Sensors and monitors (temperature, etc.)

Why is Data Analytics important?

Transforming this raw data into useful insights has major value:

Web Online advertising driven by understanding customer behaviour; [tailored google searches](#), [Google Analytics](#)

Business [Sales predictions](#), marketing promotions, manufacturing improvement

Science Scientific discoveries, new medical treatments and drugs; see some [examples](#) in healthcare.

Environmental Understanding of environmental processes to allow for changing policies and behaviours, eg [Institute of Environmental Analytics](#)

Why This Course is Important?

- ▶ For many of you, this will be your first exposure to programming and data analytics.
- ▶ Regardless of your discipline, the tools you develop throughout this course will train you to think analytically and creatively.
- ▶ Beyond University, many professional jobs of the future will involve collecting, manipulating, and analyzing data.
- ▶ People who can understand how data can be used will have better employment opportunities.

Why This Course is Important?

Important skills you may learn in this course:

- ▶ Excel Proficiency: for general data analysis and productivity.
- ▶ Databases: Understand how they work and how to use them.
- ▶ Programming and Computational Thinking: Critical thinking and the ability to clearly articulate a problem in a systematic way has applications beyond data analytics.
- ▶ Applied Statistics: Using R and other software makes your statistics training useful for real-world problems.
- ▶ Data visualization: how to display and convey information in a meaningful way.
- ▶ Real-world problem solving: learn to tackle real-world data analysis problems and understand when to use what tool.

Data Analytics Toolkit

- ▶ A data analyst has expertise in programming, statistics, data munging (transformation), and data visualization.
- ▶ In this course, you will be introduced to several tools for gaining competency in each of these skills.
- ▶ As an introductory course, the goal is to get exposure to the skills and techniques as there will not be time for mastery.
- ▶ This toolkit of systems and techniques will be useful in many jobs even if they are not considered data analyst positions.

Why is Data Analytics important?

- ▶ 90 percent of the worlds data have been generated over the last two years [[Forbes Magazine](#)].
- ▶ In 2017, Machine Learning Engineers, Data Scientists, and Big Data Engineers ranked among the top emerging jobs on LinkedIn [[Forbes Magazine](#)].
- ▶ In 2012 was Data Scientist dubbed the sexiest job of the 21st Century by [Harvard Business Review](#).
- ▶ An estimated 2.7 million job postings for Data Analytics and Data Science are projected in the United States by 2020 [[IBM](#)].

Massive Growth of Data - “Big Data”

Data facts from Forbes Magazine, May 21, 2018 [[Forbes](#)]

- ▶ An estimated 2.5 quintillion bytes (2.5 EB) generated per day.
- ▶ Google processes about 3.5 billion requests/day and stores about 10 EB of data.
- ▶ Facebook collects 500 TBs/day (~2.5 billion items) and stores 100+ PB of photos.

See [here](#) on how much data is generated every minute

- ▶ Users watch 4,146,600 videos every minute on YouTube
- ▶ Instagram users post 46,740 photos every minute.

Types of Data Analysis

Descriptive: what are the features of the data?

Exploratory: what new relationships/connections exist in the data?

- ▶ Correlation does not imply causation

Inferential: use data samples to predict about larger population

- ▶ Statistical models: estimate value and uncertainty

Predictive: use data to predict values for other objects

Causal: what variables change values of other variables?

- ▶ Randomized studies: if give a drug/treatment, does it have a positive effect?

Clicker Test

Example 1

Why are you here?

- A) I want to learn more about data analytics.
- B) I know how important data is to my work or future work.
- C) I need an upper-year elective course.
- D) I already have training in computer science/statistics and want to expand my knowledge further.
- E) I want an easy credit.

Clicker Test

Example 2

Which of the following topics are you most interested in?

- A) Excel and SQL Databases
- B) Programming and Python
- C) Data visualization and GIS
- D) R and Applied Statistics
- E) None of the above

Clicker Test

Example 3

What is your major?

- A) Math/Stat/Computer Science/Engineering
- B) Buisness
- C) Science (biology, chemistry, physics, environmental)
- D) Arts
- E) Other

Clicker Test

Example 4

What is your computer background?

- A) I can use computer and mobile applications
- B) I can write a formula in Excel
- C) I can write a simple program in some programming language (eg. Python, R, Java)
- D) I can write a query in SQL
- E) Two or more of the above

Clicker Test

Example 5

What grade are you expecting to get?

- A) A
- B) B
- C) C
- D) D
- E) F