









# **CQADS Data Analysis Workshops**

At CQADS, we know that the ideal decisionmaking environment requires an in-depth understanding of your data.

To help you make the most out of your decisions, we offer a variety of workshops targeted to all learning levels, on topics ranging from data collection to data analysis and visualization.

Selected workshops are offered at Carleton University on a regular basis. On-site sessions can also be arranged. Customizable sessions (topic- and level-specific) containing multiple workshops are available, as are spotlights on specific methods.

For pricing information, availability, and other inquiries, please contact cqads@carleton.ca.

For the workshop schedule, or to join our mailing list, please visit cgads.carleton.ca/workshops.

Theoritical (T) and practical (P) program suggestions are provided for various groups (in hours), ranging from the short and less technical (C1), to the "big picture" survey (C3), to the more technical and demanding (Full).

## I. INTRODUCTION TO DATA SCIENCE

Prelude to Data Science	2
Elements of Data Visualization	2
Mining for Information Gold	3
Text Mining and Natural Language Processing	3

## II. THE DATA SCIENTIST'S TOOLBOX

Supervised Learning Methods	4
Unsupervised Learning Methods	5
Web Scraping and Automated Data Collection	4

# III. SPECIAL TOPICS IN DATA ANALYSIS (available in January 2019)

Spotlight on Big Data Deep Learning and Artificial Intelligence Data Streams Recommender Engines Social Network Analysis Bayesian Data Analysis Putting it All Together

# **IV. STATISTICAL ANALYSIS WITH R** (available in June 2019)

Visual and Descriptive Statistics Univariate Analysis Multivariate Analysis Time Series Analysis Categorical Data Analysis Sampling Design

Participants must provide laptop computers with wireless connectivity. No software installation required.

Reaching expertise in R, Python, D3 or other programming languages is not a goal of the training modules.

Prelude to Data Science: Starting the Journey on the Right Foot	С	1	C	2	С	3	Сс	ore	Fu	ill –
Topics	Т	Ρ	Т	Ρ	Т	Ρ	Т	Ρ	Т	Ρ
01. Data science universals	1		2		2		3		3	
<b>02.</b> Understanding the world through systems and data							1		1	
03. Basics of R		1		1		1		2		2
04. Basics of Python				1		1		2		2
05. Data cleaning									1	1
06. Statistical analysis vs. data science							3		3	
07. Case studies	2		2		2		2		2	
08. Data reduction							2	1	2	1
09. Example: predicting algae blooms (with R)		2		2		2		2		2
Total	3	3	4	4	4	4	11	7	12	8

An introduction to the notions that must be mastered prior to embarking on data analysis, along with a discussion of common challenges and pitfalls. Participants will be introduced to various methods of data preparation, and to some intrinsic limitations of data and easily avoidable pre-analysis mistakes. The programming languages R and Python will be introduced.

# **INTENDED AUDIENCE**

Individuals with an interest in quantitative methods who wish to grasp the larger and broader "behind-the-scenes" issues that affect analysts' decisions and results. For the most part, this is not a workshop on formulas, computation or any specific analytical methods. Mathematical and/or statistical sophistication beyond introductory topics is not required.

<sup>&</sup>quot;Dencity" by Fathom - Previous page: "Health and Wealth of Nations, 2012" by the Gapminder Foundation



Elements of Data Visualization: Presenting Data in a New Light	C1	<b>C2</b>	<b>C3</b>	Core	Full
Topics	ΤP	ΤР	ΤP	ΤP	ΤР
10. Pre-analysis vs. post-analysis use				1.5	1.5
<b>11.</b> Simple graphical methods				1.5	1.5
12. Representations of multi-dimensional data				2	2
13. Basic visualization in R					1
14. Design suggestions			2	2	2
<b>15.</b> Storytelling with data					3 2
16. Data representation and accessibility					1
<b>17.</b> Advanced visualization in R (with ggplot2)					4
<b>18.</b> Interactive visualizations (with D3)					2
Total	0 0	0 0	2 0	7 0	11 9

# DESCRIPTION

An introduction to data visualization principles and to simple implementations with R (and/or Python) and D3. By the end of the course, participants will understand the basic ideas underlying data visualization and will be able to produce some simple visualizations using R and D3.

# **INTENDED AUDIENCE**

Individuals who have some conceptual understanding of data science, and who would like to gain some practical data visualization experience. Participation in the workshop *Prelude to Data Science* is not required, but familiarity with its contents is expected. Some exposure to R is assumed.

Mining for Information Gold: Introduction to Data Science	C1	C2	C3	Core	Full
Topics	ΤP	ΤP	ΤР	ΤР	ТР
19. Association rules mining	1.5 0.5	2 1	2 1	2 1	2 1
20. Classification and decision trees	1.5 0.5	2 1	2 1	2 1	2 1
<b>21.</b> Clustering and <i>k</i> -means	1.5 0.5	2 1	2 1	2 1	2 1
22. Data science issues and challenges	1.5	2	1	1	1
23. Naïve Bayes classifiers		1 1	1 1	1 1	1 1
24. Hierarchical clustering		1 1	1 1	1 1	1 1
25. Example: detecting anomalous transactions (with R)					2
26. Example: predicting wait times at an airport checkpoint (with R/Python)		2	2	2	2
27. Example: clustering departing flights (with R/Python)	1				2
Total	63	10 7	9 7	9 7	9 11

An introduction to the fundamental data science concepts involved in data analysis, with a detailed discussion of three common analytical concepts – classification, clustering and association rules. The application of these concepts will be illustrated through some simple toy examples, along with a discussion of common challenges and pitfalls. By the end of the course, participants will be able to appreciate the functionality of different types of data science concepts and recognize opportunities for their application when presented with real-world data.

#### **INTENDED AUDIENCE**

Individuals who wish to understand the functionality and capabilities offered by different data science concepts and methods, even if they won't be the ones implementing them. Participation in the workshop *Prelude to Data Science* is not required, but familiarity with its contents is expected. Some experience with quantitative thinking and with R/Python is assumed.





Text Mining and Natural Language Processing	C	21	C	22	C	3	Co	ore	F	ull
Topics	Т	Ρ	Т	Р	Т	Ρ	Т	Ρ	Т	Ρ
28. Text mining basics	1	1	1	1	1	1	1	1	1	1
29. Text data classification			1	1	1	1	1	1	1	1
<b>30.</b> Text data clustering			1	1	1	1	1	1	1	1
<b>31.</b> Sentiment analysis	1	1	1	1	1	1	1	1	1	1
<b>32.</b> Natural language processing basics and tasks	1		1		1	1	1	1	2	1
33. Natural language processing focus: named-entity recognition, summarization					2	2	2	2	2	2
34. Natural language processing focus: topic modeling, Markov-chain generation					2	2	2	2	2	2
<b>35.</b> Example: classification and topic modeling of UK government press releases										1
Total	3	2	5	4	9	9	9	9	10	10

# DESCRIPTION

An introduction to the basic concepts of text data analysis, illustrated with simple examples, along with a discussion of common challenges. By the end of the course, participants will be able to appreciate the unique challenges encountered in text analysis, and to apply appropriate data science models when presented with real-world text mining tasks.

#### **INTENDED AUDIENCE**

Individuals who wish to understand the functionality and capabilities offered by some text mining and natural language processing methods, even if they won't be the ones implementing them. Participation in the workshops *Prelude to Data Science* and *Mining for Information Gold* is not required, but familiarity with their content is expected. Some exposure to R/Python is assumed.

Techniques of Data Analysis I: Supervised Learning	C1	<b>C2</b>	C3	Core	Full
Topics	ΤP	ΤP	ΤP	ΤP	ΤP
36. Overview					1
37. Support vector machines			1 1	1 1	1 1
38. Artificial neural networks		1 1	1 1	1 1	1 1
<b>39.</b> Decision trees and variants (reprise)					1 1
40. Ensemble learning and gradient boosting			1 1	1 1	1 1
<b>41.</b> Performance evaluation and model selection		1	1	1 1	1 1
<b>42.</b> Value estimation methods				1 1	1 1
43. Logistic regression					1 1
44. Rare occurrence learning					1 1
<b>45.</b> High-dimensional classification					1 1
46. Practical suggestions			0.5	0.5	1
Total	0 0	2 1	4.5 3	5.5 5	11 9

A continuation of the introduction of the data science concepts started in *Mining for Information Gold*, with a detailed discussion of additional supervised learning methods/concepts commonly used by data scientists. The application of these concepts will be illustrated through some simple toy examples, along with a discussion of common challenges and pitfalls. By the end of the course, participants will be able to further appreciate the functionality of increasingly sophisticated classification methods and to recognize opportunities for their application when presented with real-world data.

## **INTENDED AUDIENCE**

Individuals who wish to understand the functionality and capabilities offered by different classification methods, even if they won't be the ones implementing them. This workshop is also designed for individuals who want to increase their data science competency by learning different data science techniques. Familiarity with the data science concepts introduced in *Mining for Information Gold* is assumed. Some experience with mathematical optimization would be beneficial, but is not necessary. Some exposure to R/Python is assumed.

"Interactive NBA Shot Charts" by Todd W. Schneider



We	eb Scraping and Automated Data Collection	C1 C:		C1		C1		C1		C1		C1		C1 (		C1 (		C1		C1		C1		C1 C2		C2		<b>C2</b>		C2		C2		<b>C2</b>		<b>C2</b>		<b>C2</b>		<b>C2</b>		C2		<b>C2</b>		<b>C2</b>		C1 C2		C1 C		C2		C	3	Сс	ore	F	ull
Тор	nics	Т	Р	Т	Ρ	Т	Ρ	Т	Ρ	Т	Ρ																																																
58.	Data analysis and web scraping									2	2																																																
59.	Web technologies 101									2	2																																																
60.	Scraping toolbox (XPath, Beautiful Soup, Selenium, etc.)									2	2																																																
61.	Specialized uses and applications									2	2																																																
62.	Practical examples with R/Python									0	4																																																
	Total	0	0	0	0	0	0	0	0	8	12																																																

#### DESCRIPTION

In the age of open data, information is at our fingertips. How can one get their hands on it? By the end of the course, participants will be able to write simple data extracting scripts in R and Python, and use these scripts to collect data from the Web and from existing electronic files.

#### **INTENDED AUDIENCE**

Individuals who may be required to prepare scripts for automated data collection and extraction. Some exposure to programming is assumed.

Techniques of Data Analysis II: Unsupervised Learning	C1	C2	C3	Core	Full
Topics	TP	TP	T P	ТР	ΤP
47. Overview					1
48. Density-based clustering		1 1	1 1	1 1	1 1
49. Spectral clustering			1 1	1 1	1 1
<b>50.</b> <i>k</i> -means and variants (reprise)					1 1
51. Cluster ensembles			1 1	1 1	1 1
52. Validation metrics		1	1 1	1 1	1 1
53. Latent Dirichlet allocation					1 1
54. Expectation-maximization				1 1	1 1
55. Fuzzy clustering					1 1
56. High-dimensional clustering					1 1
57. Practical suggestions			0.5	0.5	1
Tota	al 0 0	2 1	4.5 4	5.5 5	11 9

A continuation of the introduction of the data science concepts started in *Mining for Information Gold*, with a detailed discussion of additional unsupervised learning methods/concepts commonly used by data scientists. The application of these concepts will be illustrated through some simple toy examples, along with a discussion of common challenges and pitfalls. By the end of the course, participants will be able to further appreciate the functionality of increasingly sophisticated clustering methods and to recognize opportunities for their application when presented with real-world data.

## **INTENDED AUDIENCE**

Individuals who wish to understand the functionality and capabilities offered by different clustering methods, even if they won't be the ones implementing them. This workshop is also designed for individuals who want to increase their data science competency by learning different data science techniques. Familiarity with the data science concepts introduced in *Mining for Information Gold* is assumed. Some experience with mathematical optimization would be beneficial, but is not necessary. Some exposure to R/Python is assumed.