



# COMP47350: Data Analytics (Conv)

Dr. Georgiana Ifrim

[georgiana.ifrim@ucd.ie](mailto:georgiana.ifrim@ucd.ie)

Insight Centre for Data Analytics

School of Computer Science

University College Dublin

2016/17

# Module Overview

- Module website:  
<https://csmoodle.ucd.ie/moodle/course/view.php?id=544>
- Self-enroll with your user name and password:  
**comp47350162**
- Lectures: Tuesday@11 & Thursday@12 (B0.02)
- Labs: Thursday @14 (E2.16), **laptop required!**
- Course Lecturer: Dr. Georgiana Ifrim
  - Office hours: by appointment; office @ UCD- Insight (O'Brien Centre, E3.36)
- Course TA: Gevorg Poghosyan  
(Gevorg.Poghosyan@ucdconnect.ie )

# Learning Objectives

1. Understand the principles and the purposes of data analytics.
2. Use Python to retrieve and analyse real-world datasets.
3. Apply the process of data understanding and address data quality issues.
4. Use appropriate machine learning techniques for a given data analytics problem.
5. Design evaluation experiments for selecting the best predictive model for a given analytics problem.

# Module Topics

- **Python Environment** (Anaconda, Jupyter Notebook)

## **CRISP-DM Methodology**

- **Getting Data** (Web scrapping, APIs, DBs)
- **Understanding Data** (slicing, visualisation)
- **Preparing Data** (cleaning, transformation)
- **Modeling & Evaluation** (machine learning)

# Assessment

Marks distribution, 3 components:

- Homework: 60%
  - Starting Week3, due after 2 weeks
  - 2 assessed homeworks, 30% each
- Project: 30%
  - Starting Week7
  - Group project, 4-5 students per group
- Presentation: 10%
  - In the last week

# Late Submission Policy

- Up to 1 week delay, 10% marks lost
- Between 1-2 weeks, 20% marks lost
- 2 weeks cut-off (no submissions accepted after 2 weeks)

# Marks to grades mapping used in this module

## Pass mark 40%

Grade	Lower	Upper	Calculation Point
A+	95	100	97.5
A	90	95	92.5
A-	85	90	87.5
B+	80	85	82.5
B	75	80	77.5
B-	70	75	72.5
C+	65	70	67.5
C	60	65	62.5
C-	55	60	57.5
D+	50	55	52.5
D	45	50	47.5
D-	40	45	42.5

Grade	Lower	Upper	Calculation Point
E+	35	40	37.5
E	30	35	32.5
E-	25	30	27.5
F+	20	25	22.5
F	15	20	17.5
F-	10	15	12.5
G+	8	10	9
G	5	8	6.5
G-	2	5	3.5
NG	0	0	0



# Plagiarism & UCD Computer Science

- **Plagiarism is a serious academic offence**
  - [Student Code, section 6.2] or [UCD Registry Plagiarism Policy] or [CS Plagiarism policy and procedures]
- Our staff and demonstrators are **proactive** in looking for possible plagiarism in all submitted work
- Suspected plagiarism is reported to the CS Plagiarism subcommittee for investigation
  - Usually includes an interview with student(s) involved
  - 1st offence: **usually** 0 or NG in the affected components
  - 2nd offence: referred to the **University disciplinary committee**
- Student who enables plagiarism is equally responsible

[http://www.ucd.ie/registry/academicsecretariat/docs/plagiarism\\_po.pdf](http://www.ucd.ie/registry/academicsecretariat/docs/plagiarism_po.pdf)

[http://www.ucd.ie/registry/academicsecretariat/docs/student\\_code.pdf](http://www.ucd.ie/registry/academicsecretariat/docs/student_code.pdf)

<http://libguides.ucd.ie/academicintegrity>



# Questions & E-mail

- If you have a technical question, post to Moodle
- Only send e-mail if of personal nature (e.g., health or grade related)
- If you need to discuss, set an appointment (please do not show up at TA's or lecturer's desk without prior appointment)

# What is Big Data?

1 Zettabyte  
= 1 billion  
Terrabytes



## What Happens in an Internet Minute?

1,572,877 GB of global IP data transferred<sup>1</sup>

10 Million  
ads displayed<sup>2</sup>

347,222  
Tweets<sup>3</sup>

3.3 Million  
pieces of  
content shared<sup>4</sup>

6.9 Million  
messages sent<sup>4</sup>

Netflix + Youtube =  
more than 1/2 of  
all traffic<sup>5</sup>

\$400 Million  
during Alibaba  
peak day sales<sup>6</sup>

438,801  
Wiki page views<sup>7</sup>

10 Million  
WeChat messages at its peak<sup>9</sup>

34.7 Million  
instant messages  
(iMessage) sent<sup>8</sup>

194,064  
app downloads<sup>10</sup>

\$133,436  
in sales<sup>11</sup>

31,773  
hours of  
music played<sup>12</sup>

38,194  
photos  
uploaded<sup>13</sup>

57,870  
page views<sup>14</sup>

100  
hours of video  
uploaded<sup>16</sup>

4.1 Million  
searches<sup>15</sup>

138,889  
hours of video  
watched<sup>16</sup>

23,148  
hours of video  
watched<sup>17</sup>

And Future  
Growth is  
Staggering



By 2017, mobile  
traffic will have grown  
**13X** in just  
5 years<sup>1</sup>



In 2017, there will be  
**3X** more connected devices  
than people on Earth<sup>1</sup>

All digital data created reached  
**4 zettabytes** in 2013<sup>18</sup>

# What is Data Analytics?

**Making sense of data:**

**Knowledge**  **Decisions**



# Other Names for Data Analytics

- Data Science
- Artificial Intelligence (AI)
- Machine Learning
- Data Mining
- Statistics
- Business Intelligence
- Customer Relationship Management

# Types of Data Analytics

- **Supervised Learning (A->B learning)**
  - **Classification:** Input A: email, Output B (label): spam/nospam
  - **Regression:** Input A: past stock prices, Output B (number): future stock price
- **Unsupervised Learning**
  - **Clustering:** Find structure in input data, e.g., similar groups or hierarchy of groups (given Input A: news articles; discover groups of similar articles)
  - **Association:** Find frequent item co-occurrence (given Input A: transactions with purchased products; discover products that are frequently purchased together)

# What Machine Learning Can Do

A simple way to think about supervised learning.

INPUT A	RESPONSE B	APPLICATION
Picture	Are there human faces? (0 or 1)	Photo tagging
Loan application	Will they repay the loan? (0 or 1)	Loan approvals
Ad plus user information	Will user click on ad? (0 or 1)	Targeted online ads
Audio clip	Transcript of audio clip	Speech recognition
English sentence	French sentence	Language translation
Sensors from hard disk, plane engine, etc.	Is it about to fail?	Preventive maintenance
Car camera and other sensors	Position of other cars	Self-driving cars

SOURCE ANDREW NG

© HBR.ORG

<https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now>

- Human needs to carefully decide what A and B is
- Human needs to provide example data for learning an A->B relationship

# Data Analytics Jobs



# Popular Data Analytics Tools

## Main tools: R and Python

<http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html>

- **R**: open-source, Matlab-like language
  - Developed by [statisticians](#) (Ross Ihaka and Robert Gentleman, 1995)
  - Steep learning curve
  - Great support for stats and visualization (good IDE: Rstudio)
- **Python**: open-source, scripting language
  - Developed by a [computer programmer](#) (Guido Van Rossem, 1991)
  - Growing fast, easy to learn, good for developing full-fledge projects (good IDE: Spyder, PyCharm)

**We will use Python 3.5 in this module**



# Data Analytics Websites & Platforms

- [Kdnuggets](http://www.kdnuggets.com): A website for Data Mining, Analytics, Big Data, Data Science
  - Lots of good resources, tutorials, books:  
<http://www.kdnuggets.com/2015/09/free-data-science-books.html>
- [Kaggle](https://www.kaggle.com/wiki/Home): A platform for Data Science competitions (problems, datasets, scripts)
  - Learn from the solutions of the best competitors:  
<https://www.kaggle.com/wiki/Home>

# Module Topics

- **Python Environment (Anaconda, Jupyter Notebook)**

## **CRISP-DM Methodology**

- **Getting Data** (Web scrapping, APIs, DBs)
- **Understanding Data** (slicing, visualisation)
- **Preparing Data** (cleaning, transformation)
- **Modeling & Evaluation** (machine learning)

# References

Good free e-books for data analytics:

<http://www.kdnuggets.com/2015/09/free-data-science-books.html>

Book for Python 3:

**Dive Into Python3**, Mark Pilgrim (e-book)

<http://www.diveintopython3.net>

Online resources:

- <http://www.kdnuggets.com/2012/10/ipython-notebook-environment-for-data-science.html>
- <http://www.kdnuggets.com/2015/07/continually-updated-data-science-ipython-notebooks.html>
- <http://chrisalbon.com>
- <http://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/learning-path-data-science-python/>