Applied Big Data and Visualization

P. Healy

CS1-08
Computer Science Bldg.
tel: 202727
patrick.healy@ul.ie

Spring 2021–2022
Outline

1. Administrative Details
   - Meeting Times
   - General Issues
   - Assessment

2. Syllabus
   - Overview
   - Learning Outcomes

3. Data Deluge
Outline

1. Administrative Details
   - Meeting Times
   - General Issues
   - Assessment

2. Syllabus
   - Overview
   - Learning Outcomes

3. Data Deluge
Lectures / Labs / Tutes

Lecture Hours:  
- Mon. 11h00  D1050
- Fri. 11h00  KBG12

Lab  
- 2D  Mon. 09h00  CS3(0)04B
- 2B  Mon. 14h00  CS3(0)04B
- 2E  Tue. 13h00  CS3(0)04B
- 2C  Thu. 11h00  CS3(0)04B
- 2F  Thu. 14h00  CS3(0)04B
- 2A  Thu. 16h00  CS3(0)04B

4 contact hours → 6 non-contact hours
Outline

1. Administrative Details
   - Meeting Times
   - General Issues
   - Assessment

2. Syllabus
   - Overview
   - Learning Outcomes

3. Data Deluge
Attendance at all lectures and labs / tutes is expected
Handing up all assigned lab exercises is a good idea
Handing up other people’s work is a serious
All lectures, homeworks, past exams, etc. can be found on the class home page:

```
garryowen.csisdmz.ul.ie/~cs6502/
g```

Class lists and attendance records will also be available here
1. All lectures, homeworks, past exams, etc. can be found on the class home page:
   `garryowen.csisdmz.ul.ie/~cs6502/`

2. Class lists and attendance records will also be available here
Acknowledgements

- Andrew Ju and Nik Nikolov have made significant contributions
- Wikipedia images
- Various youtube videos / data science lectures from around the world / web
Outline

1. Administrative Details
   - Meeting Times
   - General Issues
   - Assessment

2. Syllabus
   - Overview
   - Learning Outcomes

3. Data Deluge
Assessment Instruments

Lab Assignments: \( 5 \times 6\% \)
Final: \( 70\% \) Week15
Mid-Term: \( 15\% \) Week09
Final: \( 55\% \) Week15

- Labs start next week
- September repeat exam will count for same as final
  - What this means...
<table>
<thead>
<tr>
<th>Grade</th>
<th>Score Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0 – 29</td>
</tr>
<tr>
<td>D2</td>
<td>30 – 34</td>
</tr>
<tr>
<td>D1</td>
<td>35 – 39</td>
</tr>
<tr>
<td>C3</td>
<td>40 – 47</td>
</tr>
<tr>
<td>C2</td>
<td>48 – 51</td>
</tr>
<tr>
<td>C1</td>
<td>52 – 55</td>
</tr>
<tr>
<td>B3</td>
<td>56 – 59</td>
</tr>
<tr>
<td>B2</td>
<td>60 – 63</td>
</tr>
<tr>
<td>B1</td>
<td>64 – 71</td>
</tr>
<tr>
<td>A2</td>
<td>72 – 79</td>
</tr>
<tr>
<td>A1</td>
<td>80 – 100</td>
</tr>
</tbody>
</table>
Main text (though not mandatory):

- *Lakshmanan, Valliappa*
  Data Science on the Google Cloud Platform (O’Reilly) ISBN 9781491974551
  On request in library; book’s home page
Other useful books:

- *Vijay Srinivas Agneeswaran* Big data analytics beyond Hadoop: real-time applications with Storm, Spark, and more Hadoop alternatives, Pearson Education (2014)
Linux password creation / resetting\(^1\) will be done by Liam O’Riordan, liam.oriordan@ul.ie, (CS2-004) between 14.30 - 15.30 from today onwards

Don’t forget to register online at [http://www.si.ul.ie](http://www.si.ul.ie)

\(^1\)Needed only if you’ve forgotten your old one!
**Announcements**

- Labs start in Week02
- Will need to sign up to Google Cloud (for later in the semester)
  - url is [https://cloud.google.com/free/](https://cloud.google.com/free/)
  - use UL student email
  - you will be asked for card details so make sure that auto-billing is turned off!
Outline

1. Administrative Details
   - Meeting Times
   - General Issues
   - Assessment

2. Syllabus
   - Overview
   - Learning Outcomes

3. Data Deluge
Rationale and Purpose

Introduce students to big data management and associated issues. The topics include overview of the Apache toolset (Hadoop, Spark, and others), Google Cloud, distributed file systems, big data programming models, data warehousing and big data security and protection; visualization tools and frameworks.
Overview of Syllabus

- "Big data": meaning and sources; the Vs of big data; data governance: accuracy, availability, usability and security; impacts of big data, industrial and societal;
- Big data programming frameworks and systems: distributed file systems, scalable computing, the MapReduce programming model, the Spark programming and computing model, overview of the main components of the Hadoop ecosystem;
- Data warehousing concepts: what is a data warehouse; role of a data warehouse in data management; architecture of a data warehouse; ETL: extraction, transformation, load process, data marts; operational systems vs. data warehouses
Overview of Syllabus (contd.)

- Big data security and protection challenges and practices, such as privacy-preserving data composition, encryption, granular access control, user authentication models, endpoint filtering and validation, etc.
- Relational information in a business context; visualization challenges; graph / network visualization frameworks: Sugiyama and force-directed layout methods.
Outline

1. Administrative Details
   - Meeting Times
   - General Issues
   - Assessment

2. Syllabus
   - Overview
   - Learning Outcomes

3. Data Deluge
Learning Outcomes: Cognitive

On successful completion of this module students will be able to (cognitively):

- Recognise the technological challenges in big data governance
- Summarise the Hadoop ecosystem
- Describe the MapReduce programming model
- Describe the architectural components of a data warehouse
- Discuss the big data security challenges and practices.
Learning Outcomes: Affective

On successful completion of this module students will be able to (affectively):

- Discuss the impact of big data on industry and society
- Recognise the importance of big data security
- Awareness of ethical issues associated with use and misuse of big data.
A Snapshot of Internet Activity, 2021

2021: This Is What Happens In An Internet Minute

- Facebook: 1.4 Million Scrolling, 21.1 Million Texts Sent
- YouTube: 500 Hours Content Uploaded, 414,764 Apps Downloaded
- LinkedIn: 9,132 Connections Made
- Netflix: 28,000 Subscribers Watching
- Instagram: 695,000 Stories Shared
- Twitter: 200,000 People Tweeting
- Tinder: 2 Million Swipes
- Google: 932 Smart Audio Devices Shipped
- Amazon: 197.6 Million Emails Sent
- TikTok: 5,000 Downloads
- Twitch: 2 Million Views

60 SECONDS

Created By:
@LoriLewis
@OfficiallyChadd
Comparison of Internet Activity: 2019 vs 2021

2019: This is what happens in an internet minute:
- 3.8 Million Google searches
- 1 Million YouTube videos viewed
- 1 Million Twitter tweets
- 87,500 Instagram posts
- 2.1 Million Snapchat snaps created
- 41.6 Million Facebook messages sent
- 4.8 Million GIFs served
- 188 Million emails sent
- 1 Million streaming subscriptions
- 1 Million views
- 694,444 hours watched
- 1 Million WhatsApp messages
- 1 Million Alexa commands
- 1 Million Snapchats sent
- 2.1 Million Instagram posts
- 596,956 spent online

2021: This is what happens in an internet minute:
- 4.5 Million Google searches
- 4.5 Million YouTube videos viewed
- 3.4 Million Twitter tweets
- 211 Million Instagram posts
- 1 Million Snapchat snaps created
- 418 Million Facebook messages sent
- 1.4 Million GIFs served
- 1 Million TikTok videos
- 1 Million views
- 2.1 Million hours watched
- 1 Million WhatsApp messages
- 1 Million Alexa commands
- 1 Million Snapchats sent
- 2 Million Instagram posts
- 197.6 Million spent online

Created by: @LoriLewis @OfficiallyChadd
Big Data – what is it?

- **Wikipedia**: “Big data is a field that [considers] ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.”

- Predictive analytics

- Data increasingly gathered by cheap and numerous information-sensing devices

- “Global data volume will grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020. By 2025, IDC predicts there will be 163 zettabytes of data”

- $1 \text{ zettabyte} = 10^{12} \text{ GB} = 10^{21} \text{ B}$

- mega (M); giga (G); tera (T); peta (P); exa (E); zetta (Z); yotta (Y)
Netflix

- over 100 million subscribers; customer responses / selections drive their recommendation system
- developed in 2009 but constantly updated / revised
- recommendation system influences $\sim 80\%$ of streamed content

UOB bank from Singapore (Andreas, 2014) (link)

- uses big data to drive risk management; huge potential for incurring losses if not well thought out
- recently tested a risk management system (RMS) that is based on big data
- RMS enabled bank to reduce calculation time of the value at risk from about 18 hours to a few minutes
- the bank will possibly be able to carry out real-time risk analysis in the near future
The Flood

- Netflix
  - over 100 million subscribers; customer responses / selections drive their recommendation system
  - developed in 2009 but constantly updated / revised
  - recommendation system influences ~80% of streamed content

- UOB bank from Singapore (Andreas, 2014) (link)
  - uses big data to drive risk management; huge potential for incurring losses if not well thought out
  - recently tested a risk management system (RMS) that is based on big data
  - RMS enabled bank to reduce calculation time of the value at risk from about 18 hours to a few minutes
  - the bank will possibly be able to carry out real-time risk analysis in the near future
Formula One Racing:

- “On each lap, an F1 machine generates 35MB of data to analyse” \(\Rightarrow\) \(\sim\)30GB over a grand prix weekend with two cars
- 600 channels (streams) of data per car, sampling (recording) sensors at 200Hz – 10kHz

Google Flu Trends

- nowcasting
Words of Caution

- Big structures are full of spurious correlations
- More spurious correlations [here](#)
- Law of truly large numbers [Wikipedia](#)
- Google Flu Trends failed to deliver good predictions in recent years, overstating the flu outbreaks by a factor of two; has been mothballed
- Corporate surveillance
Words of Caution

- Big structures are full of spurious correlations

More spurious correlations [here](#)
- Law of truly large numbers [Wikipedia](#)
- Google Flu Trends failed to deliver good predictions in recent years, overstating the flu outbreaks by a factor of two; has been mothballed
- Corporate surveillance
Big structures are full of spurious correlations

More spurious correlations [here](#)

- Law of truly large numbers [Wikipedia](#)
- Google Flu Trends failed to deliver good predictions in recent years, overstating the flu outbreaks by a factor of two; has been mothballed
- Corporate surveillance
Words of Caution

- Big structures are full of spurious correlations
- More spurious correlations here
- Law of truly large numbers Wikipedia
- Google Flu Trends failed to deliver good predictions in recent years, overstating the flu outbreaks by a factor of two; has been mothballed
- Corporate surveillance
Words of Caution

- Big structures are full of spurious correlations
- More spurious correlations [here](#)
- Law of truly large numbers [Wikipedia](#)
- Google Flu Trends failed to deliver good predictions in recent years, overstating the flu outbreaks by a factor of two; has been mothballed
- Corporate surveillance
Words of Caution

- Big structures are full of spurious correlations
- More spurious correlations [here](#)
- Law of truly large numbers [Wikipedia](#)
- Google Flu Trends failed to deliver good predictions in recent years, overstatement the flu outbreaks by a factor of two; has been mothballed
- Corporate surveillance