FIT1043 Introduction to Data Science

Week 1, Data Science

Ian K T Tan

School of Information Technology
Monash University Malaysia

Overview of Data Science

ePub Section 1.1

MONASH University

Learning Outcomes

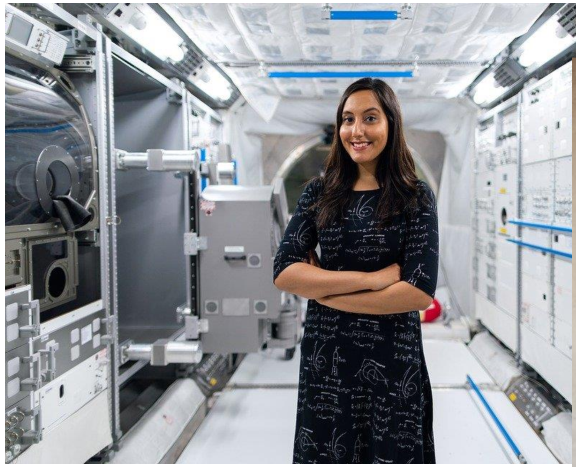
Week 1

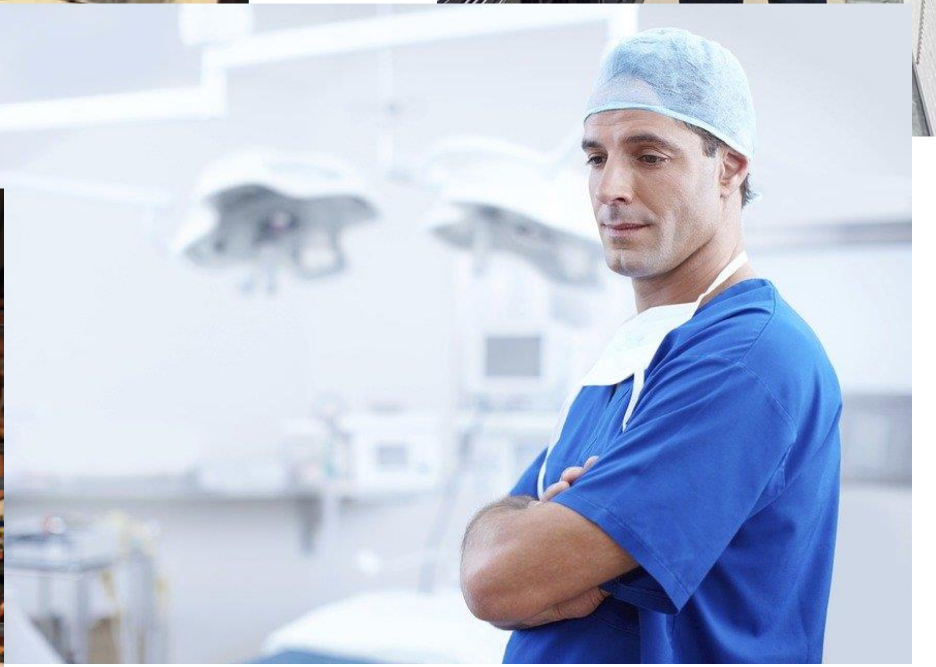
By the end of this week you should be able to:

- Explain what is data science and Drew Conway’s Venn diagram
- Comprehend the usefulness of machine learning
- Explain different components of a data science process
- Differentiate data science from other related disciplines
- Learn how to install and start coding in Python with Jupyter Notebook
    - To be achieved in your tutorial / laboratory session

Data Scientist

Different professions





What is Data Science?

“name contains the word ‘science’, so it can’t be one”

- This is a very old joke as most of us know sciences as Biology, Physics, Chemistry and so on.
- They don’t have the word science in them ☺

“data science is what a data scientist does”

- A circular definition, which is practically useless.
- Please do not answer like this.

What is Data Science?

“data science is the technology of handling and extracting value from data”

- This is a less circular definition and slightly more useful.
- It isn’t a bad attempt at it.

“machine learning on big data”

- Useful, but it is too narrow.

What is Data Science?

Data Science and Big Data (links to Wikipedia)

Data Science

- Data Science is the extraction of knowledge from data, which is a continuation of the field data mining and predictive analytics.
- *Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.*
- Data science is related to data mining and big data.

Big Data

- Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate.
- *Big data is a field that treats 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.*

What is Data Science?

Quote from Hal Varian

The ability to **take data** and;

- to be able to **understand** it,
- to **process** it,
- to **extract value** from it,
- to **visualize** it,
- to **communicate** it

That's going to be a hugely important skill in the next decades.

What is Data Science?

In Summary

Examples

- narrow: machine learning on big data
- broad: extraction of knowledge/value from data through the complete data lifecycle process

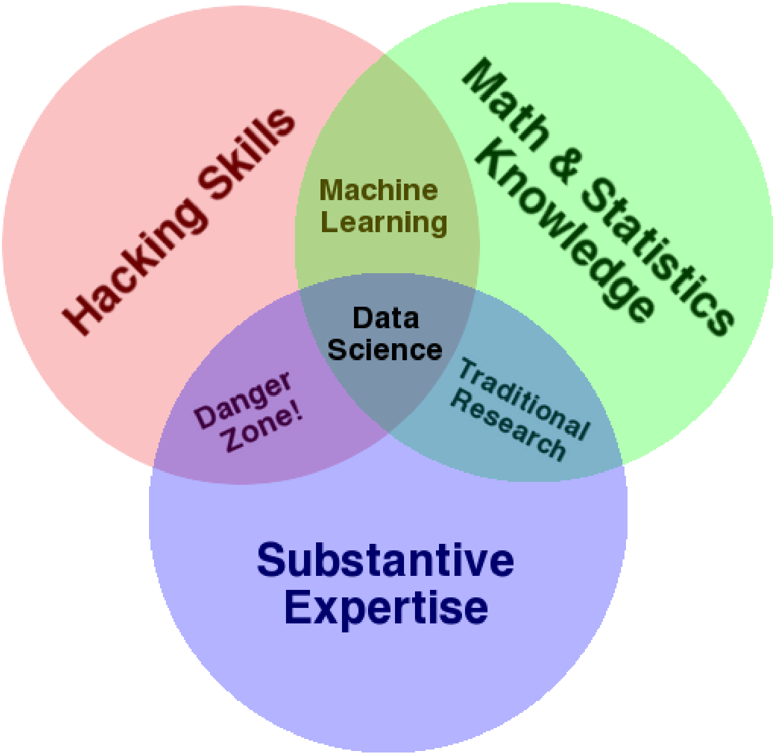
Should include

- broad concern with the *different stages*
- focus on the learning/*knowledge discovery*

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

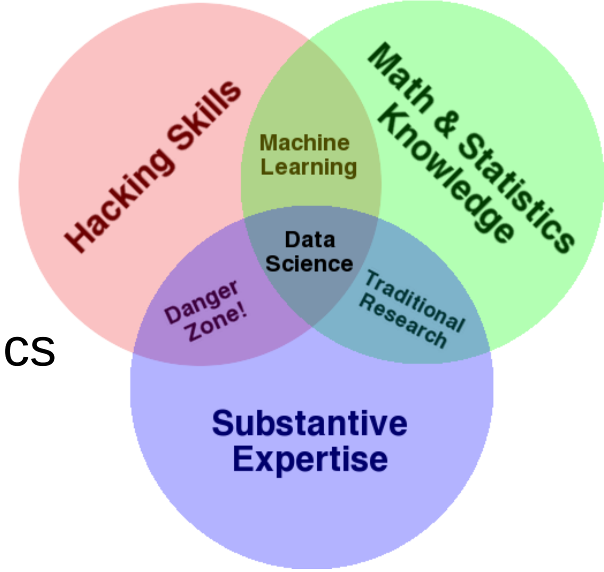
Data Science Venn Diagram

Drew Conway (http://drewconway.com/)



MONASH University

Data Science Venn Diagram



- Able to obtain the data (using your hacking skills) plus mathematics & statistics gets you machine learning.
    - Many CS graduates starts with the title **Machine Learning** Engineer

- Substantive expertise plus mathematics & statistics is **traditional research**.
    - Doctoral level researchers spend a lot of their time acquiring expertise in specific areas, but very little time learning about technology.

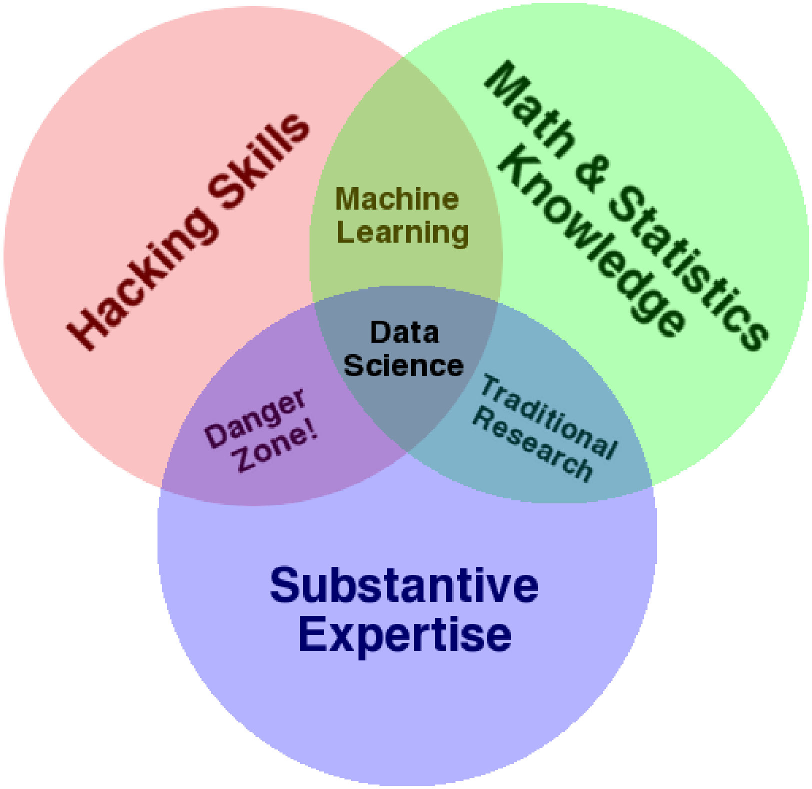
- Capable of extracting and structuring data and has good knowledge of the field equates a **danger zone**.
    - Appears to be a legitimate analysis without any understanding of how they got there or what they have created

Data Science Venn Diagram

Conclusion

- Combination of different skill sets

- Diverse skills are needed



MONASH University

Data Science Applied

Examples of Data Science being applied

Microsoft predictive analytic for traffic forecast

iOS predictive text

Google translation engine

Amazon recommender system

Health research studies on saturated fat

Machine Learning

Machine Learning Definition

Well understood definition and widely agreed upon:

Machine Learning is concerned with the development of algorithms and techniques that allow computers to learn.

- Concerned with building computer programs that can learn, often with computational output.
- With the underlying theory in statistics

Further reading at A Gentle Guide to Machine Learning

Why Use Machine Learning?

Machine learning is useful when:

- Human expertise is not available, e.g. Martian exploration

- Humans cannot explain their expertise (as a set of rules), or their explanation is incomplete and needs tuning, e.g. speech recognition

- Many solutions need to be adapted automatically, e.g. user personalisation

image sources: theconversation.com, meduim.com, blog.prioridata.com

Why Use Machine Learning?

Machine learning is useful when:

- Situation changes over time, e.g. junk email

- There are large amounts of data, e.g. discover astronomical objects

- Humans are expensive to use for the work, e.g. handwritten zipcode recognition

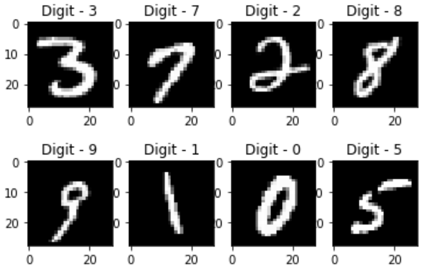


image sources: lifewire.com, clealyexplained.com, medium.com

Why Use Machine Learning?

Summary

- Humans are incapable of it
- Automation
  - Large Amount of Data
  - Humans are too expensive to be used in such a situation
- Experts Systems are generally rule based.
  - But many situations change all the time.

you do not want to be this poor guy!

sifting through all the data by hand

Recap: Learning Outcomes

Week 1

By the end of this week you should be able to:

- **Explain what is data science and Drew Conway's Venn diagram**
- **Comprehend the usefulness of machine learning**
- Explain different components of a data science process
- Differentiate data science from other related disciplines
- Learn how to install and start coding in Python with Jupyter Notebook
    - To be achieved in your tutorial / laboratory session

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram