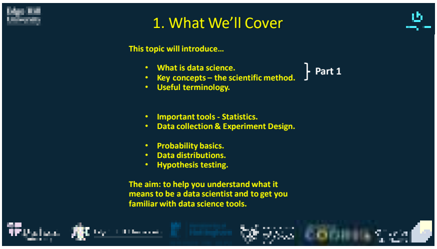
Advanced Data Science

Topic 11b – Part 1



1. What We'll Cover

This topic will introduce…

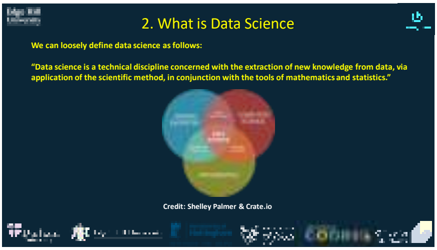
- What is data science.
- Key concepts – the scientific method.
- Useful terminology.

} Part 1

- Important tools - Statistics.
- Data collection & Experiment Design.

- Probability basics.
- Data distributions.
- Hypothesis testing.

The aim: to help you understand what it means to be a data scientist and to get you familiar with data science tools.

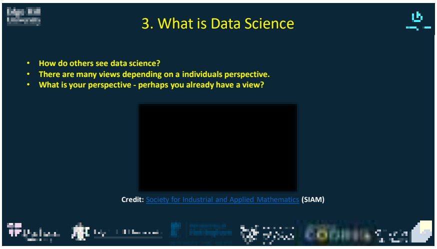


2. What is Data Science

We can loosely define data science as follows:

"Data science is a technical discipline concerned with the extraction of new knowledge from data, via application of the scientific method, in conjunction with the tools of mathematics and statistics."

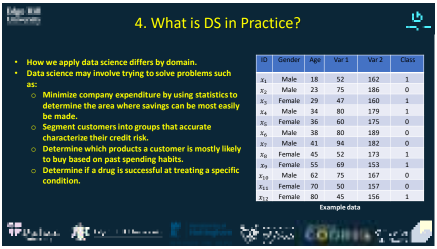
Credit: Shelley Palmer & Crate.io



3. What is Data Science

- How do others see data science?
- There are many views depending on a individuals perspective.
- What is your perspective - perhaps you already have a view?

Credit: Society for Industrial and Applied Mathematics (SIAM)



4. What is DS in Practice?

- How we apply data science differs by domain.
- Data science may involve trying to solve problems such as:
  - Minimize company expenditure by using statistics to determine the area where savings can be most easily be made.
  - Segment customers into groups that accurate characterize their credit risk.
  - Determine which products a customer is mostly likely to buy based on past spending habits.
  - Determine if a drug is successful at treating a specific condition.

ID	Gender	Age	Var 1	Var 2	Class
x_1	Male	18	52	162	1
x_2	Male	23	75	186	0
x_3	Female	29	47	160	1
x_4	Male	34	80	179	1
x_5	Female	36	60	175	0
x_6	Male	38	80	189	0
x_7	Male	41	94	182	0
x_8	Female	45	52	173	1
x_9	Female	55	69	153	1
x_{10}	Male	62	75	167	0
x_{11}	Female	70	50	157	0
x_{12}	Female	80	45	156	1

Example data

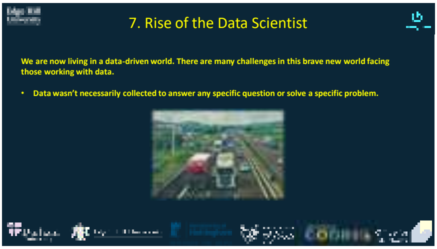
5. A New Discipline?

- Data Science is not a new discipline.
- Humans have been colleting/analysing data for centuries.
- Astronomy provides a famous example.

"Data science is a technical discipline concerned with the extraction of new knowledge from data, via application of the scientific method, in conjunction with the tools of mathematics and statistics."

6. What's Changed?

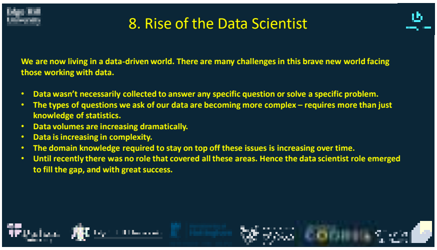
- For most of history "data science" activities were undertaken by experimental scientists.
- Contrast this to the modern world we live in. Everything has changed - data is generated about us everyday.
- Clearly data is everywhere. But little of it was collected with the aim of revealing new knowledge.



7. Rise of the Data Scientist

We are now living in a data-driven world. There are many challenges in this brave new world facing those working with data.

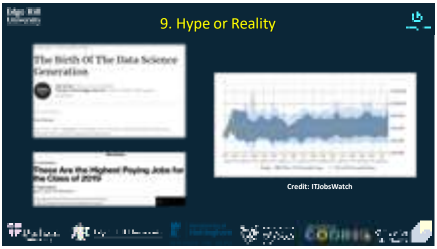
- Data wasn't necessarily collected to answer any specific question or solve a specific problem.



8. Rise of the Data Scientist

We are now living in a data-driven world. There are many challenges in this brave new world facing those working with data.

- Data wasn't necessarily collected to answer any specific question or solve a specific problem.
- The types of questions we ask of our data are becoming more complex – requires more than just knowledge of statistics.
- Data volumes are increasing dramatically.
- Data is increasing in complexity.
- The domain knowledge required to stay on top off these issues is increasing over time.
- Until recently there was no role that covered all these areas. Hence the data scientist role emerged to fill the gap, and with great success.

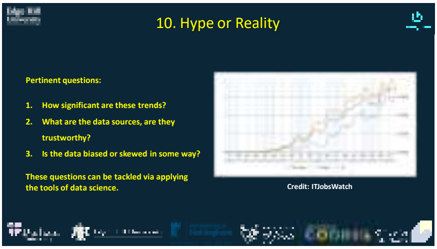


9. Hype or Reality

The Birth Of The Data Science Generation

These Are the Highest Paying Jobs for the Class of 2019

Credit: ITJobsWatch



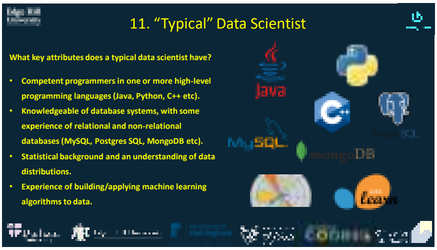
10. Hype or Reality

Pertinent questions:

1. How significant are these trends?
2. What are the data sources, are they trustworthy?
3. Is the data biased or skewed in some way?

These questions can be tackled via applying the tools of data science.

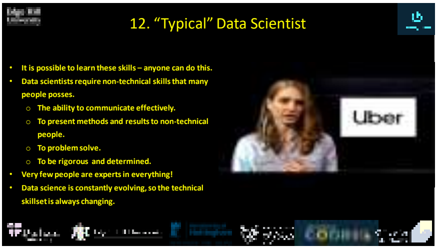
Credit: ITJobsWatch



11. “Typical” Data Scientist

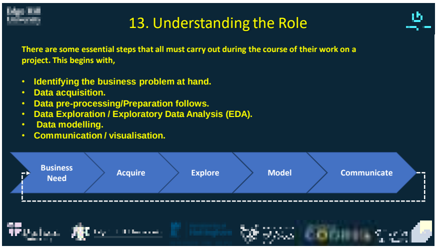
What key attributes does a typical data scientist have?

- Competent programmers in one or more high-level programming languages (Java, Python, C++ etc).
- Knowledgeable of database systems, with some experience of relational and non-relational databases (MySQL, Postgres SQL, MongoDB etc).
- Statistical background and an understanding of data distributions.
- Experience of building/applying machine learning algorithms to data.



12. "Typical" Data Scientist

- It is possible to learn these skills – anyone can do this.
- Data scientists require non-technical skills that many people posses.
  - The ability to communicate effectively.
  - To present methods and results to non-technical people.
  - To problem solve.
  - To be rigorous and determined.
- Very few people are experts in everything!
- Data science is constantly evolving, so the technical skillset is always changing.

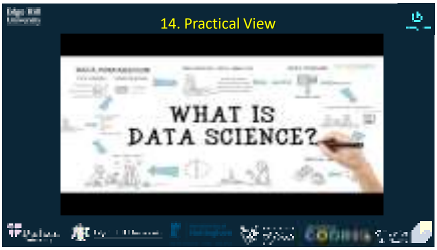


13. Understanding the Role

There are some essential steps that all must carry out during the course of their work on a project. This begins with,

- Identifying the business problem at hand.
- Data acquisition.
- Data pre-processing/Preparation follows.
- Data Exploration / Exploratory Data Analysis (EDA).
- Data modelling.
- Communication / visualisation.

Business Need → Acquire → Explore → Model → Communicate



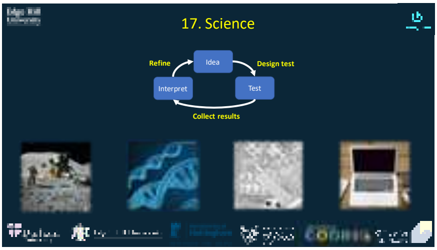
14. Practical View

15. Practical View

Credit: Chevron

16. Why Important?

- Data science is becoming an increasingly important discipline.
- As a society we continue to collect more and more data.
- This data is being used to make decisions. Such decisions can affect the lives of people in unforeseen ways, both good and bad.
- It is crucial that we train competent and professional data scientists capable of owning such responsibility.
- If we do this well, data scientists have the potential to help society in many.
- For all the positive possibilities there are an equal number of potentially negative outcomes
- This is why data science, and principled data science, is crucially important.



17. Science

Idea

Design test

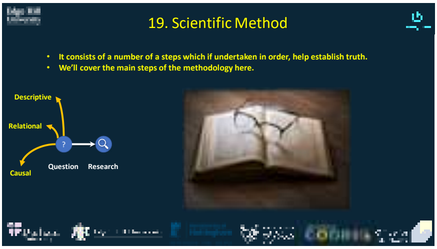
Test

Collect results

Interpret

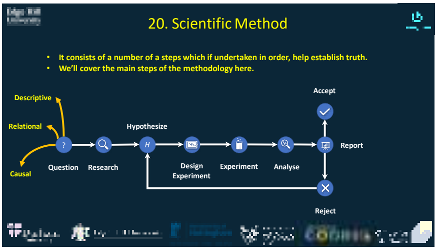
Refine

18. The Method



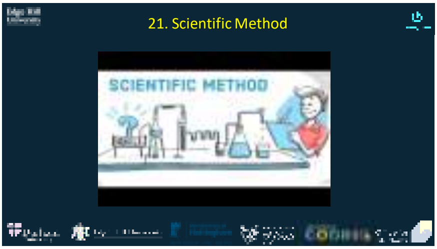
19. Scientific Method

- It consists of a number of a steps which if undertaken in order, help establish truth.
- We'll cover the main steps of the methodology here.

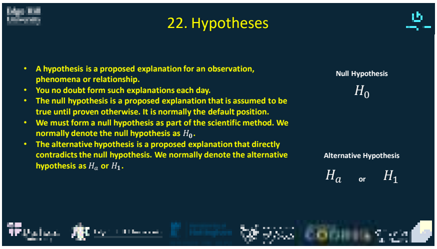


20. Scientific Method

- It consists of a number of a steps which if undertaken in order, help establish truth.
- We'll cover the main steps of the methodology here.



21. Scientific Method



22. Hypotheses

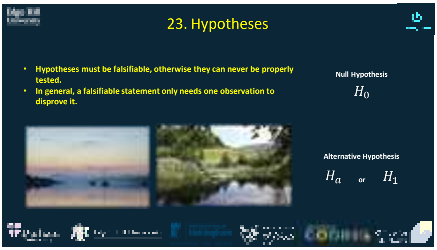
- A hypothesis is a proposed explanation for an observation, phenomena or relationship.
- You no doubt form such explanations each day.
- The null hypothesis is a proposed explanation that is assumed to be true until proven otherwise. It is normally the default position.
- We must form a null hypothesis as part of the scientific method. We normally denote the null hypothesis as H_0.
- The alternative hypothesis is a proposed explanation that directly contradicts the null hypothesis. We normally denote the alternative hypothesis as H_a or H_1.

Null Hypothesis

H_0

Alternative Hypothesis

H_a or H_1



23. Hypotheses

- Hypotheses must be falsifiable, otherwise they can never be properly tested.
- In general, a falsifiable statement only needs one observation to disprove it.

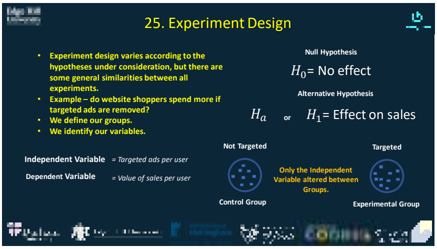
Null Hypothesis

H_0

Alternative Hypothesis

H_a or H_1

24. Hypotheses



25. Experiment Design

- Experiment design varies according to the hypotheses under consideration, but there are some general similarities between all experiments.
- Example – do website shoppers spend more if targeted ads are removed?
- We define our groups.
- We identify our variables.

Null Hypothesis

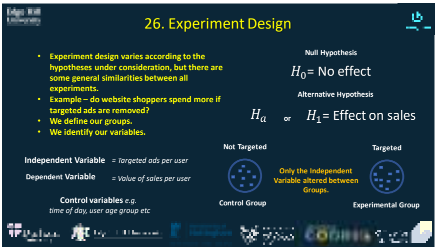
H_0= No effect

Alternative Hypothesis

H_a or H_1= Effect on sales

Independent Variable = *Targeted ads per user*

Dependent Variable = *Value of sales per user*



26. Experiment Design

- Experiment design varies according to the hypotheses under consideration, but there are some general similarities between all experiments.
- Example – do website shoppers spend more if targeted ads are removed?
- We define our groups.
- We identify our variables.

Null Hypothesis

H_0= No effect

Alternative Hypothesis

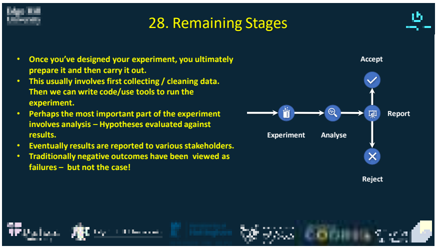
H_a or H_1= Effect on sales

Independent Variable = *Targeted ads per user*

Dependent Variable = *Value of sales per user*

Control variables *e.g. time of day, user age group etc*

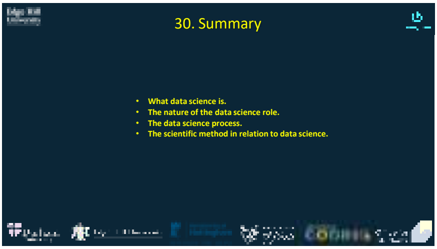
27. Experiment Design



28. Remaining Stages

- Once you've designed your experiment, you ultimately prepare it and then carry it out.
- This usually involves first collecting / cleaning data. Then we can write code/use tools to run the experiment.
- Perhaps the most important part of the experiment involves analysis – Hypotheses evaluated against results.
- Eventually results are reported to various stakeholders.
- Traditionally negative outcomes have been viewed as failures – but not the case!

29. Case Studies



30. Summary

- What data science is.
- The nature of the data science role.
- The data science process.
- The scientific method in relation to data science.