irds

Welcome

to the house of fun.

Welcome

to the lions den.

Sung by whom?

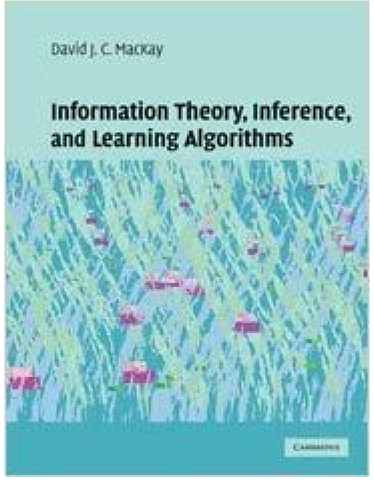
Why What How?

- Data Science Research is moving faster than we have ever seen.
- Top Data Science Research requires incredible breadth and depth of knowledge, understanding, experience and application.
- You have been accepted to CDT Data Science because of the potential you have in this field.
- IRDS is about taking the first steps to make that potential real…

Data Science

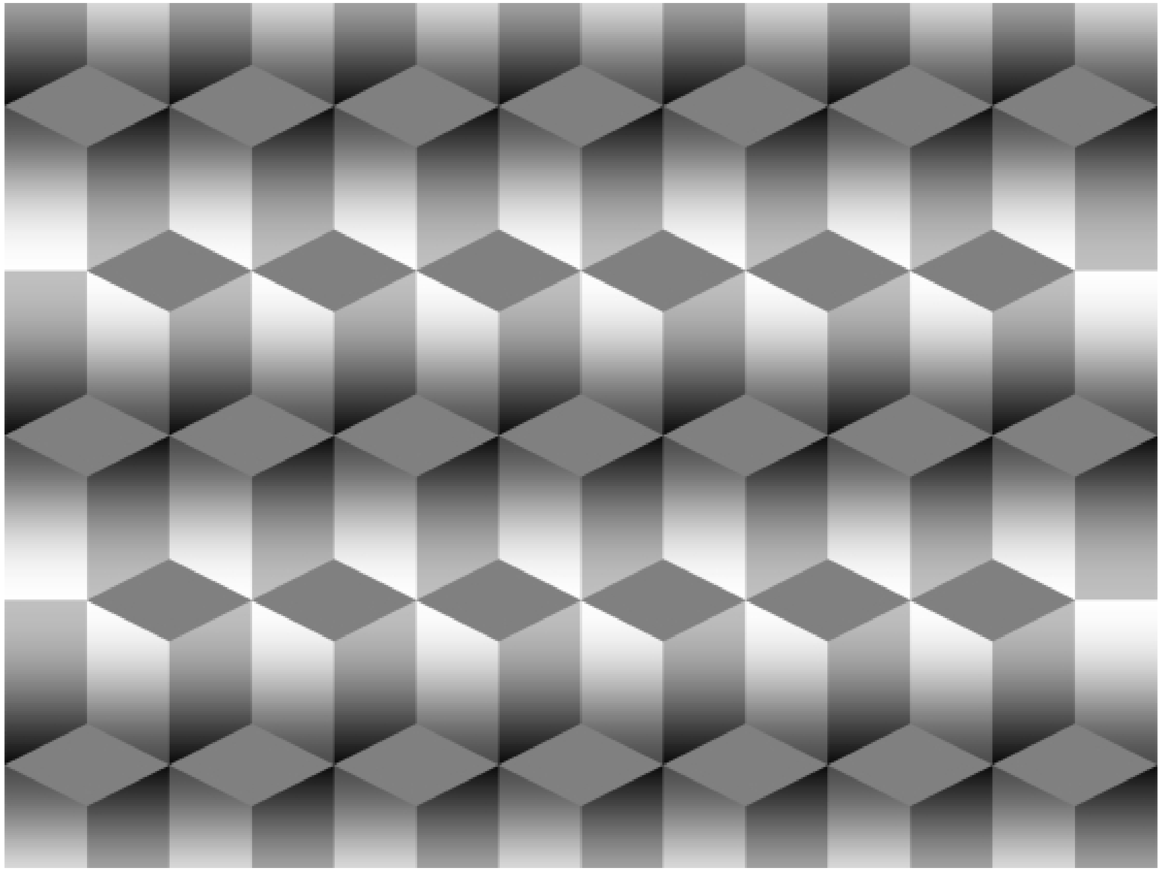
- has theory of information intuition at heart.

- Machine Learning: information coding
- Databases: information efficiency and accessibility
- Data collection: information sourcing
- Inference: combining prior information and data information.



- Mackay: Information Theory, Inference and Learning Algorithms. Good resource.

An aside: Priors in the Brain

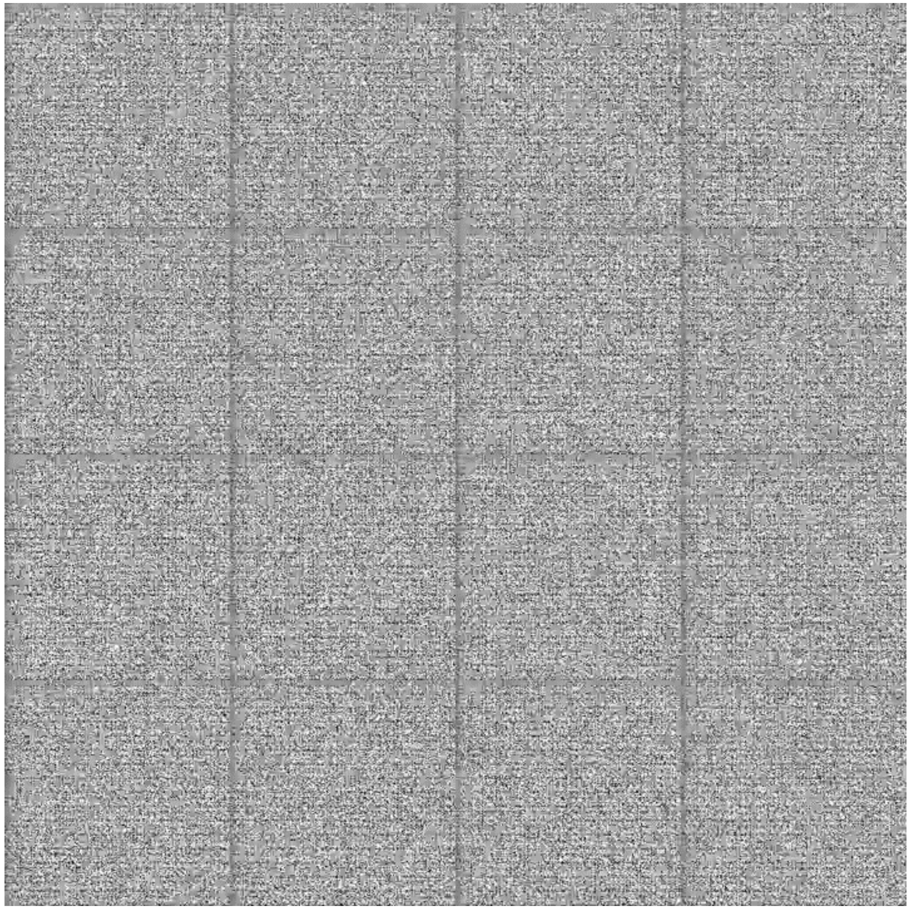


Logvinenko Illusion

GARDNER

Data Science…

- …is more than just cool machine learning.



Benjamin Meier

It is about making
cool methods actually
useful in reality not
just contrived situations.

Today

- A simple quiz.
- I’ll present a problem.
- You write down your answer.
- You discuss it with your neighbour(s).
- We discuss it as a class.
- Next problem….

- At end, I will tell you what I think.

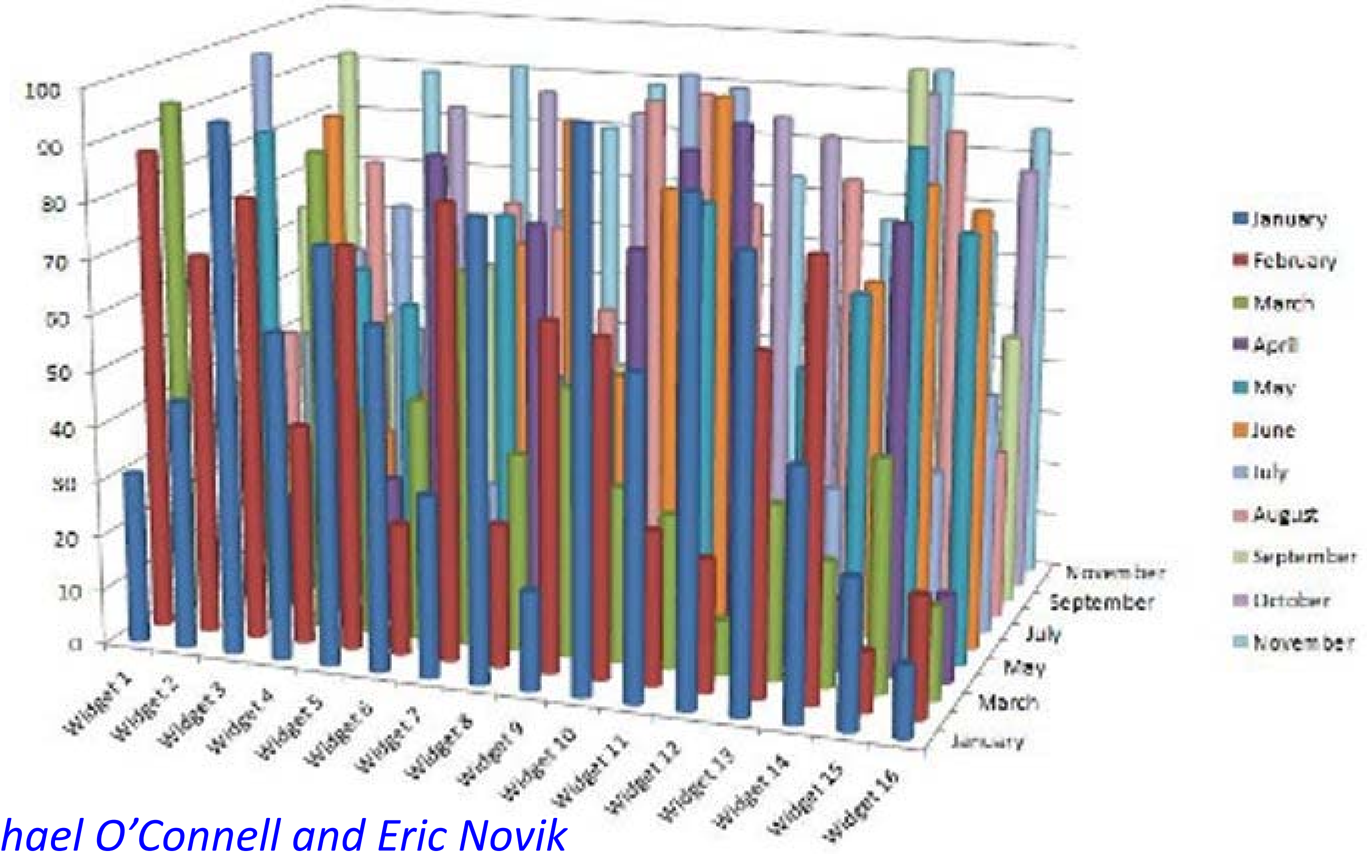
Question 1

A potential client comes to you with an interesting proposition. They will be getting some data which will provide a large array of cell counts from repeated in-vitro timed trials of different cancerous cells under the treatment of different cancer drugs. Some drugs will suppress cell growth for a period, but then it returns. Others are slower acting.

They are interested in analysing this data to determine the contribution of each drug, model the effects of each drug and to help determine the optimum mixture of drugs for given cancerous cell lines. They are interested in if you would join them to help them in this project as a consultant.

Q: What is the most important thing for you to note from the above as a potential consultant?

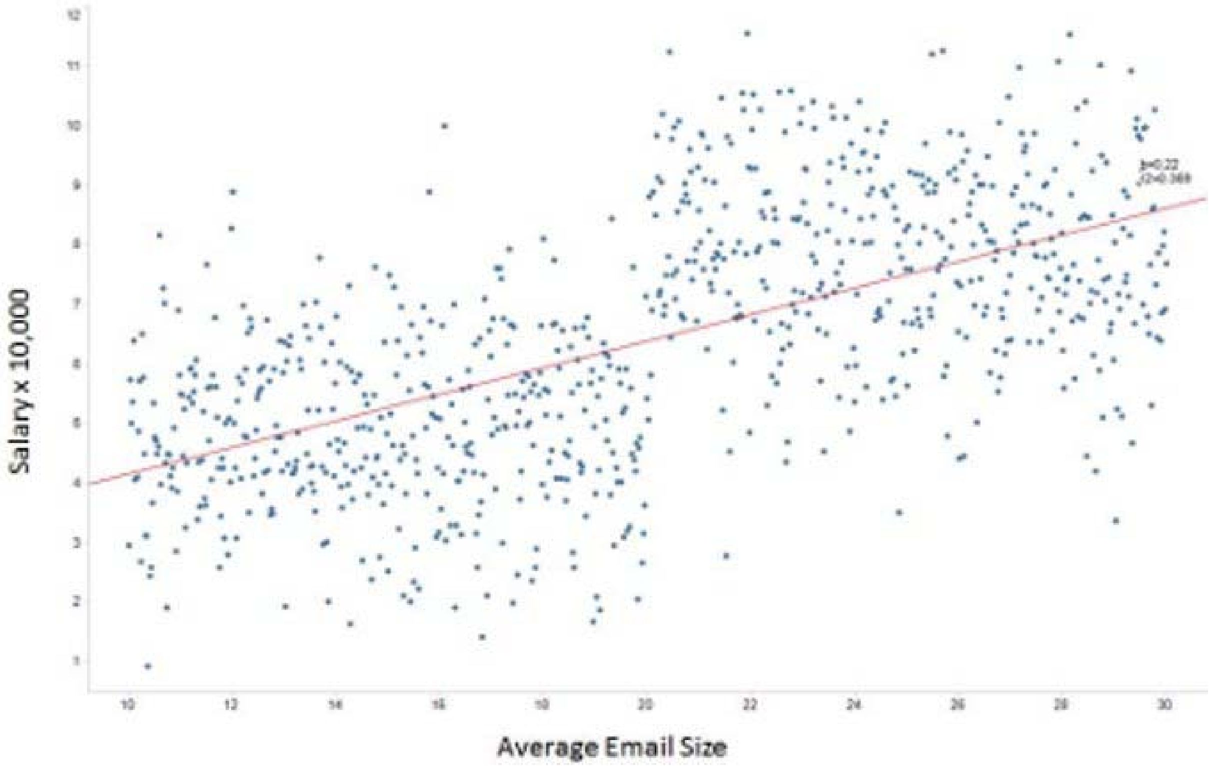
Question 2



From Michael O'Connell and Eric Novik

- The above visualisation of the sale of widgets over the year is appalling. Why? How would you do it better?

Question 3



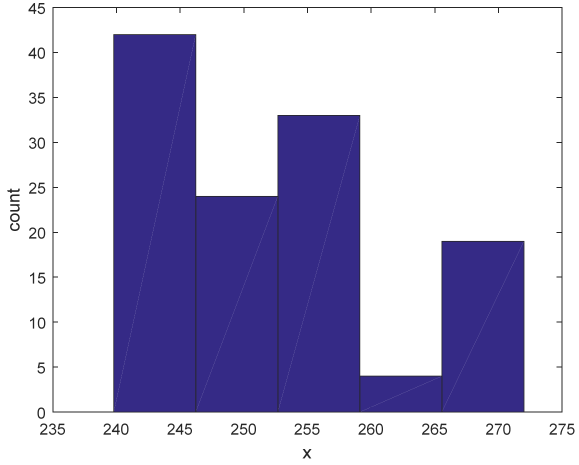
From Michael O'Connell and Eric Novik

- The above graph displays (hypothetical) data plotting email size against salary. What comments do you have?

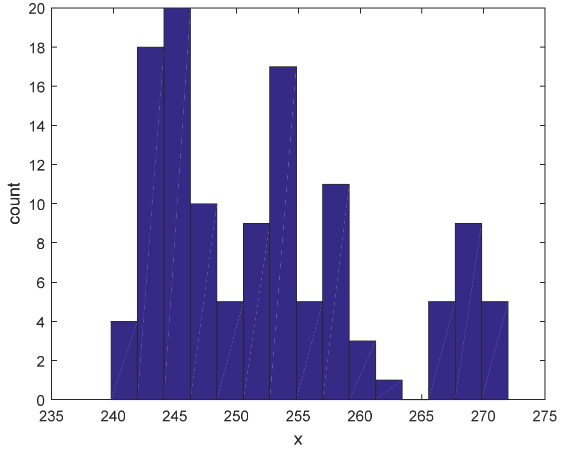
Question 4

- You are plotting a histogram of some data as part of exploratory data analysis. Which of (a), (b), (c) or (d) should you choose to plot?

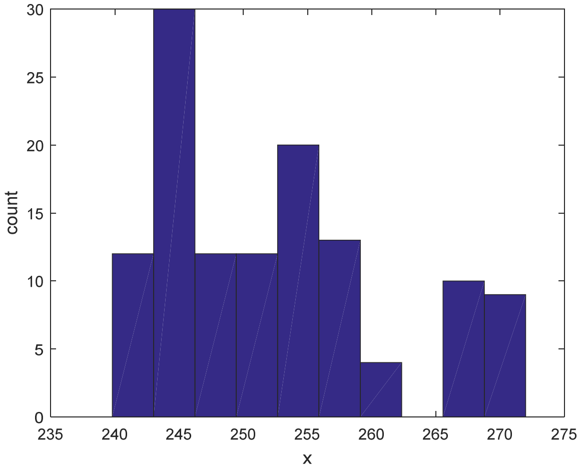
a



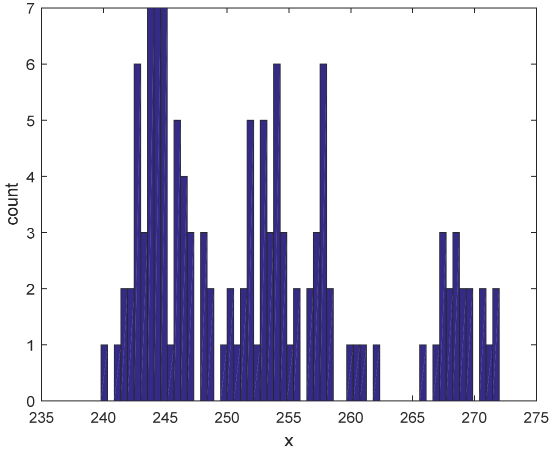
c



b



d



Question 5

- You are doing a predictive analysis of multivariate time series data, predicting the 10 future time steps from the past. The data has substantial missing information for some variables. You are using a model that required complete data (e.g. a neural network)
- How do you fill in for the incomplete data to use it? How do you change what you learn to account for it?

Question 6

- You have a massive database that you require to index on a variable with a 3D coordinate.
- You will have queries of the sort: find the nearest neighbour, find the k nearest neighbours, find all data points within a radius.
- How should you organise your database.

Question 7

- The papers announced that eating a bacon sandwich every day increased the risk of bowel cancer by 20%.
  Can you now suggest a more useful way of describing the increased risk?

Thanks: David Spiegelhalter

Question 8

- You have to present a case to an NHS assessment board, as to whether an excess death rate at a particular city centre hospital is a concern.
- What sorts of analyses do you want to be sure you have done before making your case?
- What simple argument would you want to use to make your case?

Question 9

- You have a data-driven method of managing power in a datacentre that will save a power bill of 10%
- You want to persuade a client to use this method.
- What questions will the client want to ask?

Question 10

- You are planning a deep learning research project. You already have the data, splits into training, validation and test sets. You are going to arrange the project in 5 stages. What should those 5 stages be?

Question 11

- You are writing a data science paper exploring the development of a new technique and its application on a specific problem.
- It will have 8 section headings. What should they be, and in what order?

Summary

- You will learn many technical details on many of the courses around you.
- But being a good Data Science Researcher is much more than knowing the tech.
- This course is about
  - the *nous* about what to do
  - Gauging the right information about a problem
  - Explaining/presenting things to people
  - Organisation and making good choices
- We explore the broader scene for datascience research.
- See DME lecture notes for the supplementary info:
- http://www.inf.ed.ac.uk/teaching/courses/dme/2017/lecture-notes.pdf