ECMM443/COM2015
Introduction to Data Science

Dr Xiaoyang Wang

Department of Computer Science

x.wang7@exeter.ac.uk

Assessments

Module information page: Introduction to Data Science - 2024 entry

The assessments are in two parts:

- Coursework - 20%: Data analysis practice using Python; The coursework document will be published at least 4 weeks before the deadline; **Deadline: ∼Week 9**
- Coursework submissions should be **anonymous** - please put your student number, but not your name.
- Exam - 80%

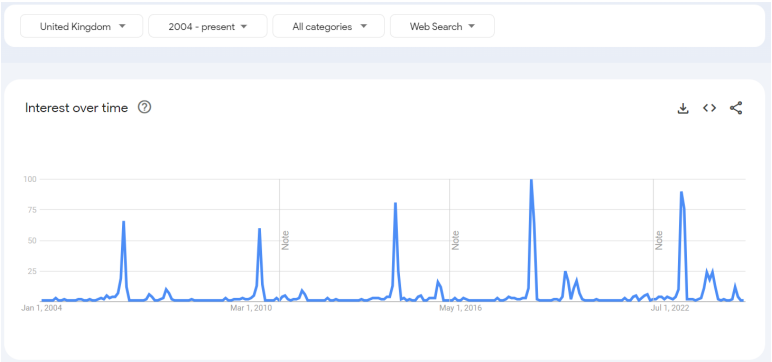
Look at Some Data

Google Trends: `https://trends.google.com/trends/`

Look at Some Data

Google Trends: `https://trends.google.com/trends/`

Search for "World Cup"



Look at Some Data

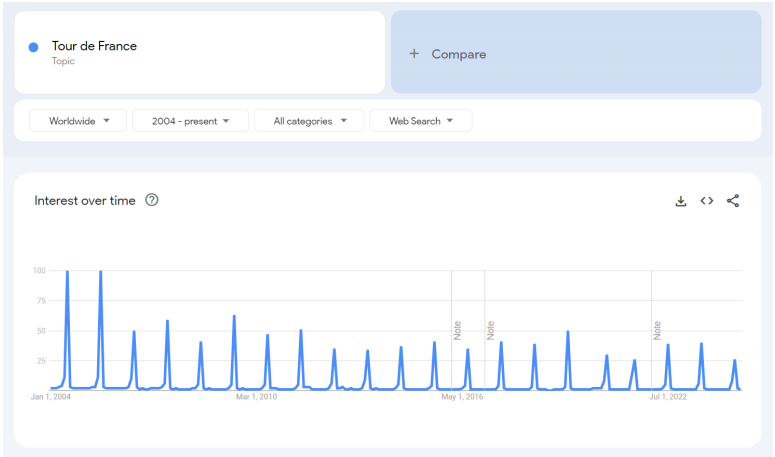
Search for "Tour de France"

Stage 21

https://www.happy.rentals/blog/323-tour-de-france

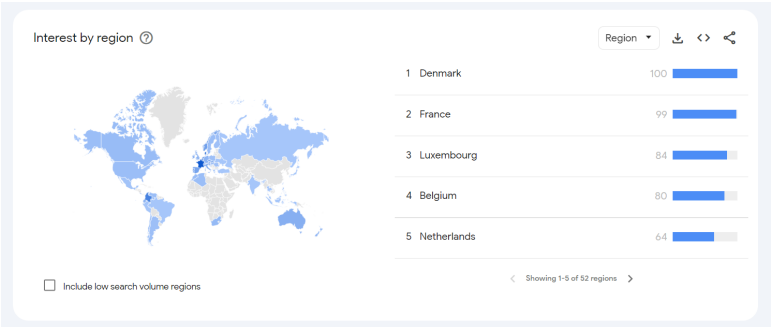
Look at Some Data

Search for “Tour de France”



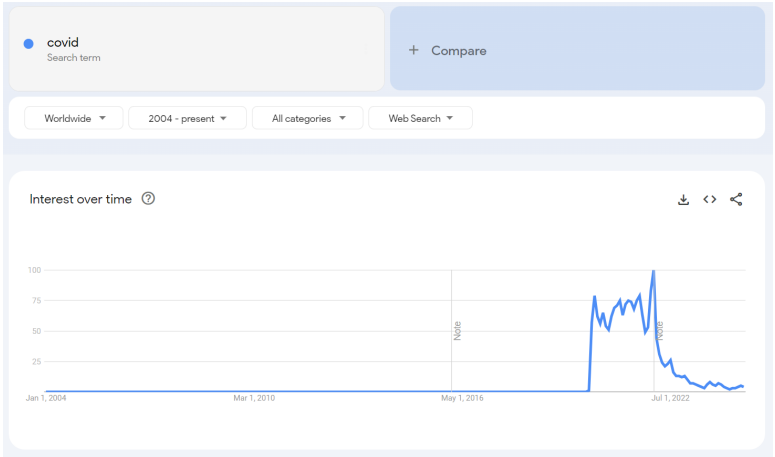
Look at Some Data

Search for “Tour de France”



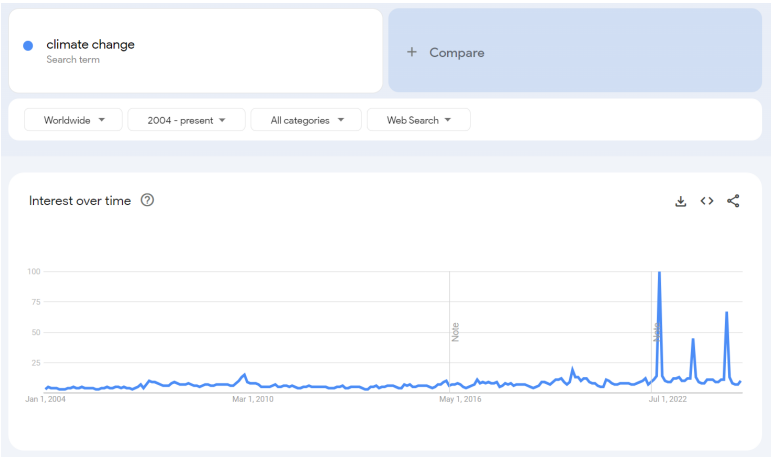
Look at Some Data

Search for “Covid”



Look at Some Data

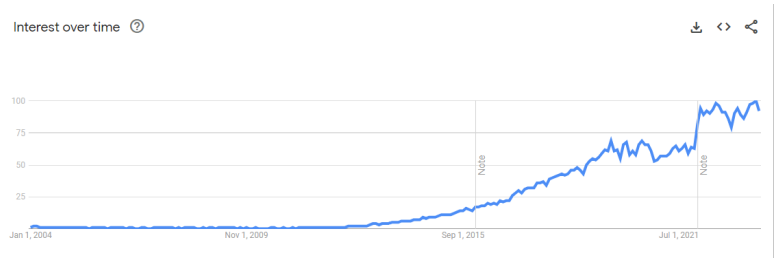
Search for “Climate change”



Look at Some Data

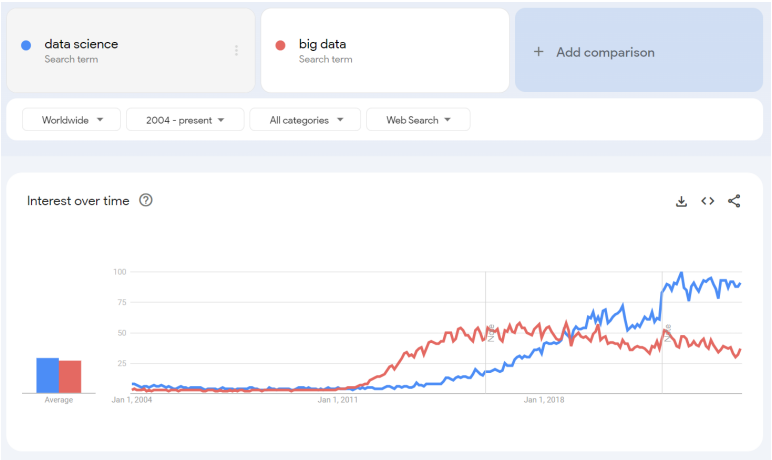
University of Exeter

Search for "Data Science"



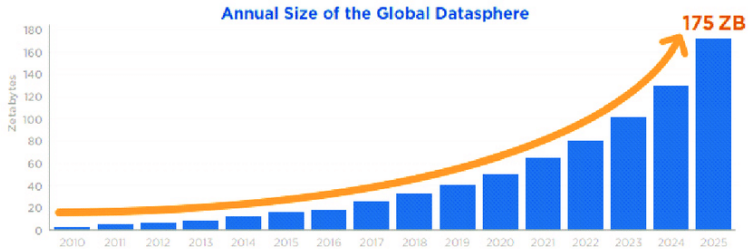
Look at Some Data

Compare “Data Science” and “Big Data”



Big Data

Annual Size of the Global Datasphere



$1\ \text{ZB} = 10^9\ \text{TB}$

A 4TB hard drive is $\sim$ £90

$1\text{ZB} \approx$?

https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Big Data

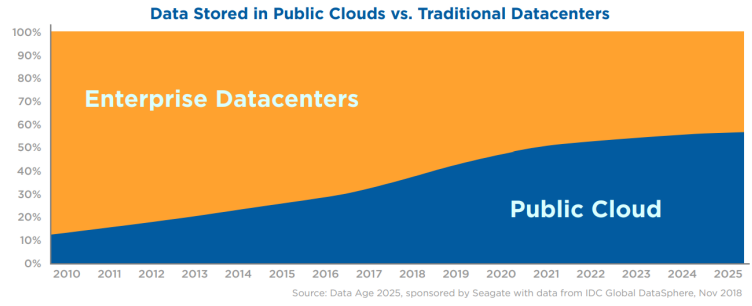
University of Exeter

What is it?
In a single day ...

- ∼500 million Tweets
- ∼720,000 hours of video on Youtube
- ∼5.6 billion Google searches
- ∼350 million Amazon Sales (in the US)
- ∼258 million active users in Weibo (daily) (2023, Q2)
- ∼300 billion emails are sent
- That's ∼100 trillion emails per year!

https://www.statista.com/statistics/1058070/china-sina-weibo-dau/

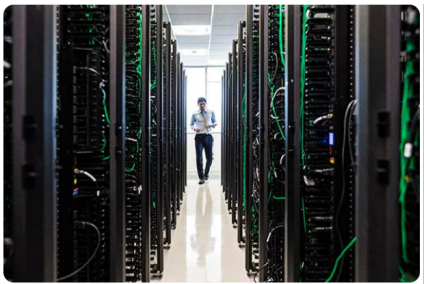
Public Clouds vs Private Sectors



https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Computer Science/IT: Data Center

A data centre is a physical facility that organisations use to house their critical applications and data. The key components of a data centre design include routers, switches, firewalls, storage systems, servers, and application-delivery controllers.



https://www.cisco.com/c/en_uk/solutions/data-center-virtualization/what-is-a-data-center.html

Sustainability

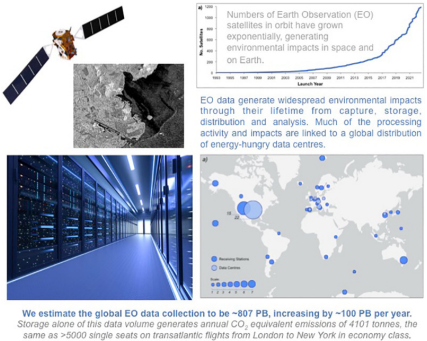


Figure: Earth Observation (EO) data storage alone generates annual CO_2 emissions equal to 41,000 individual London-Paris airplane journeys.

Wilkinson, R., Mleczko, M.M., Brewin, R.J.W., Gaston, K.J., Mueller, M., Shutler, J.D., Yan, X. and Anderson, K., 2024. Environmental impacts of earth observation data in the constellation and cloud computing

Example: The "Green Algorithms" Project

University of Exeter

"Green Algorithms": `https://www.green-algorithms.org/`

- "Carbon intensity will vary considerably depending on the type of data centre, and with the specific geographical location of the server used."
- Data can be processed on the cloud, generating widespread environmental impacts.
- Be transparent about the environmental impacts of big data.

Think about LLMs?

Lannelongue, L., Grealey, J. and Inouye, M., 2021. Green algorithms: quantifying the carbon footprint of computation. Advanced science, 8(12), p.2100707.

LLMs, Carbon Footprint

Published as a conference paper at ICLR 2024

LLMCarbon: Modeling the End-To-End Carbon Footprint of Large Language Models*

Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Osi†, Prateek Sharma, Fan Chen, Lei Jiang
Indiana University †Jackson State University
{afaiz,skaneda,ruhwang,prateeks,fc7,jiang60}@iu.edu
†j00967039@students.jsums.edu

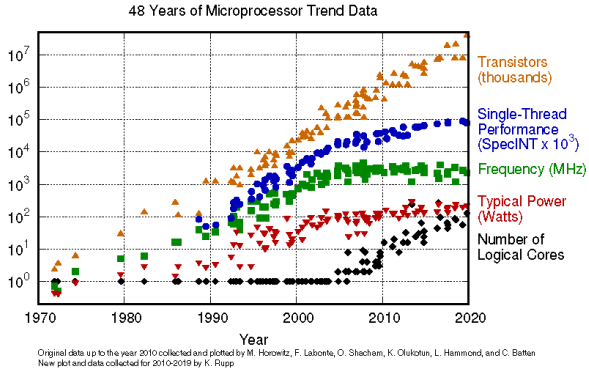
Abstract

The carbon footprint associated with large language models (LLMs) is a significant concern, encompassing emissions from their training, inference, experimentation, and storage processes, including operational and embodied carbon emissions. An essential aspect is accurately estimating the carbon impact of emerging LLMs even before their training, which heavily relies on GPU usage. Existing studies have reported the carbon footprint of LLM training, but only one tool, mlco2, can predict the carbon footprint of new neural networks prior to physical training. However, mlco2 has several serious limitations. It cannot extend its estimation to dense or mixture-of-experts (MoE) LLMs, disregards critical architectural parameters, focuses solely on GPUs, and cannot model embodied carbon footprints. Addressing these gaps, we introduce *LLMCarbon*, an end-to-end carbon footprint projection model designed for both dense and MoE LLMs. Compared to mlco2, LLMCarbon significantly enhances the accuracy of carbon footprint estimations for various LLMs. The source code is released at https://github.com/SotaroKaneda/MLCarbon.

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F. and Jiang, L., 2023. LLMcarbon: Modeling the end-to-end carbon footprint of large language models. arXiv preprint arXiv:2309.14393.

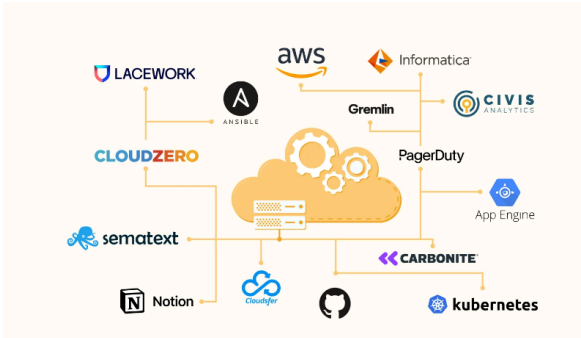
Computational Power and Moore's Law

Moore's law is the observation that the number of transistors in an integrated circuit (IC) doubles about every two years, in 1965.



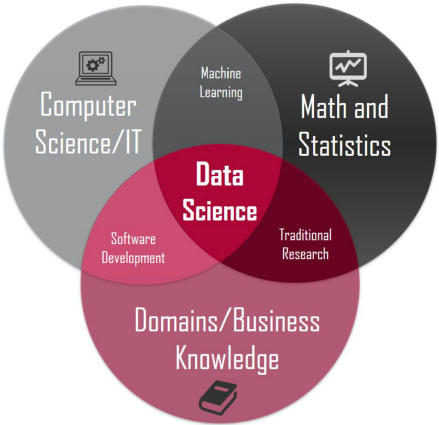
https://www.semianalysis.com/p/a-century-of-moores-law

Cloud Services



https://www.cloudzero.com/blog/cloud-computing-tools

Data Science



This figure is from Dr Rudy Arthur's slides

Statistics

University of Exeter

In this module, we will learn

- Linear Regression
- Hypothesis Testing
- Dimensionality Reduction
  - PCA
- Clustering
  - K-means
- Graph Theory

Some old but very effective (and reliable) approaches!

Machine Learning

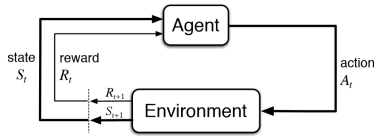
Some famous machine learning tasks

- Recognise handwritten digits
- Find anomalies in big data
- Speech to text
- Robot auto-control

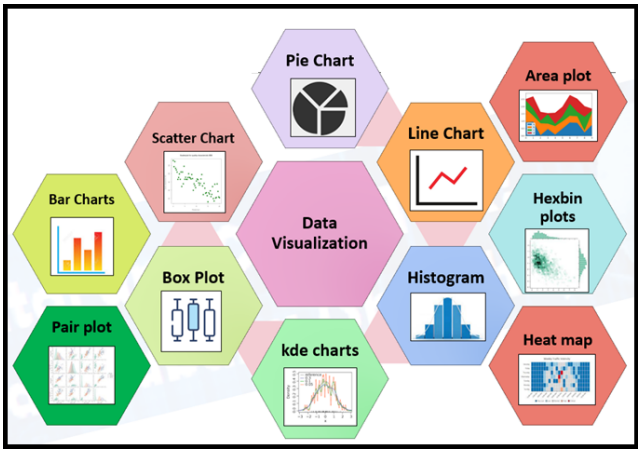
How do we get the data?

Yann LeCun's cake

- cake: unsupervised learning
- icing: supervised learning
- cherry: reinforcement learning



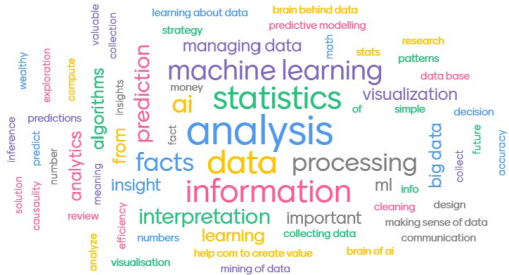
Visualisation



https://www.analyticsvidhya.com/blog/2021/08/effective-data-visualization-techniques-in-data-science-using-python/

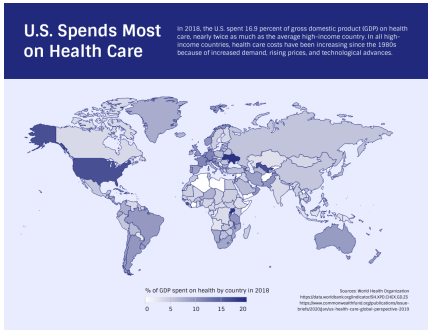
Visulisation

Word clouds



From lecture 1, 'how would you define data science'

Visualisation: Maps



A Choropleth map is a statistical thematic map that uses pseudocolour, meaning colour corresponding with an aggregate summary of a geographic characteristic within spatial units.

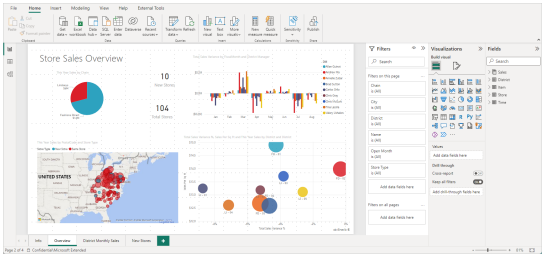
Q: Is it always a good way to visualise?

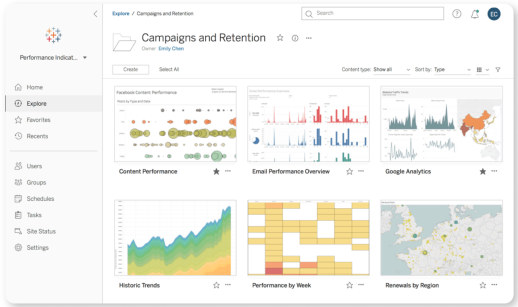
https://venngage-wordpress.s3.amazonaws.com/uploads/2022/05/United-States-Health-Care-Spending-Map-Chart-Template.png

Visualisation, Interactive

There are different tools including

- PowerBi
- Tableau
- and much more





What is Data Science

Next lecture: Matplotlib