HDAT9400 Data Management: S, M, L, XL Data

Malcolm Gillies

https://github.com/mbg-unsw/hdat9400

19 October 2021

UNSW SYDNEY

CENTRE FOR
BIG DATA RESEARCH
IN HEALTH

main@d5b2252 2021-10-11

Lecture outline

About me

- When I studied computer science (1990), a PC had
  - 4MB RAM (1000th today’s phones)
  - 200MB Disk (1000th today’s phones)
  - 33MHz Processor (100th today’s phones)

Health data I’ve worked with

- NPS MedicineWise: GP electronic medical records (MedicineInsight)
- NSW Ministry of Health: hospital, ambulance, births
- CBDRH: Pharmaceutical Benefits Scheme (PBS)
- SAS, R, MS SQL Server, PostgreSQL, SQLite, DuckDB

Why does data size matter?

- Time and space are finite
- We have budgets and deadlines
- Two times bigger can take more than twice the time

What can we do about it?

- Work smarter, not harder
- Relax, people have been thinking about this for a long time!

How big are health data sets?

Data	Records	Gigabytes
NSW congenital conditions (5 years)	10 000	0.001
NSW perinatal (20 years)	1 000 000	1
NSW Admitted patients (20 years)	100 000 000	15
AU Pharmaceutical benefits (20 years)	1 000 000 000	400
XXXX Data Lake??		

20–200 variables per record

Examples of different data processing technologies

Method	**Max size**	**Rec per sec**	**Notes**
In memory [R]	xx	xx	Simple!
Disk streaming [SAS]	1TB	xx	Slower
Relational database [PostgreSQL]	1TB	xx	Complicated
Column-store database [DuckDB]	1TB	xx	Specialised
NoSQL [Apache Spark]	????	????	Don’t ask

Starting simple: process all the data

- Sometimes you need to look at every record aka *table scan*
    - e.g. What is the total length of stay of all NSW admissions?
- All else being equal, twice the data takes twice the time
- Most important distinction: scan in memory (RAM) or disk?

main@d5b2252 2021-10-11

Making things more complicated

- Sort all prescriptions by date of prescription
- Analyse all data from hospitals in Sydney
- For each antibiotic prescription, find the corresponding doctor visit
- Build a regression model for risk of low birth weight based on maternal characteristics

main@d5b2252 2021-10-11

Experiment: sorting in SAS

- XXXX

Time (and space) complexity

- Asymptotic complexity

Speed of common algorithms

Sort	$\mathcal{O}(n \log n)$
(Binary) search	$\mathcal{O}(\log n)$
Matrix inversion	$\mathcal{O}(n^2 \log n)$
XXXX	

Speeding up WHERE using an index

- XXXX

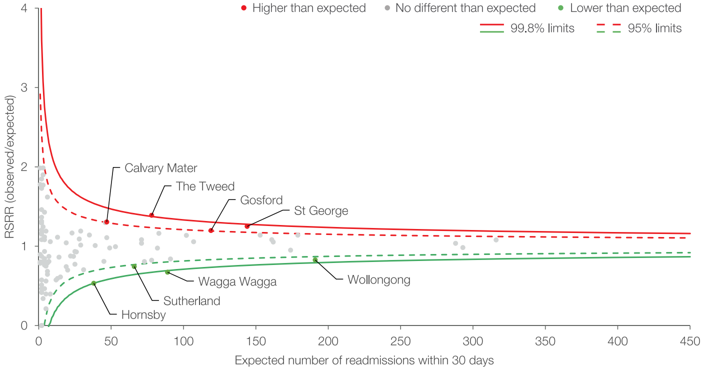
Real world example: NSW hospital readmission rates I

- Bureau of Health Information
- Quarterly report on hospital performance
- Mixed models, SAS
- Run time for the analysis: 1 minute

main@d5b2252 2021-10-11

Real world example: NSW hospital readmission rates II

Acute myocardial infarction 30-day risk-standardised readmission ratio, NSW public hospitals, July 2015 – June 2018



Real world example: C****-19 cases daily reporting

Bonus round: what about *big data*?

- Parallel processing e.g. Google MapReduce

Thanks

- Sadaf Marashi-Pour (Bureau of Health Information)
- Sandy Sa (NSW Ministry of Health)
- Juan Quiroz Aguilera (CBDRH)
- Oisin Fitzgerald (CBDRH)

Further reading

main@d5b2252 2021-10-11

References

main@d5b2252 2021-10-11