

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@5c0722d2021-10-12



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

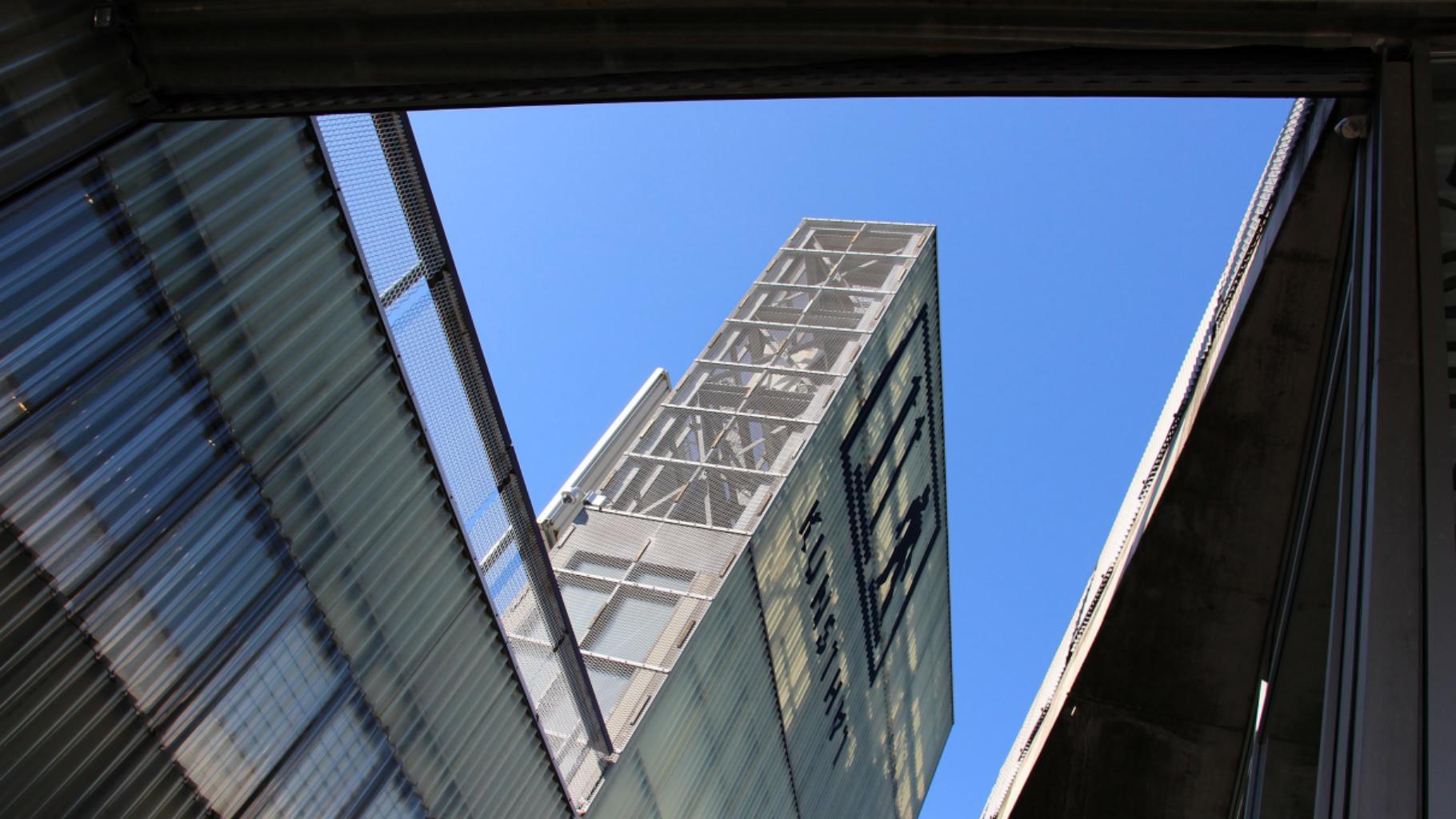
- NPS MedicineWise: GP electronic medical records (MedicineInsight)
- NSW Ministry of Health: hospital, ambulance, births
- CBDRH: Pharmaceutical Benefits Scheme (PBS)
- SAS, R, MS SQL Server, PostgreSQL, SQLite, DuckDB

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!



10:10

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
Health “data lake”	1 000 000 000 000	1 000 000

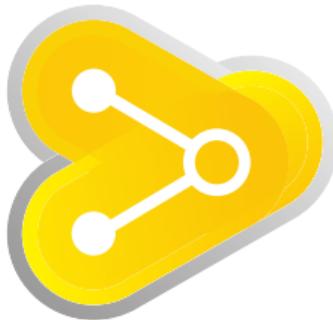
20–200 variables per record

What is “big data”?

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.

— Wikipedia

So...



CENTRE FOR
BIG DATA RESEARCH
IN HEALTH

~~MODESTLY~~ SIZED

main @ 5c0722d 2021-10-12

Examples of different data processing technologies

Method	Max size	Rec per sec	Notes
In memory [R]	GB	25M	Simple!
Disk streaming [SAS]	TB	5M	Slower
Relational database [PostgreSQL]	TB	1M	Complicated
Column-store database [DuckDB]	TB	50M	Specialised
NoSQL [Apache Spark]	PB	????	Don't ask

Starting simple: process all the data

- Sometimes you need to look at every record aka *full 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?

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



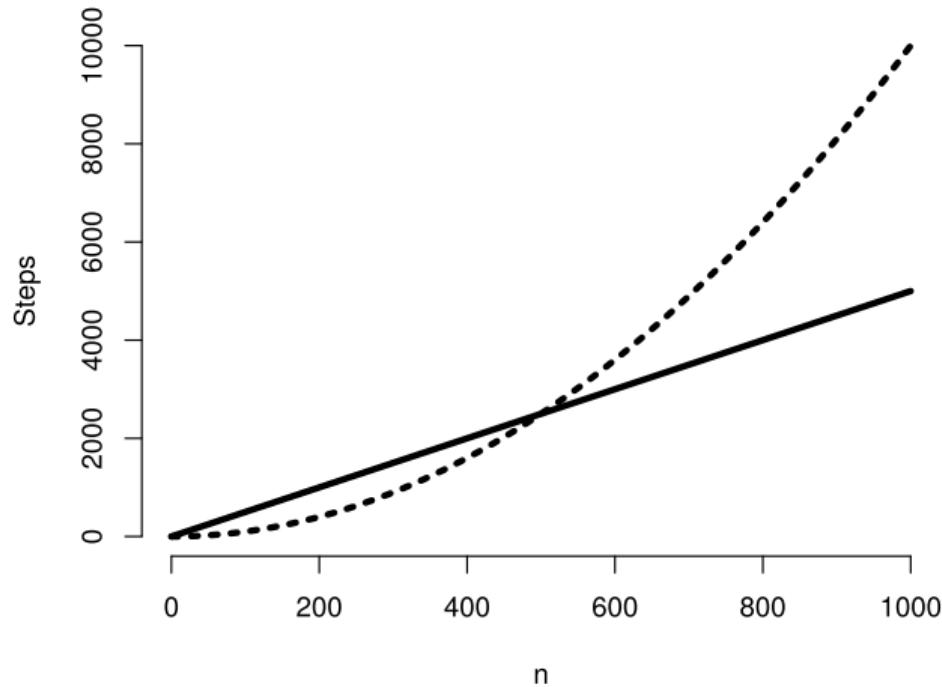
Experiment: sorting in SAS

- XXXX

Time (and space) complexity

- How does an algorithm “scale” with the amount of data?
- Ignore details of the implementation, specific computer etc
- Insight: how does the time/space cost change as $n \rightarrow \infty$?
- *Asymptotic complexity*

$O(n)$ vs $O(n^2)$



Speed of common algorithms

Hash lookup	$O(1)$
(Binary) search	$O(\log n)$
Sort	$O(n \log n)$
Matrix inversion	$O(n^2 \log n)$

Speeding up WHERE using an index

- XXXX

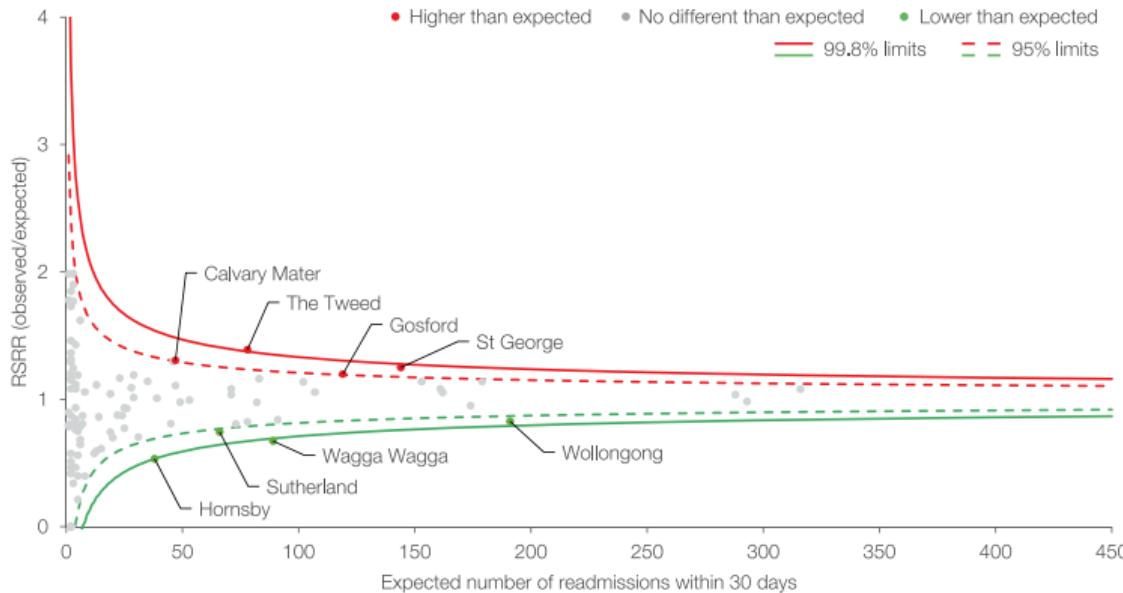


Real world example: NSW hospital readmission rates I

- Bureau of Health Information
- Quarterly report on hospital performance
- Mixed models, SAS
- 32 000 records, 177 hospitals
- Run time for the analysis: 1 minute

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

main@5c0722d2021-10-12

Bonus round: what about *big data*?

- Parallel processing e.g. Google MapReduce

Summary

main@5c0722d 2021-10-12

Thanks

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

Further reading

- A Gentle Introduction to Algorithm Complexity Analysis by Dionysis Zindros
- Database-like ops benchmark by `data.table` developers



Picture credits (Creative Commons)

- Nieuwbouw IJplein by Bart Molendijk / Anefo
- Kunsthal, Rotterdam by Fred Romero
- Seattle Public Library by Moody 75
- De Rotterdam by Fred Romero
- Kunsthal, Rem Koolhaas portrait by Jacco van Giessen