COMP20008 Elements of Data Processing

Peer Review

Data at scale: Distribution, integration and privacy

Announcements

- Anzac Day (25th April) holiday: no workshops
  - Running an additional workshop Alan-Gilbert-111 15:15-17:05 on Thursday 27th April, or
  - May attend one of the other workshops this week
  - Helpful if you can bring your own laptop in case of high attendance numbers

Announcements

- Extension on Phase 2A
  - New Phase 2A submission deadline **09:00am Wednesday 26 April**

- Phase 2B
  - expected to be open shortly after above deadline for Phase 2A closes
  - New Phase 2B: submission deadline **09:00am Monday 1st May**
  - Reviews of your own work then available later that day

Announcements

- Phase 1 marks
  - Were released before Easter break
  - Mark + comments
  - General feedback document
  - Sample solution
  - Contact Donia with any questions

Plan today

- Phase 2B advice
- Data at Scale – introduction

Project Phase 2B

- Provide brief peer review about Phase 2A submissions of two other students

- Is this relevant for data wrangling? What are the objectives?

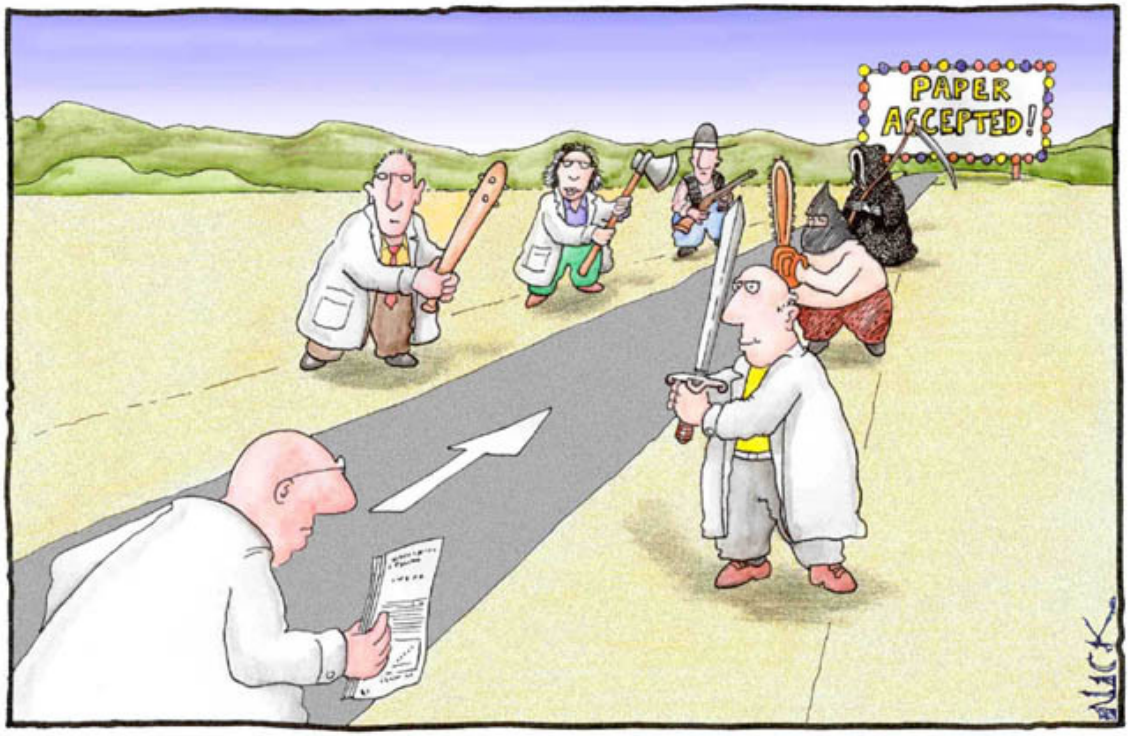
  - *Information sharing*: see what are others in the class are doing in their project. Get insights for your own approach for your Phase 3.

  - *Data scientist training*: Every chunk of functionality you implement may go through multiple rounds of (internal) peer review. Practice the skills of giving and receiving feedback.

My (academic) experience …

Conduct study → Write manuscript → Submit to journal → ***External*** **Peer review**

Accept
Revise
Reject



Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'

Phase 2B: How will it work?

Step 1: Prepare & submit Phase 2A

Step 2: Review two Phase 2A assignments of other students

Step 3: Receive feedback on own assignment. Receive a mark for feedback you provide.

Step 4: Incorporate feedback for your Phase 3

*Process is **"double blind"** (anonymity of giving and receiving feedback)*

Writing feedback

When writing feedback:

- **Aim for balance** – highlight strengths as well as areas for improvement
- **Be specific** – include explanations & examples (question, page, paragraph or line numbers)
- **Prioritise** – attend major issues first (message, structure, organisation) then move onto finer detail if space permits
- **Focus –** on material & content (NOT the writer)
- **Be diligent & respectful** – take care & think about how you would feel if you received the review

Helpful vs. unhelpful feedback

Helpful feedback is:

- ✓ Constructive
- ✓ Specific
- ✓ Balanced
- ✓ Succinct
- ✓ Respectful

vs.

Unhelpful feedback is:

- ✘ Too positive or too negative
- ✘ General & unspecific
- ✘ Rambling
- ✘ Aggressive – makes reader feel ‘attacked’

Helpful vs. unhelpful feedback

1. *What are the main strengths of this report?*

✗ **Unhelpful comment:**

"Your report was really good! I enjoyed reading it."

Author's response: "I'm flattered you liked my report, but I don't have a sense of *what* you thought was good about it."

✓ **Helpful comment:**

"This report was succinct and well written. The aims of the report were clear and I found it easy to identify your take-home messages..."

Helpful vs. unhelpful feedback

2. *Where are the main areas for improvement?*

✗ **Unhelpful comment:**

"Your report was poorly written and hard to read!"

Author's response: "This comment doesn't really help me to improve anything!"

✓ **Helpful comment:**

"There are a few areas that might make this report stronger. Expanding the Introduction to include more background information would help set the scene a little more (para 2). The arguments could also be strengthened by adding additional references, for examples lines 3, 16 and 55..."

Helpful vs. unhelpful feedback

3a. *Is the balance between the sections about right?*

× **Unhelpful comment:**

"No – there wasn't enough space left for covering the background of the study."

✓ **Helpful comment:**

"The balance feels very good; however you may consider the possibility of expanding the background section with greater information on theoretical concepts being tested"

Author's response: "Although stating good and bad points, none of it was portrayed negatively. The comments were given helpfully, with clear points for me to follow."

Helpful vs. unhelpful feedback

3b. *Is the balance between the sections about right?*

✗ **Unhelpful comment:**

"The overall balance was good, with no section out-weighing any other at all."

Author's response: "Very positive review, but not much given that I can improve on - I highly doubt it was almost perfect."

✓ **Helpful comment:**

"Not the best balance: The introduction and rationale sections were too lengthy. While very clear, they could be trimmed down quite a bit and made to be much more concise. For example, I think lines 108 to 113 are unnecessary..."

Helpful vs. unhelpful feedback

4a. *Did you feel the article had good flow and structure?*

× **Unhelpful comment:**

"The paper flows really well from one section to the next and there is a logical progression."

✓ **Helpful comment:**

"It had good flow and structure from paragraphs 1-5, but somewhat lost it's flow from then on. This can be fixed by adjusting the order in which you present your points. For instance, in paragraph 2..."

Author's comment: "Thanks for this comment – it was a good mix of positive comments and suggestions for improvement. It was insightful and helped me improve my paper."

Helpful vs. unhelpful feedback

4b. *Did you feel the report had good flow and structure?*

✗ **Unhelpful comment:**

"The article flowed really nicely and it was easy to follow the author's train of thought"

✓ **Helpful comment:**

"Not the best balance: The introduction and rationale sections were too lengthy. While very clear, they could be trimmed down quite a bit and made to be much more concise. For example, I think lines 108 to 113 are unnecessary..."

Author's comment: "This comment is much more helpful because it gives me specific areas I can improve."

Example review 1

"I like the writing style, and I think the article is relatively easy to follow and the paragraphs are well linked. The article might be stronger if some of the sentences were more simple and succinct such as line 1 and 7 in paragraph 1, and line 3 in paragraph 4."

Specific?

Constructive?

Balanced?

Clear?

Example review 2

"This report has poor structure and flow. There are several grammatical and spelling errors and some of the paragraphs should be shortened. I got confused about what you were trying to say at some points."

Specific?

Constructive?

Balanced?

Clear?

Example review 3

"Some sentences lacked commas where there should have been one, or were too long at times (e.g. line 34 and line 41). Otherwise, the article as a whole had a smooth flow and the intent behind each paragraph clear and understandable."

Specific?

Constructive?

Balanced?

Clear?

Review form

- *Is the proposed question clear? Why or why not?*
  - Is it clear in terms of what is being investigated and why?

- *Does the remainder of the project appear feasible and likely to yield interesting results, in light of the initial investigations? Why or why not?*
  - Consider investigation completed so far, as well as expected value that can be added by further investigation. Authors have been asked to address this directly – do you agree with what they have written – why/why not?

Time management

- Read 2 assignments (thoroughly) = 40min
- Annotate / make notes
- Decide on the good / bad points

} = 40min

- Complete “Review Forms” = 40min

Receiving feedback

When you receive a review:

- **Understand** that reviews will vary in quality
- **Take time** to gather your thoughts & digest the comments
- **Think** about every comment – even if you disagree, consider if it will be an issue for other readers
- **Recognise** the review as an opportunity for reflection & improvement

Receiving feedback: tips

- ✓ Don’t panic!
- ✓ Read **all** the comments & make notes
- ✓ Take time to **reflect**
- ✓ Address any **major** issues for Phase 3. Tackle smaller points as needed

New Topic: *Data at Scale*

Data at Scale: Big Data

- Explosion of sites with huge data needs
  - Google, Facebook, Twitter, Instagram, Tumblr, Youtube, Reddit, Dropbox, Maps. ..
  - Scientific and commercial applications
    - Physics experiments, health imaging data, …
- *High demands on scalability and availability*
- Benefits from opportunities to integrate information
- Risks in regard to privacy

Data at scale: lecture plan

- Data distribution (3 lectures)
  - How do we distribute big datasets?
  - What is Google doing? (guest lecturer Scott Thomson)
  - How can data be distributed so it is verifiable?
    - Blockchain
- Data integration (3 lectures)
  - How do we integrate/link (distributed) datasets?
- Data privacy and ethics (3 lectures)
  - How do we ensure an appropriate balance between utility and privacy?

To realise those Big Data benefits…

- You need to
  - Run analyses as we've described in previous lectures
    - Correlations, predictions, visualisations, etc.
  - Make sure you're aware of sample sizes
  - Draw conclusions as you would with any other data-set

- A difference with Big Data is how you collect and store the data…
  - Split the load across multiple computers

- Tech companies providing a range of solutions to meet the scalability and availability needs

Big Infrastructure

- Cluster Computing (90's and earlier)
  - Set of computers that work together and can be viewed as a single system
- Grid Computing (00's)
  - Collection of computer resources to achieve common goal
  - Modelled on the electricity grid
  - Process grids / Information grids
  - SETI@home was one of the original concepts
- Cloud Computing (10's)
  - Infrastructure as a service (IAAS)
    - Google Compute Engine
  - Platform as a service (PAAS)
    - Google App Engine
  - Software as a service (SAAS)
    - Gmail

Big data: Two key issues

- *Issue 1: Data consistency and availability*
- *Issue 2: Size vs Structure*

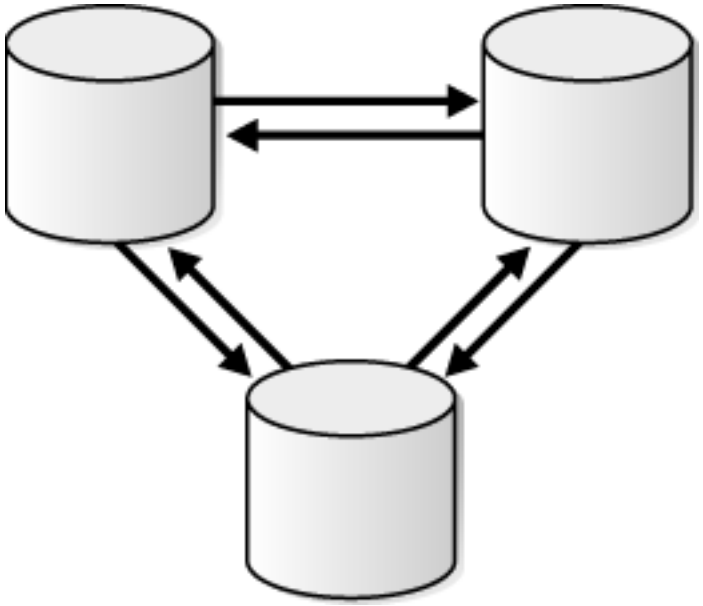
- These are problems that Big Data tries to fix so that large volumes of data can be collected, stored and accessed easily.

- They **DO NOT** impact the science that you are trying to do.

Big data issue 1: Data availability and consistency

- Availability: To make data always available, want to replicate or spread the data (and hence the load) across multiple computers

- Consistency: If data is being updated in parallel across multiple computers, needs smart methods for propagating changes across computers while ensuring consistency of information (same response to queries happening at the same time)



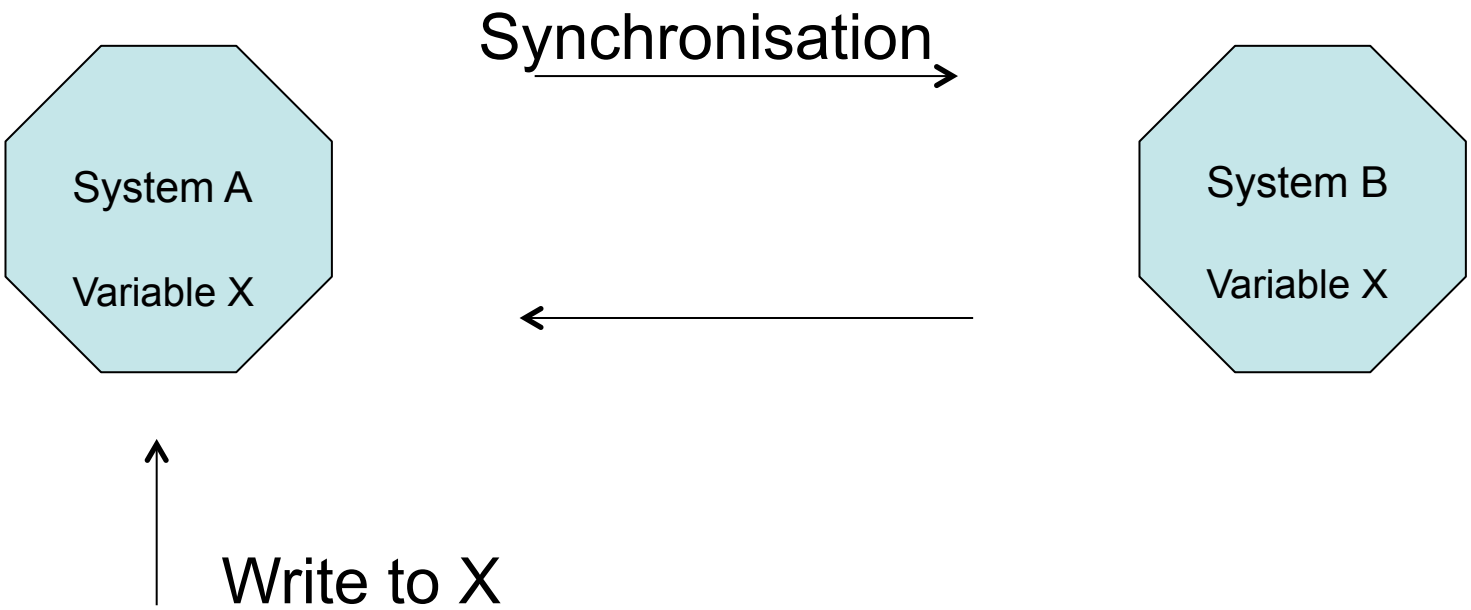
https://docs.oracle.com/cd/E11882_01/server.112/e17516/tdpii_repcont.htm#TDPII065

Availability and consistency

- Requirements
  - Consistency: Each computer always has same view of the data
  - Availability: All computers can always read and write
  - Partition tolerance: System continues to operate in presence of network partitions (failures)
- *We can't achieve all three at the same time!*
  - Known as the "CAP" theorem

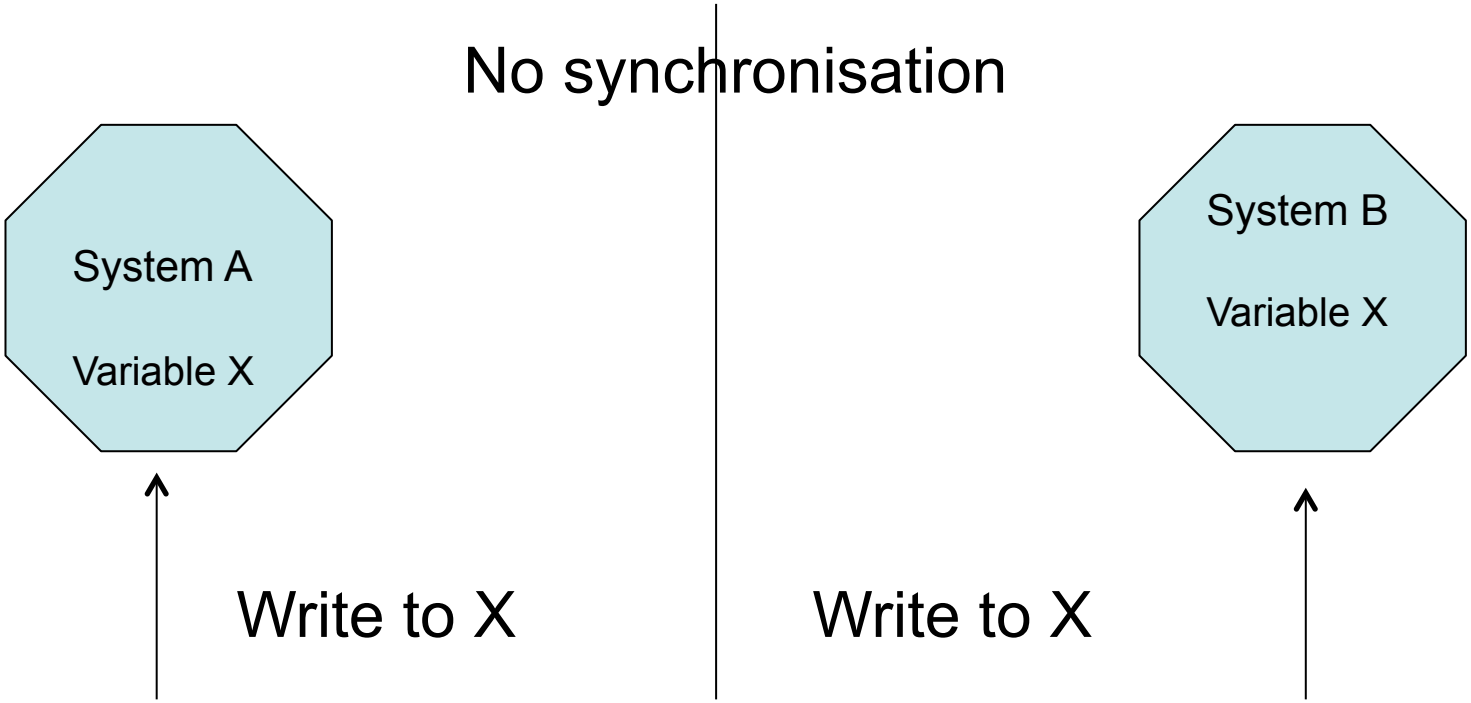
Consistency and availability

- **Consistency: Each computer always has same view of the data**
- **Availability: All computers can always read and write**
- Partition tolerance: System continues to operate in presence of network partitions (failures)



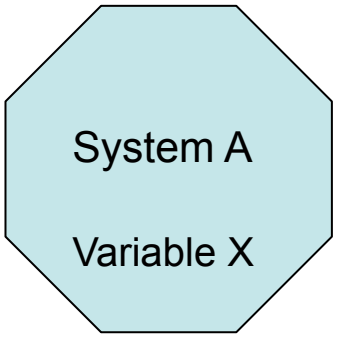
Availability and partition tolerance

- Consistency: Each computer always has same view of the data
- **Availability: All computers can always read and write**
- **Partition tolerance: System continues to operate in presence of network partitions (failures)**

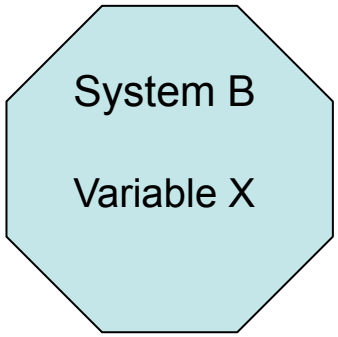


Availability and partition tolerance

- **Consistency: Each computer always has same view of the data**
- **Availability: All computers can always read and write**
- **Partition tolerance: System continues to operate in presence of network partitions (failures)**



Delay writing



Delay writing

Data replication and sharding

- Data Replication
  - System maintains multiple copies of data, stored in different computers, for faster retrieval and fault tolerance.

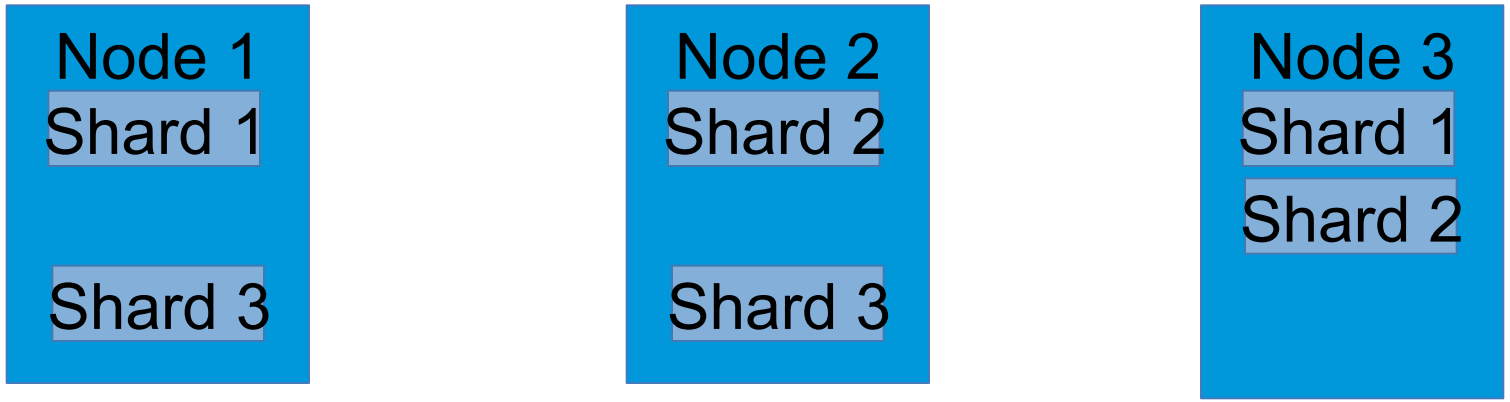
- Data sharding
  - Data is divided (horizontally) into several partitions stored in distinct computers
    - improvement of performance through the distribution of computing load

- Replication and sharding combined
  - Data is divided into multiple partitions across multiple computers; system maintains several identical replicas of each such partition.

Replication and sharding

Replication and sharding can be combined with the objective of maximizing availability while maintaining a minimum level of data safety.

For instance, in a cluster of 3 computers, a replication factor of 2 and a number of shards equal to 3 would look like:



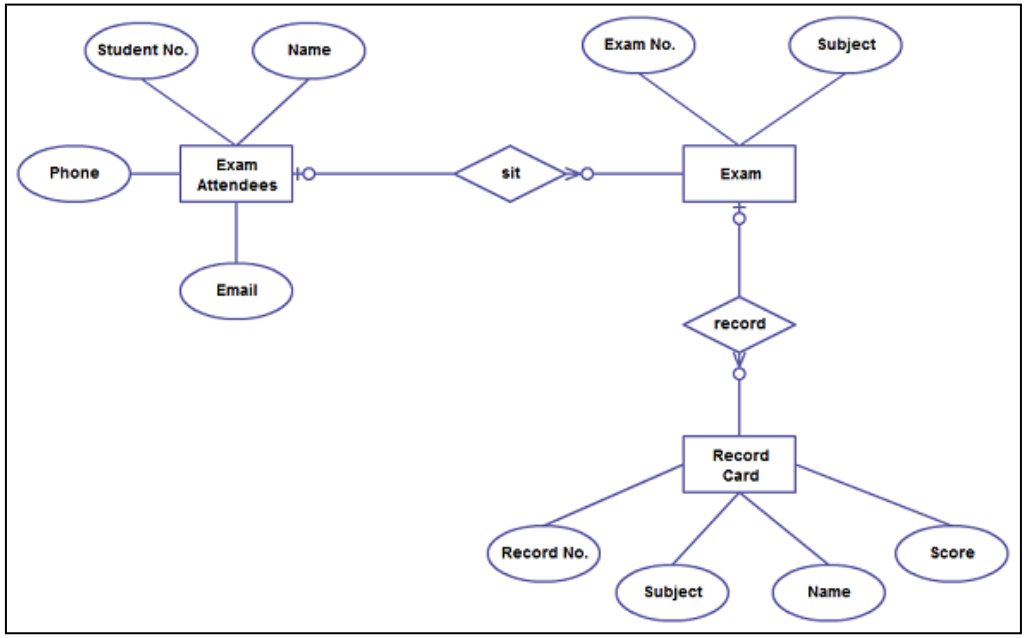
Sharding strategies

- There are different sharding strategies. A key objective is to spread data as evenly as possible across computers, whilst avoiding hotspots (areas of unbalanced demand). E.g.
  - Geographic: customers from a particular geographic region have their data stored on a particular set of computers
  - Range partitioning: customers with ID 0000-1000 on machine A, ID 1001-2000 on machine B, ID 20001-3000 on machine C, etc

Big Data Issue 2: Size vs Structure:

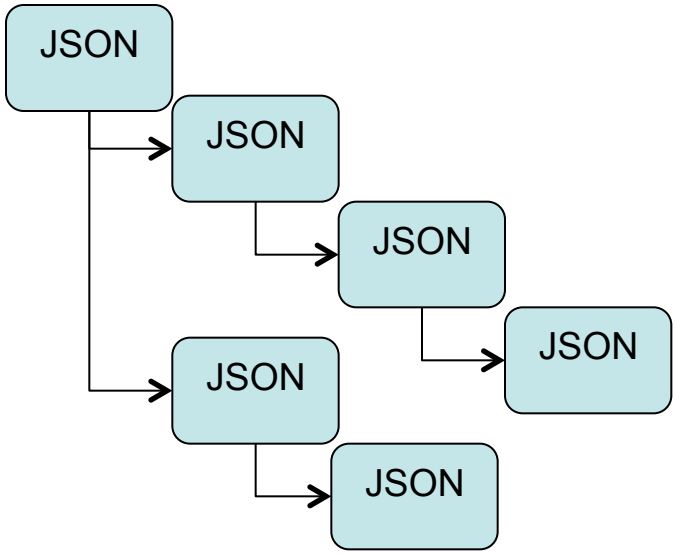
- Can you store all that data in an organised and easily accessible way

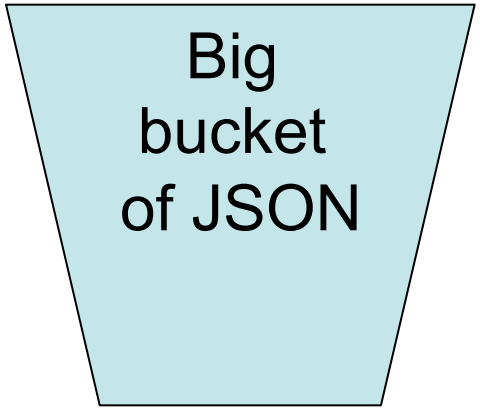
- Strategy 1: We can have a (smaller) structured database organised into tables:
  - Query using SQL (Standard Query Language)
  - Organisation allows optimisation of speed to “finding useful stuff”



Size vs Structure: NoSQL

- Strategy 2: Or we can have a (larger) repository that follows a simple syntax and grows by simple addition:



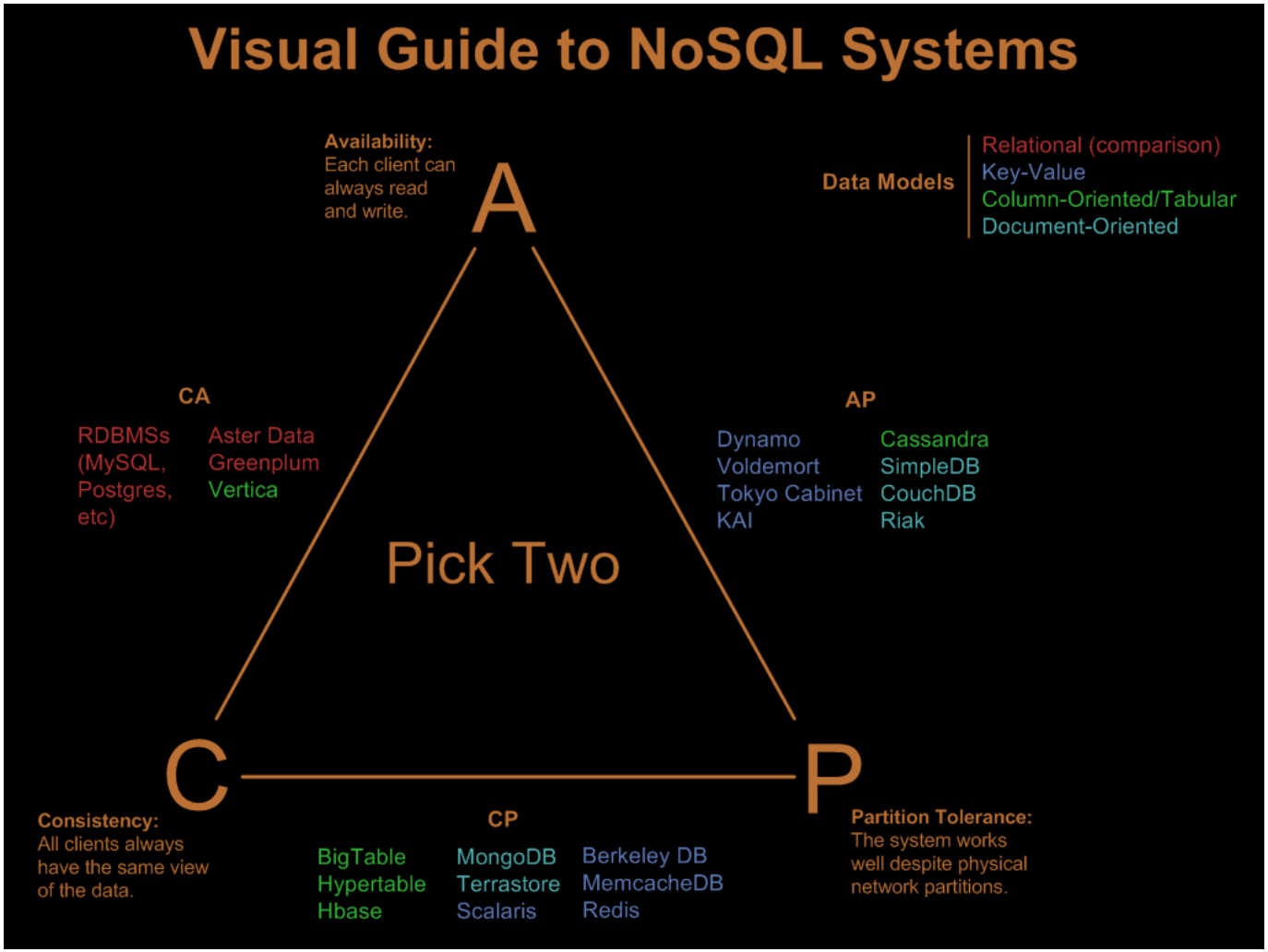


Example NoSQL application: CouchDB

- Coming into fashion at the same time as Big Data:
  - NoSQL databases
- Large-scale data repositories
  - Structure is JSON syntax
  - The emphasis is on scale of storage
  - Does not perform well with random access

- Allows sharding and replication across multiple computers
- Other examples
  - MongoDB Google BigTable and BigQuery, Apache Drill, Amazon Dyamo, Apache Cassandra, …

Landscape of systems



http://blog.nahurst.com/visual-guide-to-nosql-systems

Plan for coming lectures

- Friday 28th: Data linkage and integration
- Monday 1st: Data linkage and integration
- Friday 5th : Guest lecture: Scott Thomson from Google
- Monday 8th: Data linkage and integration
- Friday 12th: Blockchain