INFS4205/7205 Advanced Techniques for High Dimensional Data

Course Introduction

A/Prof. Hongzhi Yin

Semester 1, 2021

The University of Queensland

+ UQ INFS Courses

- INFS1200/7900 Information Systems
- INFS7901 Database Principle
- INFS2200/7903 Relational Database Systems
- INFS3200/7905 Advanced Database Topics
- INFS3202/7202 Web Information Systems
- INFS3208/7208 Cloud Computing
- INFS4203/7203 Data Mining
- **INFS4205/7205 Adv. Techniques for High-Dimensional Data**
- INFS7450 Social Media Analytics
- INFS7410 Information Retrieval

These courses are offered by the **Data Science** Research Group in ITEE
One of the strongest data science groups in the world,
ERA 5 (Well above world standard)

+ About Us

- **Course Coordinator and Lecturer**
  - A/Prof. Hongzhi Yin
  - Director of Master of Computer Science (Management)
  - Field leader of Data Mining and Analysis in Australia
  - Office: 78-639
  - Email: h.yin1@uq.edu.au
  - https://sites.google.com/view/hongzhi-yin/home
- **Lecturer**
  - Dr. Miao Xu (Machine Learning Expert)
  - Email: miao.xu@uq.edu.au
  - Office: 78-633

About Us

- **Tutors:**
  - Mr. Ziyi Liu (78-624, ziyi.liu@uq.edu.au)
  - Miss. Fengmei Jin (78-624, fengmei.jin@uq.edu.au)
  - Miss. Yuting Sun (78-540, skye.sun@uq.edu.au)
  - Any question, seek help from your tutors first.
  - Face-to-face consultation in non-tutorial time: to book an appointment by email
- **Students**
  - INFS4205: 19 (External)+35(Flexible) (BInfTech, BCompSc, BSci…)
  - INFS7205: 93 (External)+65(Flexible) (MCompSci, MDataSc…)

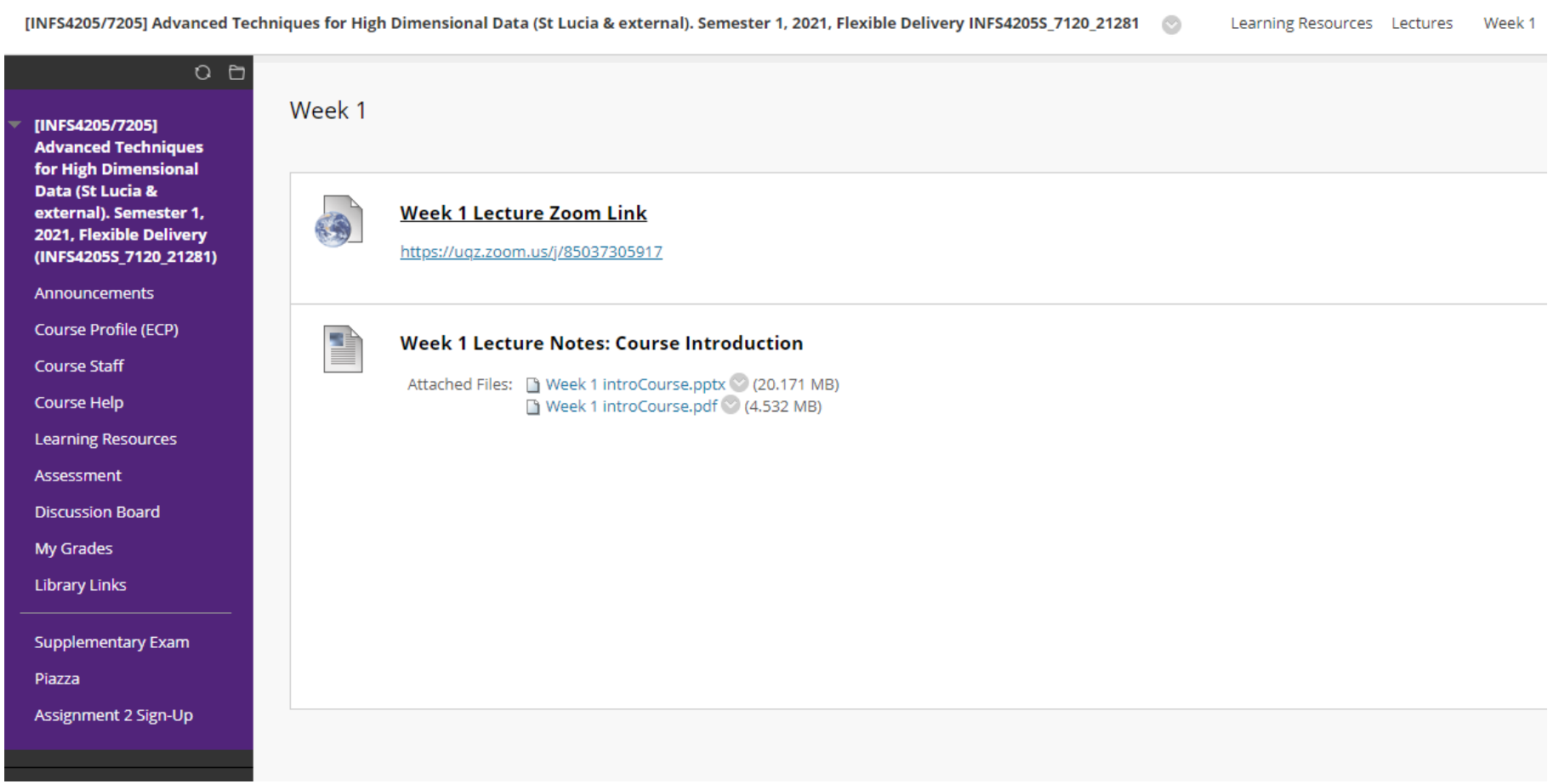
+ Learning Activities

	Lectures	Tutorials
When	Wednesday 4:00-5:50PM	On-campus Session 1: Thu 1:00-1:50PM On-campus Session 2: Thu 2:00 -3:50 PM Online Session 1: Fri 11:00-11:50 AM Online Session 2: Fri 12:00-12:50 PM
Where	Zoom	On-campus Sessions (Yuting Sun): 69-110 Online Sessions (Fengmei Jin): Zoom

- **Tutorials:**
  - Week 2-13, starting from week 2
  - 2 Components in Tutorials
    - Problem Solving/Question Answering
    - Lecture Extension by Demos, Algorithm Examples, Analysis,…

Attendance is not compulsory for both lectures and tutorials although your are encouraged to do so.

+ Blackboard


[INFS4205/7205] Advanced Techniques for High Dimensional Data (St Lucia & external). Semester 1, 2021, Flexible Delivery INFS4205S_7120_21281
Learning Resources
Lectures
Week 1
[INFS4205/7205] Advanced Techniques for High Dimensional Data (St Lucia & external). Semester 1, 2021, Flexible Delivery (INFS4205S_7120_21281)
Announcements
Course Profile (ECP)
Course Staff
Course Help
Learning Resources
Assessment
Discussion Board
My Grades
Library Links
Supplementary Exam
Piazza
Assignment 2 Sign-Up
Week 1
Week 1 Lecture Zoom Link
https://uqz.zoom.us/j/85037305917
Week 1 Lecture Notes: Course Introduction
Attached Files:
Week 1 introCourse.pptx (20.171 MB)
Week 1 introCourse.pdf (4.532 MB)

+ Learning Materials

- No Textbook
- Lecture Notes
- Tutorial Notes
- Research Papers
- Course Forum - Piazza
  - *If you have not received the Piazza invitation email for this course, please contact the tutor Mr Ziyi Liu (*ziyi.liu@uq.edu.au*)*

+ Assessment

- Individual Assignments 20% + 20%
  - A1
    - Spatial Database Design
  - A2
    - Proposal (5%) + Report (15%)
    - Two options
    1. Research Oriented
       - Paper reading and report writing
    2. Engineering Oriented
       - Algorithm implementation and experiment report
- In-class Quiz 10%
  - Multiple Choices
- Closed-Book Final Exam 50%
  - **ProctorU** online invigilated exam
- **In order to pass this course, you need "double pass"**
  - **At least 50 marks in total, and**
  - **At least 25 marks in the final exam**

+ Assessment

- **ProctorU** online invigilated exam
  - https://my.uq.edu.au/information-and-services/manage-my-program/exams-and-assessment/online-supervised-invigilated-exams
  - https://my.uq.edu.au/information-and-services/student-support/proctoru-faq

+ Timeline

W#	Date	Lecture	Tutorial	Assignments
1	24/2	Course Introduction (Hongzhi)		
2	3/3	Introduction to Spatial Databases (Hongzhi)	T1: Introduction to Spatial Databases	
3	10/3	Spatial Data Organisation I (Hongzhi)	T2: Spatial Data Organisation I	A1 out
4	17/3	Spatial Data Organisation II (Hongzhi)	T3: Spatial Data Organisation I I	
5	24/3	Spatial Query Processing I (Hongzhi)	T4: Spatial Query Processing I	
6	31/3	Spatial Query Processing II (Miao)	T5: Spatial Query Processing I I	A2 out
	7/4	No lecture - Mid semester break	No Tutorial - Mid semester break	
7	14/4	Online Quiz (mid-exam)	T6: Quiz Review	14 Apr 21 16:00 - 14 Apr 21 17:30
8	21/4	Managing Spatiotemporal Data (Miao)	T7: Managing Spatiotemporal Data	A1 due (21 Apr 16:00)
9	28/4	Managing High Dimensional Data (Miao)	T8: Managing High Dimensional Data	
10	5/5	Managing Multimedia Data (Miao)	T9: Managing Multimedia Data	A2 Proposal due (05 May 16:00)
11	12/5	(Guest Lecture) Route Planning in Road Network (Mengxuan)	T10: Shortest Path Finding	
12	19/5	Advanced Topic: When AI Meets High Dimensional Data (Miao, not included in the final exam)	T11: Q & A	
13	26/5	Course Review (Miao)	T12: Q & A	A2 Report due (28 May 16:00)
		Revision Period		
		Examination Period		

+ Pre-Requisite

- INFS2200/7903: Relational Database Systems
  - What is DBMS?
    - Databases (schema with constraints + data)
    - Management Software
    - Applications
  - What is SQL?
    - Declarative, Non-procedure
  - Indexing and Processing Optimization
    - B-Tree, Hash, and Bitmap
    - Push down the selection
      - Reduce the intermediate data size, reduce the disk visiting time
- COMP3506/7505 - Algorithms & Data Structures
  - Tree
  - Graph
  - Dynamic Programming

+ Database's Spirit

- Database is not only about storage!
  - How to retrieve and use the data
- View everything in this world as data
  - Digitalizing everything with IoT and Web techniques
  - Big data in digital world
- View every problem in this world as query
  - Computing with algorithm
    - Bounded by the time complexity, boomed by data size
  - Precomputed
    - Extremely fast data retrieval, huge data size
  - How to answer the problem efficiently with less storage cost?
    - Index

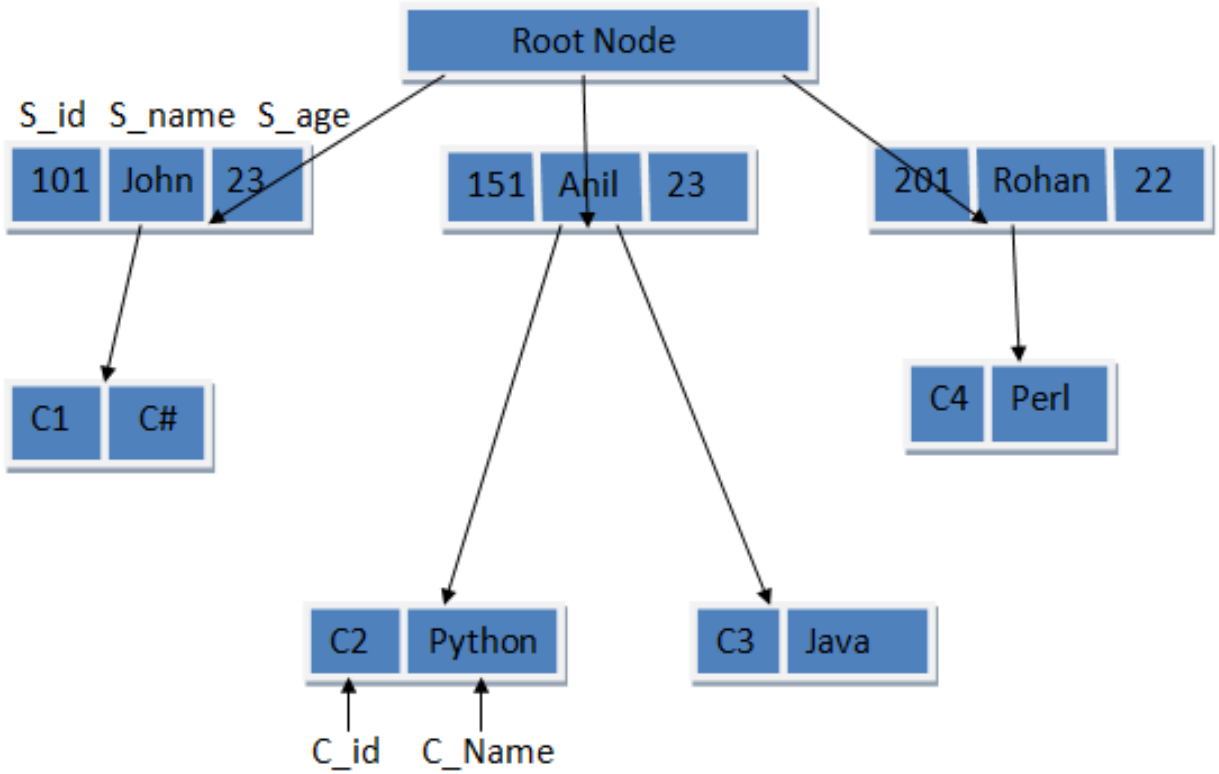
+ Data Model

- Model: Different ways to achieve a goal
  - Computation Model
    - In-memory, External memory, Distributed computing…
  - Machine Learning Models
    - Bayesian Model, Hidden Markov Model, Neural Network Model…
  - …
  - Data Models
    - A data model is the way we represent and manipulate data in a database.
    - For people using a database, the data model describes how we interact with the data in a database.

+ 60s: Hierarchical Model

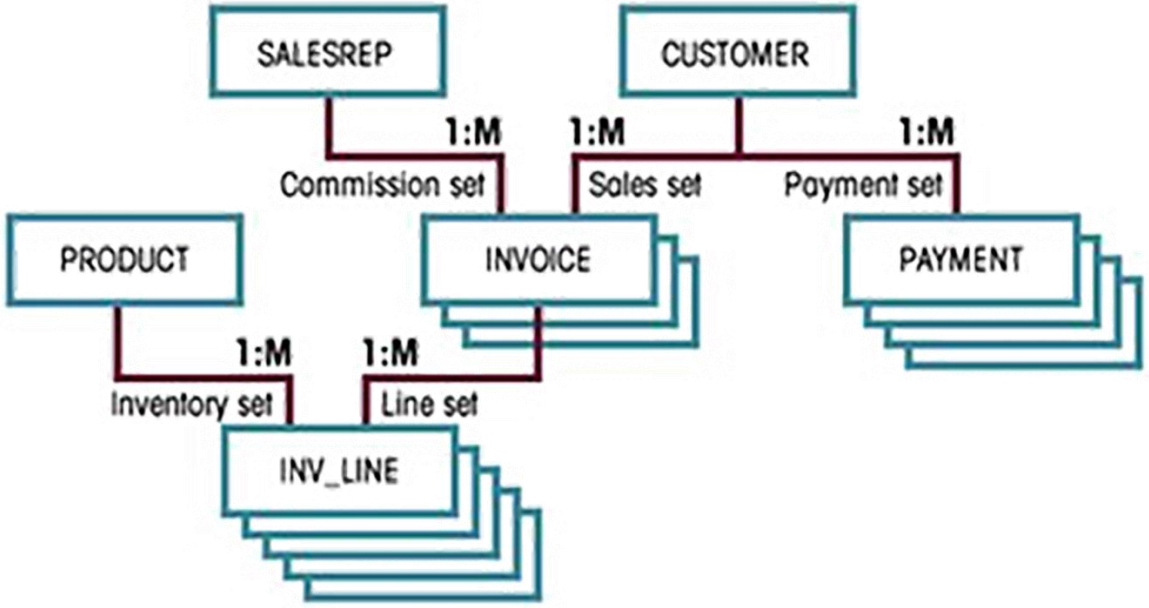
- Why starting from 60s?
  - Early day computers / programs
  - *"If we could put a man on the Moon, could we also create a computer program to track the millions of rocket parts it takes?"*
    - IBM Information Management System (*IMS)*
- Tree-like structure
  - A hierarchical database model is a data model in which the data are organized into a tree-like structure.
  - The data is stored as records which are connected to one another through links
  - A record is a collection of fields, with each field containing only one value
  - Each child record has only one parent, whereas each parent record can have one or more child records

+ Hierarchical Model



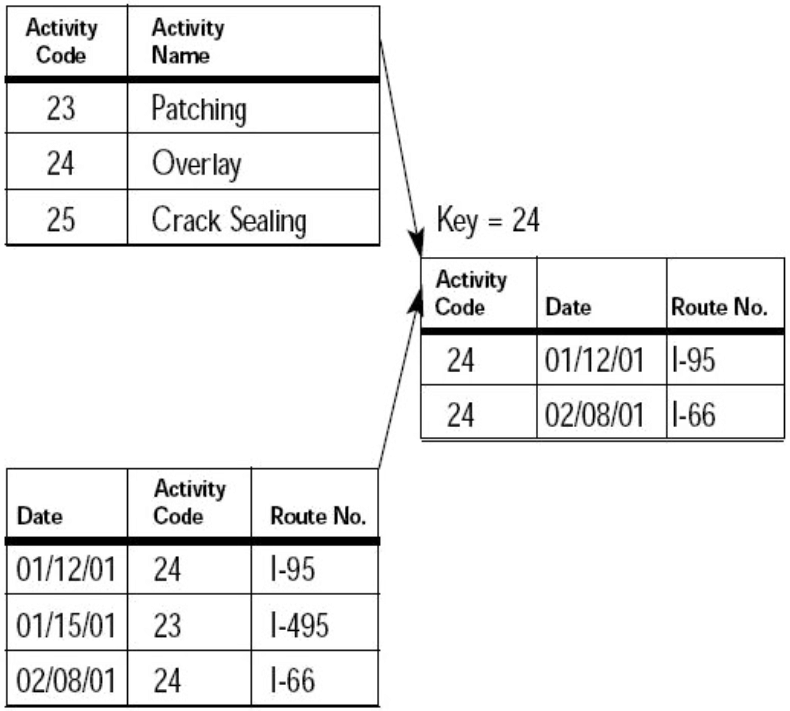
60s: Network Model

- Graph-like structure
- Allows a more natural modeling of relationships between entities



Relational Model

- Codd "*A Relational Model of Data for Large Shared Data Banks*", 1970
  - Table is the foundation
    - 1NF, 2NF, 3NF,…
  - Relational algebra / Calculus / Theory
  - Join
  - Theory and research
    - Ingres/Postgres



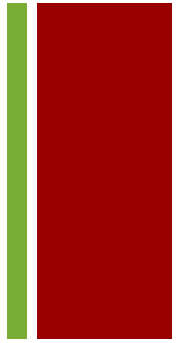
Activity Code	Activity Name
23	Patching
24	Overlay
25	Crack Sealing

Key = 24

Activity Code	Date	Route No.
24	01/12/01	I-95
24	02/08/01	I-66

Date	Activity Code	Route No.
01/12/01	24	I-95
01/15/01	23	I-495
02/08/01	24	I-66

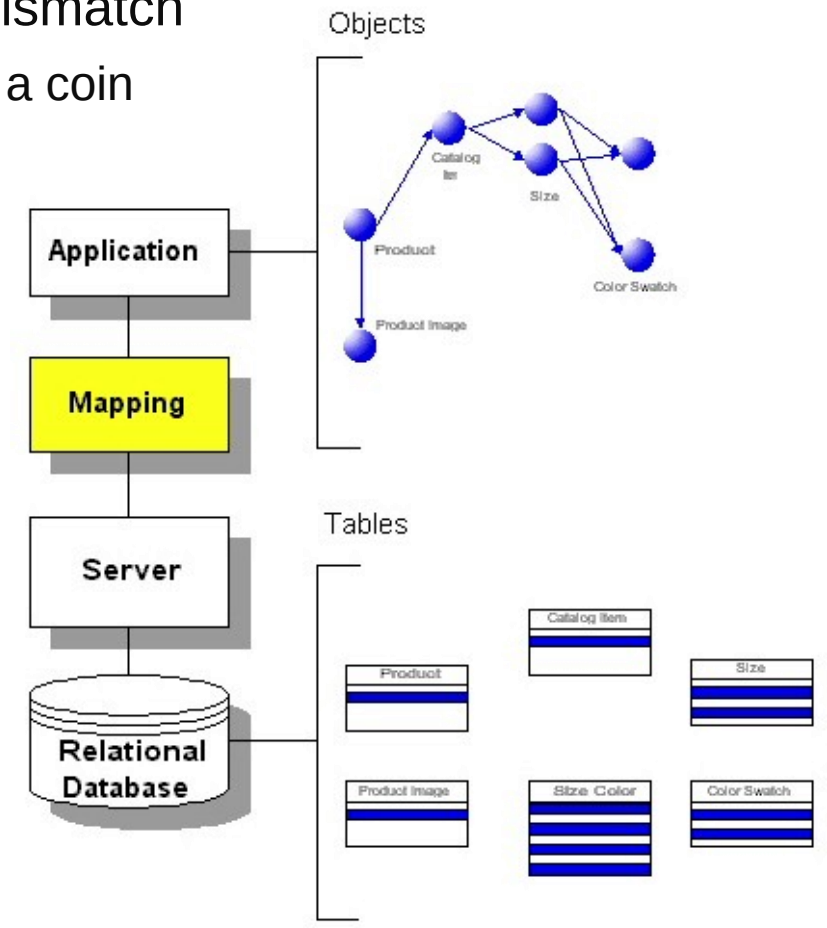
80s: Relational Model



- Concurrency
  - With ACID
- Integration
- Standard data model
  - Schema
    - Predefined
    - Constraint
  - Separate with programs
- SQL
- Commercial systems

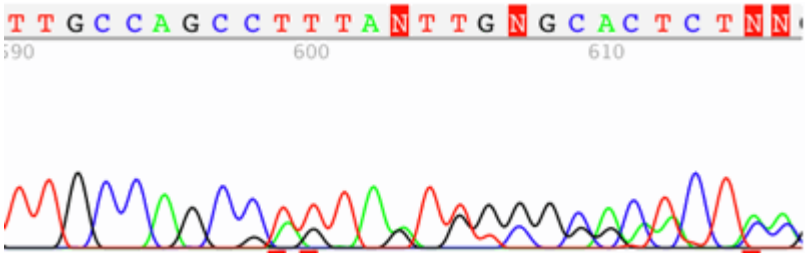
+ Relational Model

- Disadvantage
  - Impedance mismatch
    - Two sides of a coin



+ Relational Model

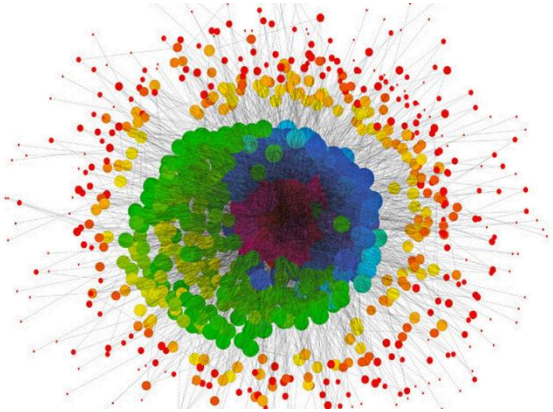
- Disadvantages
  - New data types
    - Documents
    - Sequences
    - Graphs
    - XML data
    - Spatial
    - Multimedia data

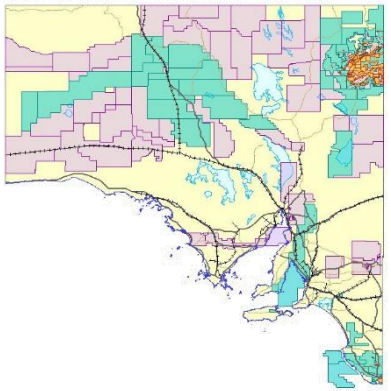



```
<Countries>

  <Country Code="AR" Name="Argentina" Regions="3" />
  <Country Code="FR" Name="France" Regions="3" />
  <Country Code="US" Name="United States" Regions="3" />

</Countries>
```



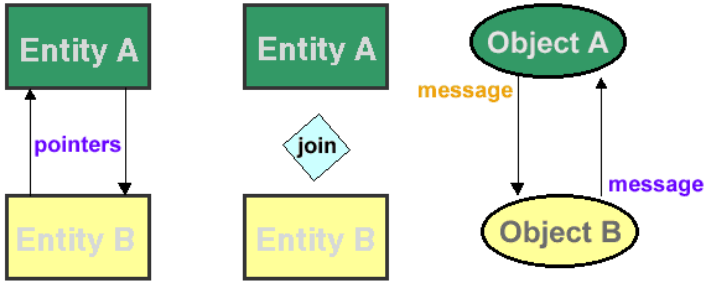


+ Relational Model

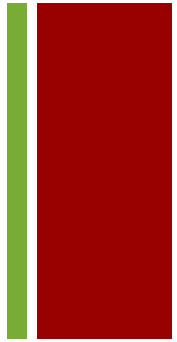
- Disadvantages
  - New applications
    - Anything that contains the new data types
    - Big data
      - Hard to Scale out on clusters
      - Low availability
      - Slow
    - What if only needs a column in a table with hundreds of columns
    - Store flexible information that violates the predefined schema
    - …
  - Can only handle 1-Dimensional Data
    - Bitmap, Hash, B-Tree to index higher dimension?

+ 90s: Object-Oriented Data Model

- Information is represented in the form of objects as used in object-oriented programming
- Add new object storage capabilities to the relational systems
  - Time-series data
  - **Geospatial data**
  - Diverse binary media such as audio, video, images, and applets
- Encapsulate methods with data structures
- Data and Program are mixed together



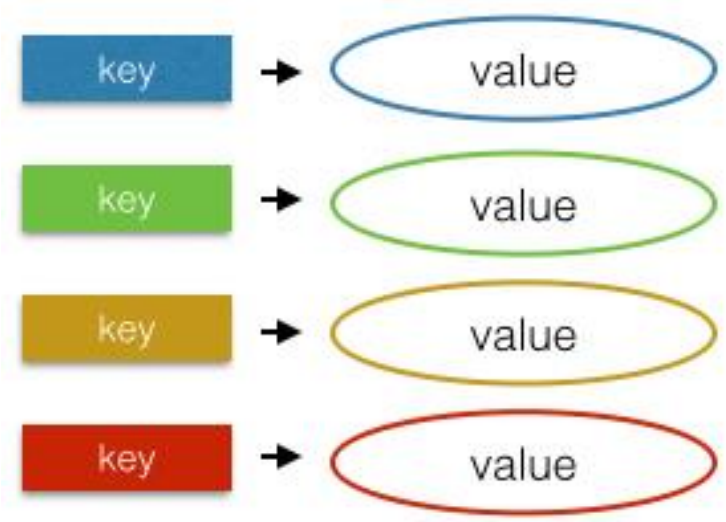
+ 2010s: NoSQL for Big Data



- Not only SQL
- Key-Value store, Document, Column store, Graph,…
  - Non-relational, Schema-less
    - Able to dynamically add new attributes
  - Built from the ground up to handle new data challenges
    - To store flexible documents → Document store database
    - To retrieve column data efficiently → Column store database
    - To store large amount of different kinds of data → K-V database
  - Open source
  - Cluster friendly
    - Horizontally scale simple operations throughput over many servers
    - Able to replicate and partition data over many servers

+ Key-Value Store Model

- Hash Key
- Value can be anything
  - BLOB
    - Binary Large Object
    - Document
    - Videos
    - …
  - Aggregate



Document Model

- Built around JSON-like documents, document databases are both natural and flexible for developers to work with.
- Unlike SQL databases, where you must determine and declare a table's schema before inserting data,

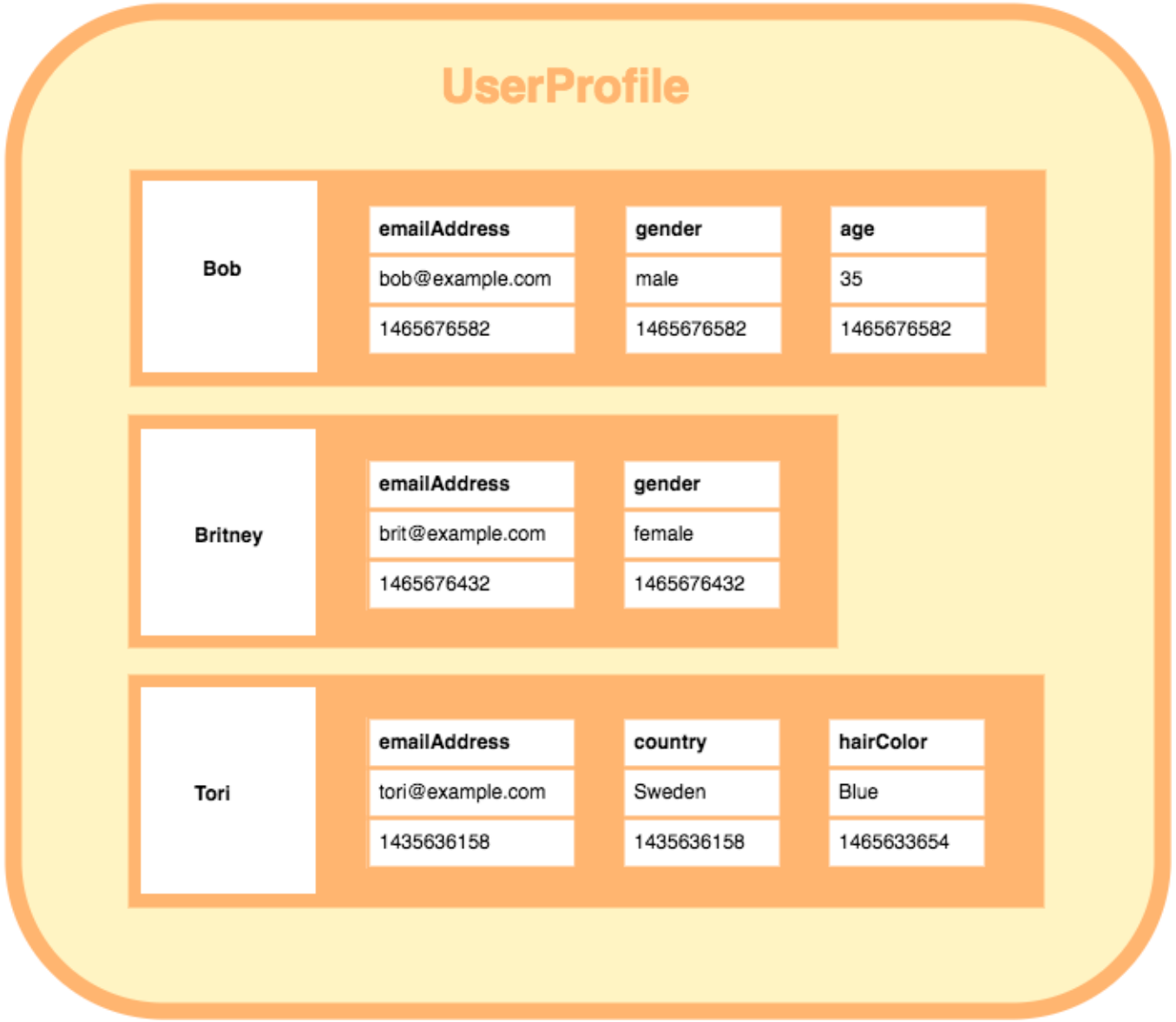
```
{
  "_id": "5cf0029caff5056591b0ce7d",
  "firstname": "Jane",
  "lastname": "Wu",
  "address": {
    "street": "1 Circle Rd",
    "city": "Los Angeles",
    "state": "CA",
    "zip": "90404"'
  }
  "hobbies": ["surfing", "coding"]
}
```

mongoDB.

+ Column Family Store Model

- A column family consists of multiple rows.
- Each row can contain a different number of columns to the other rows.
- And the columns don't have to match the columns in the other rows (i.e. they can have different column names, data types, etc).
- Each column is contained to its row. It doesn't span all rows like in a relational database.
- Each column contains a name/value pair, along with a timestamp. Note that this example uses Unix/Epoch time for the timestamp.

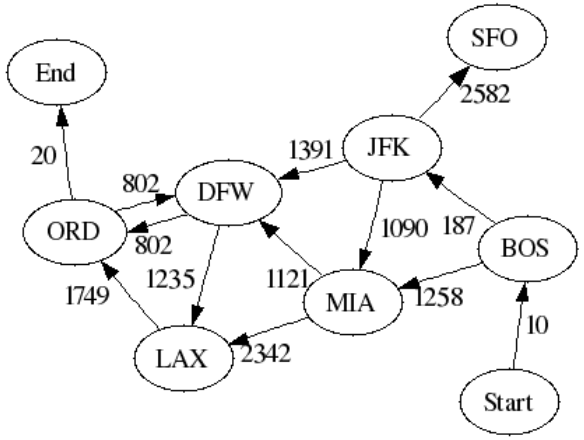
Column Family Store Model



- Cassandra
- HBase

+ Graph Model

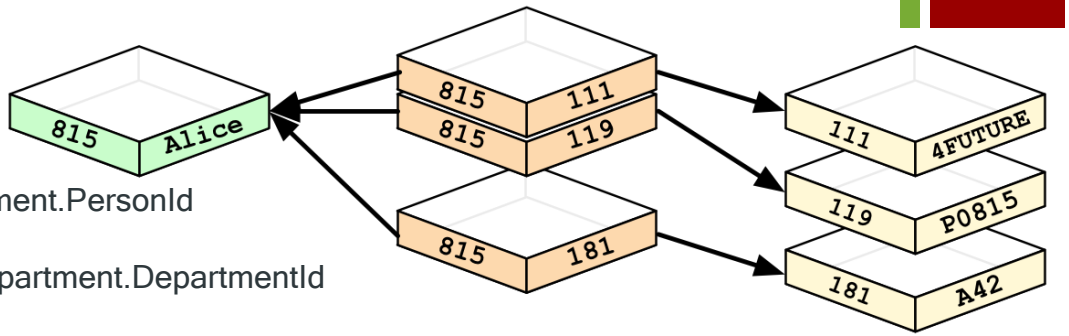
- Relational Model
  - Complex Entity + Simple Relation
- Graph Model
  - Simple Entity + Complex Relation
  - Node: Entity
  - Edge: Relation
  - Property: properties of entities and relations



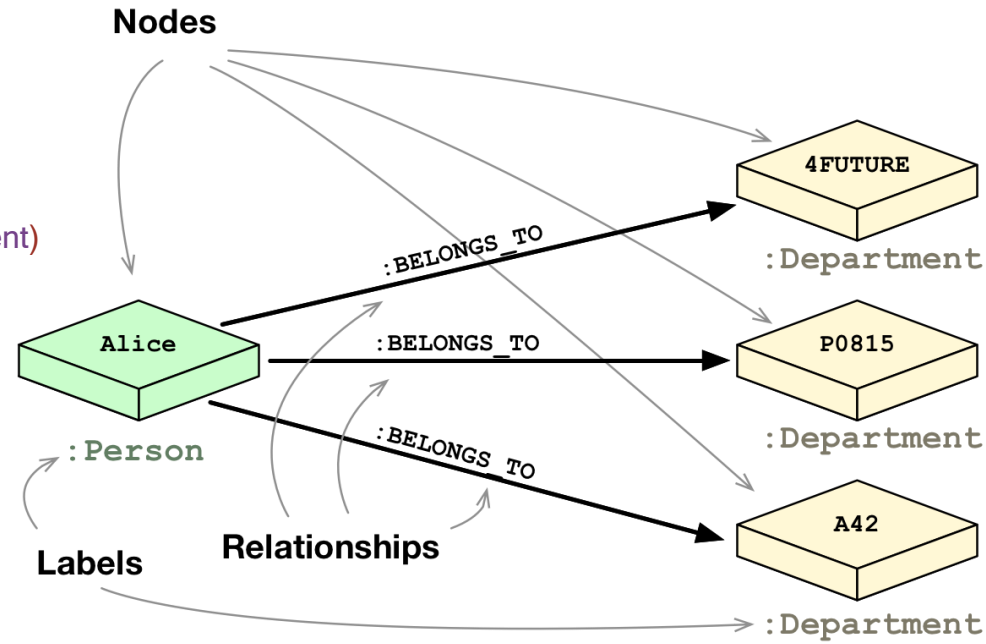
+ Graph Model

- Neo4j

```
SELECT name
FROM Person
LEFT JOIN Person_Department
                ON Person.Id = Person_Department.PersonId
LEFT JOIN Department
                ON Department.Id = Person_Department.DepartmentId
WHERE Department.name = "IT Department"
```



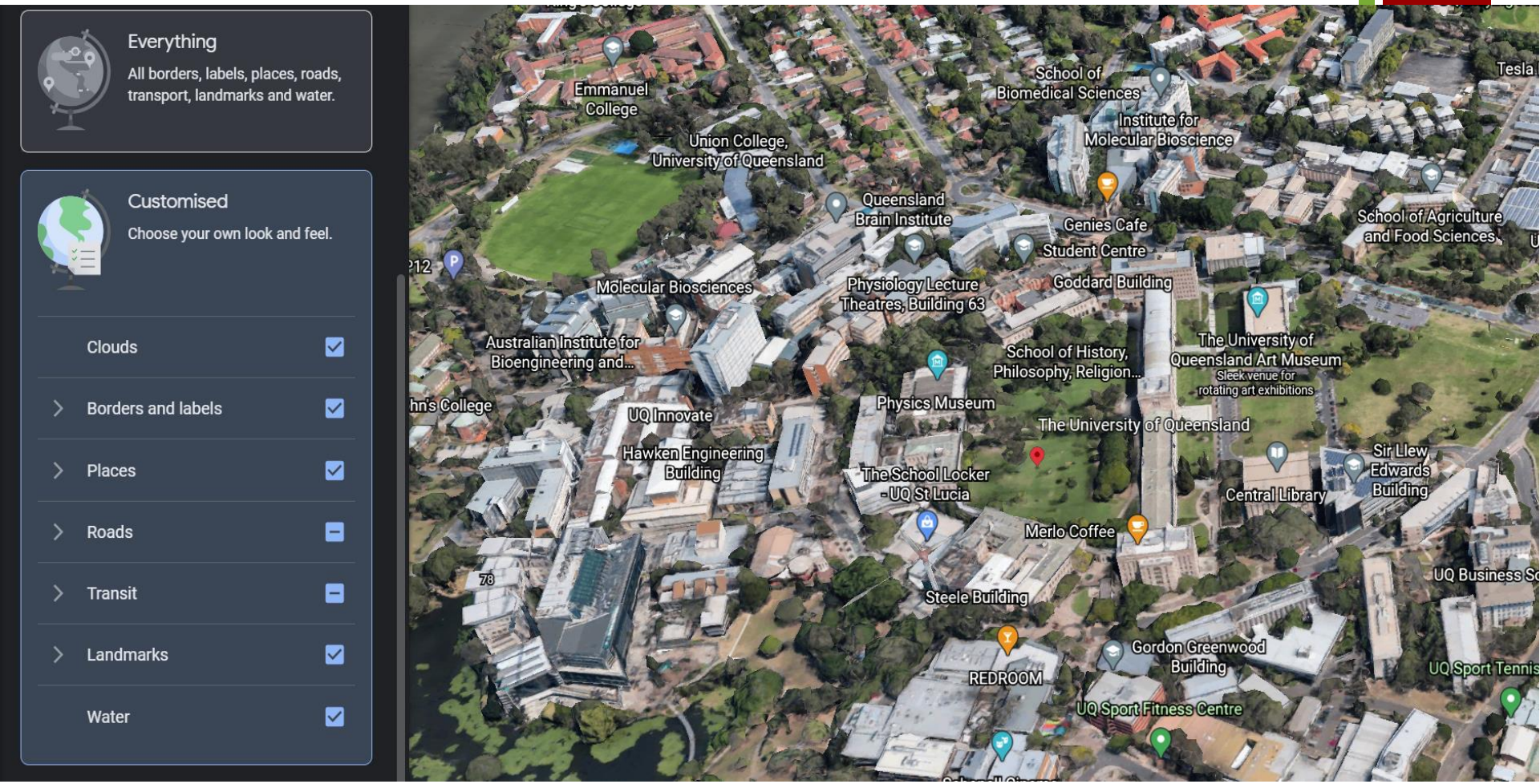

```
MATCH (p:Person)<-[:EMPLOYEE]-(d:Department)
 WHERE d.name = "IT Department"
RETURN p.name
```



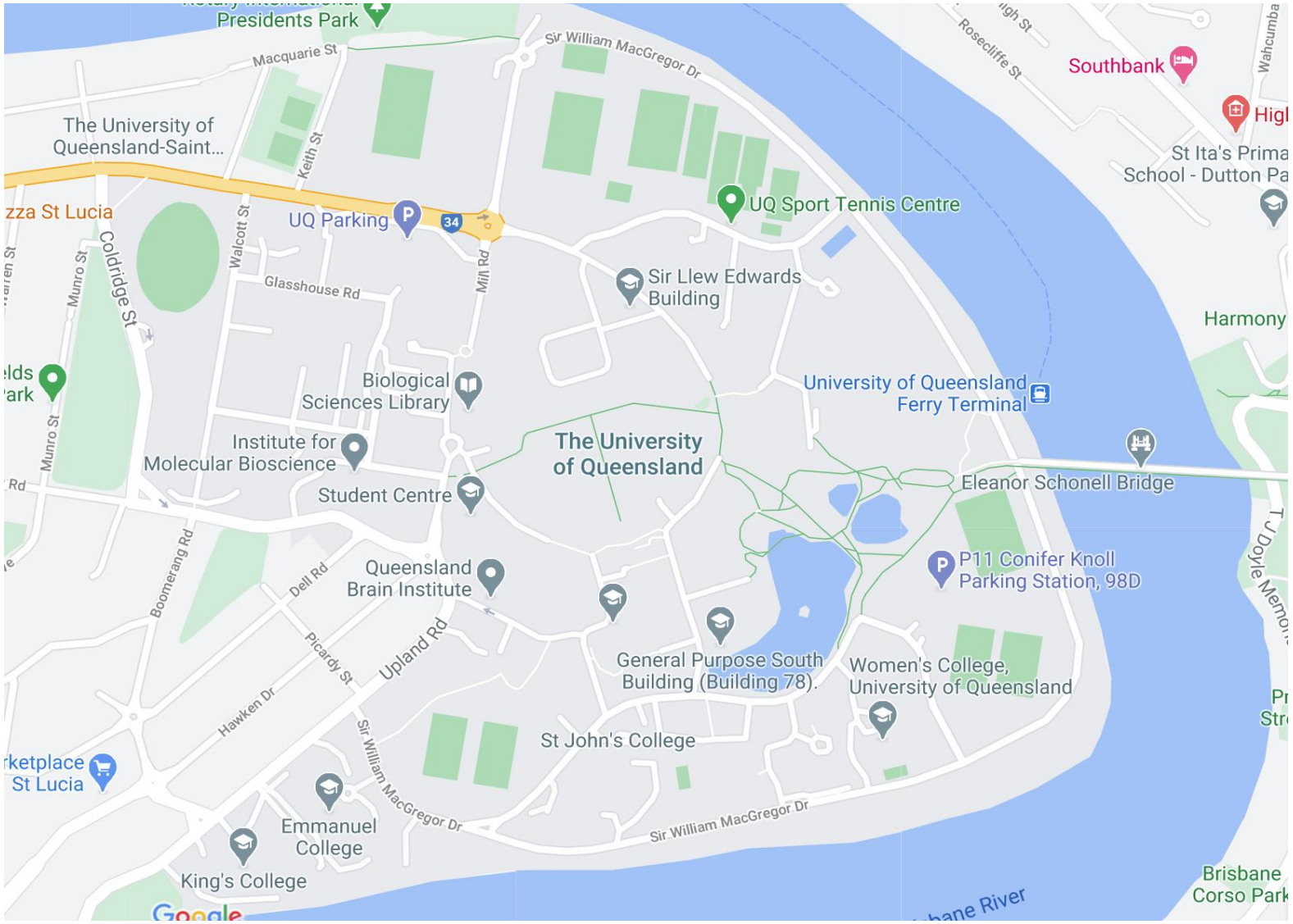
+ Beyond RDBMS

- History of DBMS
  - The 60s: Tree, Network
  - The 70s, 80s: Relation
  - The 90s: Object-Oriented
  - New millennium: Big Data, NoSQL
- Beyond RDBMS
  - Spatial and temporal attributes are ubiquitous
    - But they are poorly support by RDBMS…
  - Images and videos are becoming more popular
    - Mainstream media to represent and communicate information
  - One fundamental change:
    - #dimensions increases from 1, to 2, to several, to many…

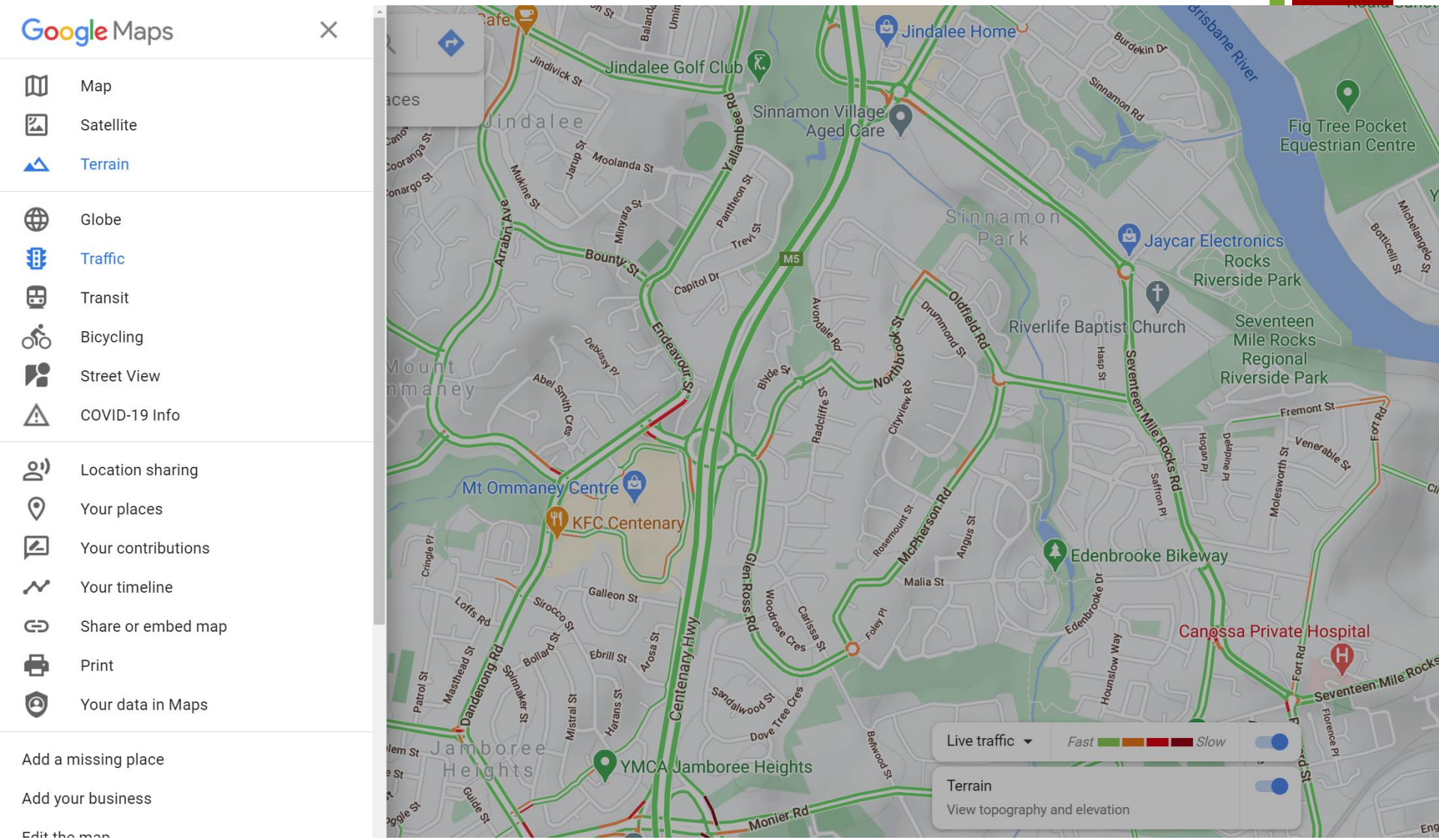
+ Google Earth



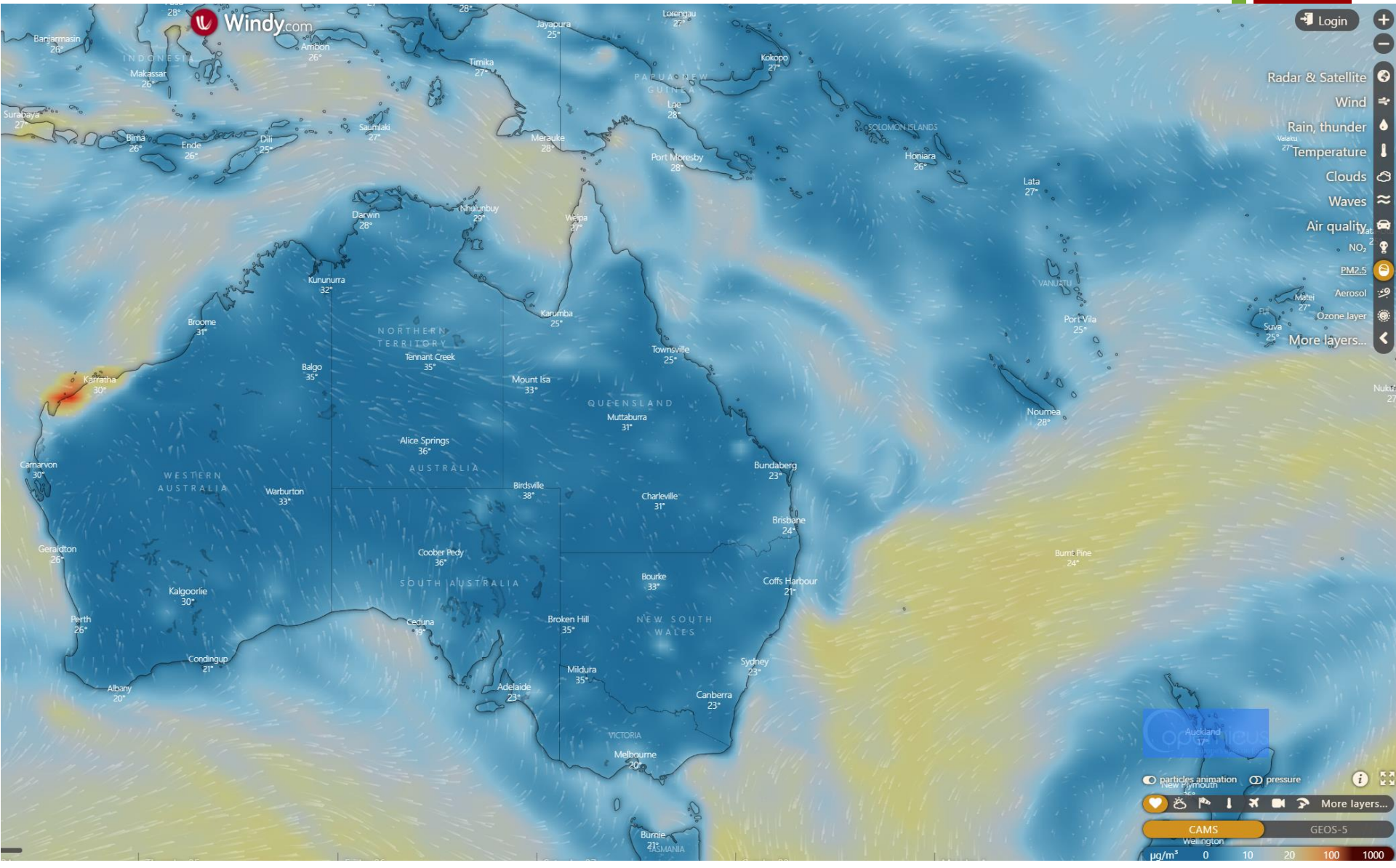
+ Google Map


Presidents Park
Macquarie St
Sir William MacGregor Dr
The University of Queensland-Saint...
Keith St
UQ Parking
34
UQ Sport Tennis Centre
Sir Llew Edwards Building
Glasshouse Rd
Mill Rd
Walcott St
Coldridge St
Munro St
Biological Sciences Library
Institute for Molecular Bioscience
The University of Queensland
Student Centre
University of Queensland Ferry Terminal
Eleanor Schonell Bridge
Harmony
Southbank
St Ita's Prima School - Dutton Pa
Roseclife St
Boomerang Rd
Dell Rd
Queensland Brain Institute
Upland Rd
Picardy St
Hawken Dr
General Purpose South Building (Building 78).
Women's College, University of Queensland
P11 Conifer Knoll Parking Station, 98D
St John's College
Emmanuel College
King's College
Brisbane Corso Park
Google

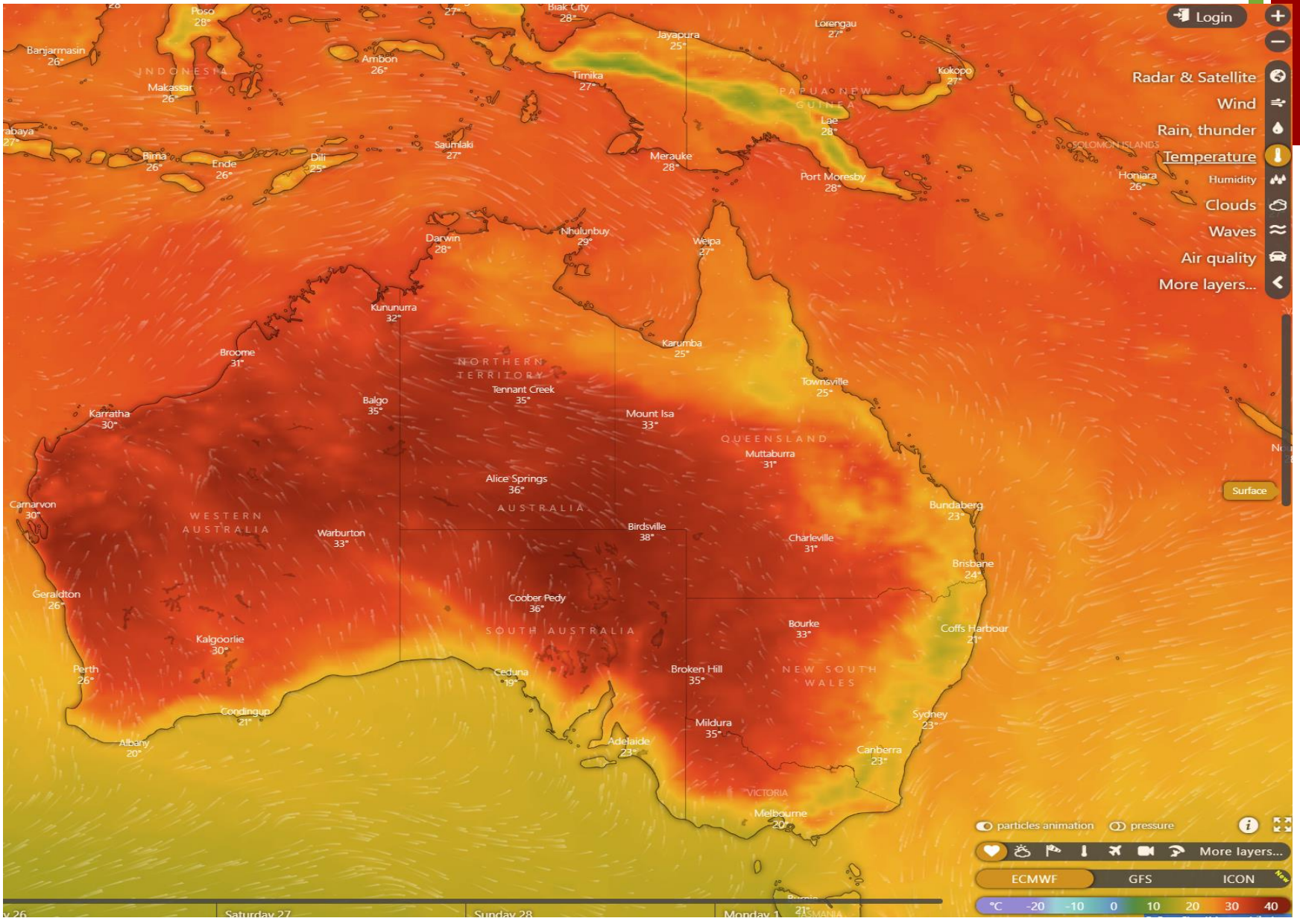
+ Google Map


Google Maps
Map
Satellite
Terrain
Globe
Traffic
Transit
Bicycling
Street View
COVID-19 Info
Location sharing
Your places
Your contributions
Your timeline
Share or embed map
Print
Your data in Maps
Add a missing place
Add your business
Jindalee Golf Club
Jindalee Home
Sinnamon Village Aged Care
Fig Tree Pocket Equestrian Centre
Brisbane River
Jaycar Electronics
Rocks Riverside Park
Riverlife Baptist Church
Seventeen Mile Rocks Regional Riverside Park
Mt Ommaney Centre
KFC Centenary
Edenbrooke Bikeway
Canossa Private Hospital
YMCA Jamboree Heights
Live traffic
Fast
Slow
Terrain
View topography and elevation

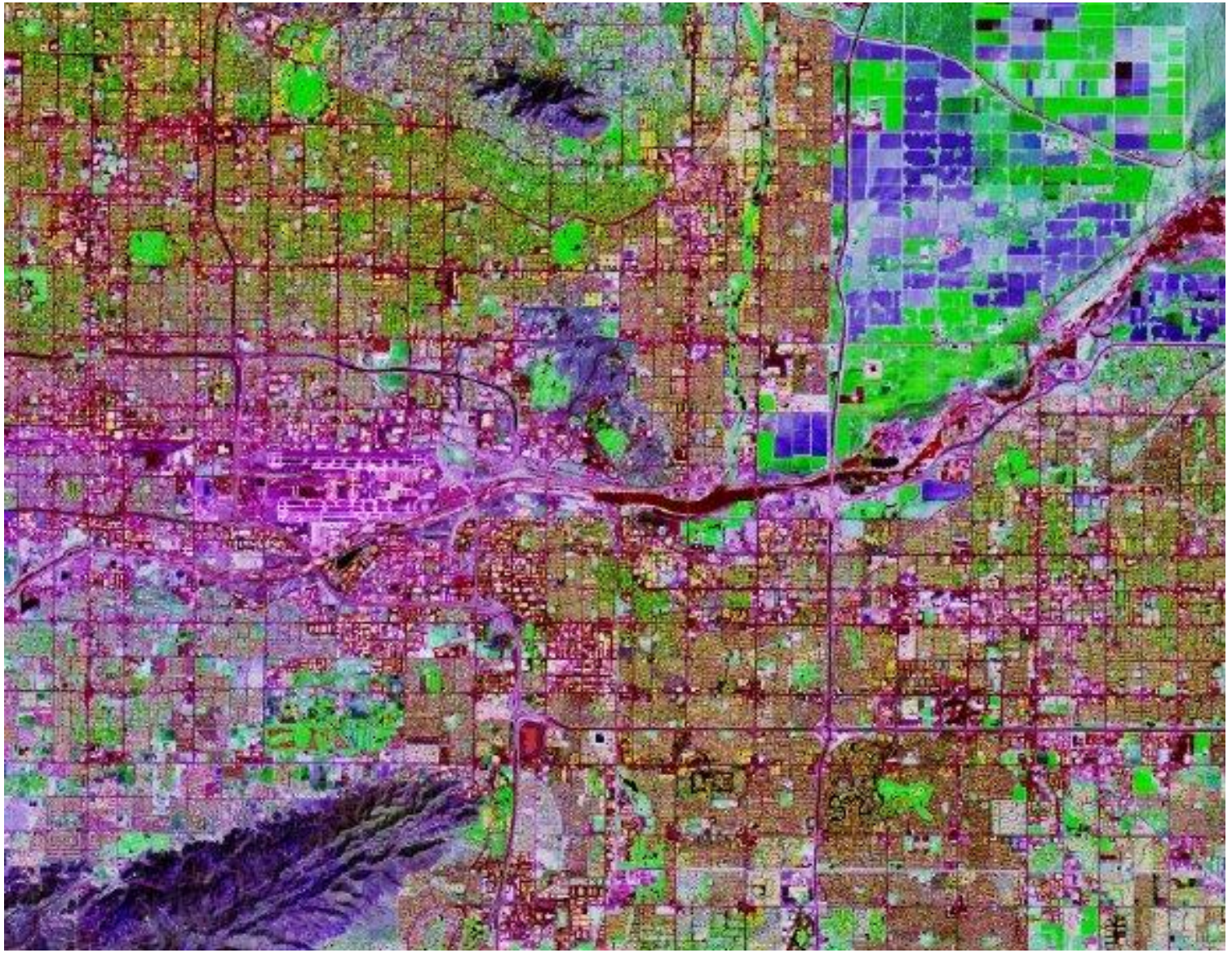
+ Windy.com


Windy.com
Login
Radar & Satellite
Wind
Rain, thunder
Temperature
Clouds
Waves
Air quality
NO2
PM2.5
Aerosol
Ozone layer
More layers...
Banjarmasin 26°
Ambon 26°
Makassar 26°
Timika 27°
Jayapura 25°
Lorengau 27°
Kokopo 27°
Lae 28°
Merauke 28°
Port Moresby 28°
Honiara 26°
Lata 27°
Saumlaki 27°
Dili 25°
Ende 26°
Bima 26°
Darwin 28°
Nhulunbuy 29°
Weipa 27°
Kununurra 32°
Broome 31°
Karratha 30°
Tennant Creek 35°
Balgo 35°
Karumba 25°
Mount Isa 33°
Townsville 25°
Muttaburra 31°
Alice Springs 36°
Noumea 28°
Port Vila 25°
Suva 25°
Carnarvon 30°
Warburton 33°
Birdsville 38°
Charleville 31°
Bundaberg 23°
Brisbane 24°
Geraldton 26°
Coober Pedy 36°
Bourke 33°
Coffs Harbour 21°
Burnt Pine 24°
Kalgoorlie 30°
Perth 26°
Ceduna 19°
Broken Hill 35°
Condingup 21°
Albany 20°
Mildura 35°
Adelaide 23°
Sydney 23°
Canberra 23°
Melbourne 20°
Burnie 21°
Auckland 17°
particles animation
pressure
More layers...
CAMS
GEOS-5
µg/m³ 0 10 20 100 1000

+ Windy.com


Login
Radar & Satellite
Wind
Rain, thunder
Temperature
Humidity
Clouds
Waves
Air quality
More layers...
Surface
particles animation
pressure
More layers...
ECMWF
GFS
ICON
°C -20 -10 0 10 20 30 40
Darwin 28°
Nhulunbuy 29°
Weipa 27°
Kununurra 32°
Broome 31°
Karumba 25°
Tennant Creek 35°
Balgo 35°
Mount Isa 33°
Townsville 25°
Karratha 30°
Muttaburra 31°
Alice Springs 36°
Carnarvon 30°
Bundaberg 23°
Birdsville 38°
Warburton 33°
Charleville 31°
Brisbane 24°
Geraldton 26°
Coober Pedy 36°
Bourke 33°
Coffs Harbour 21°
Kalgoorlie 30°
Broken Hill 35°
Perth 26°
Ceduna 19°
Sydney 23°
Condingup 21°
Mildura 35°
Albany 20°
Adelaide 23°
Canberra 23°
Melbourne 20°
Jayapura 25°
Lorengau 27°
Timika 27°
Kokopo 27°
Lae 28°
Merauke 28°
Port Moresby 28°
Honiara 26°
Saumlaki 27°
Ambon 26°
Makassar 26°
Banjarmasin 26°
Poso 28°
Biak City 28°
Bima 26°
Ende 26°
Dili 25°
INDONESIA
PAPUA NEW GUINEA
SOLOMON ISLANDS
NORTHERN TERRITORY
QUEENSLAND
WESTERN AUSTRALIA
SOUTH AUSTRALIA
NEW SOUTH WALES
VICTORIA
Saturday 27
Sunday 28
Monday 1

+ Remote Sensing Data



+ Modern Real-World Applications

- Very large datasets
  - Many with streaming nature too
- Complex data structures
  - Often represented as high dimensional data
  - Many with spatial and temporal attributes (timestamp)
  - One application often needs to integrate heterogeneous data from many sources
- Data analytics queries
  - Data Warehouses, Data Mining, Machine Learning…
- Again: Database is not only about storage
  - How to retrieve and use the data

+ INFS 4205/7205:

- **What** will we learn?
  - The latest in large-scale data management and processing: **complex data types** and **similarity-based queries**
  - How to **manage** complex data to enable efficient **processing** of advanced queries
  - Applicable to a wide range of **applications**, including spatial, temporal, network and multimedia data…
- What is not covered?
  - How to learn the **representation** of the complex data
  - Everything to vectors, such as word2vec, node2vec, region2vec…
- **How** will we learn?
  - Attending lectures
  - Attending tutorials
  - Studying selected research papers
  - Discuss on Piazza
  - Doing individual assignments

What to be Covered

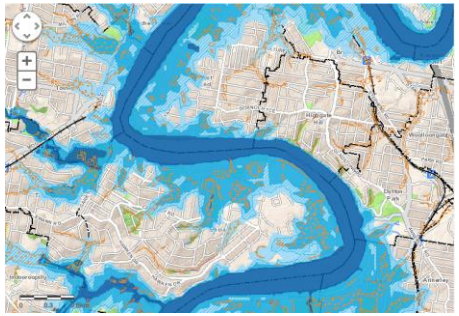
1. Spatial Databases Introduction
   - Spatial Data Types, Data Structures and Spatial Databases
2. Spatial Indexing Mechanisms
3. Spatial Algorithms and Query Processing
4. Spatiotemporal Data Management
5. High-Dimensional Indexing, Search and Applications
6. Routing Planning

+ Multidimensional Data

- An object is ***k*-dimensional** means that the object is described by k attributes as a whole
  - Not by each attribute individually
    - (Name, Age, Weight, Height, …) vs.
    - (Longitude, Latitude, Elevation)
- Examples:
  - Point(x, y) - 2D: geographic data (GIS, LIS…)
  - Point(x, y, z) - 3D: the universe, brain, molecule structure…
  - Point(x, y, t) – 3D: spatiotemporal
  - Person(Age, Weight, Height) – 3D?
  - Image Color Histogram(c1, c2, … c128) – 128D
  - Image(texture, shape, colour) – high dimension!
  - CNN features
  - Video – a sequence of image frames

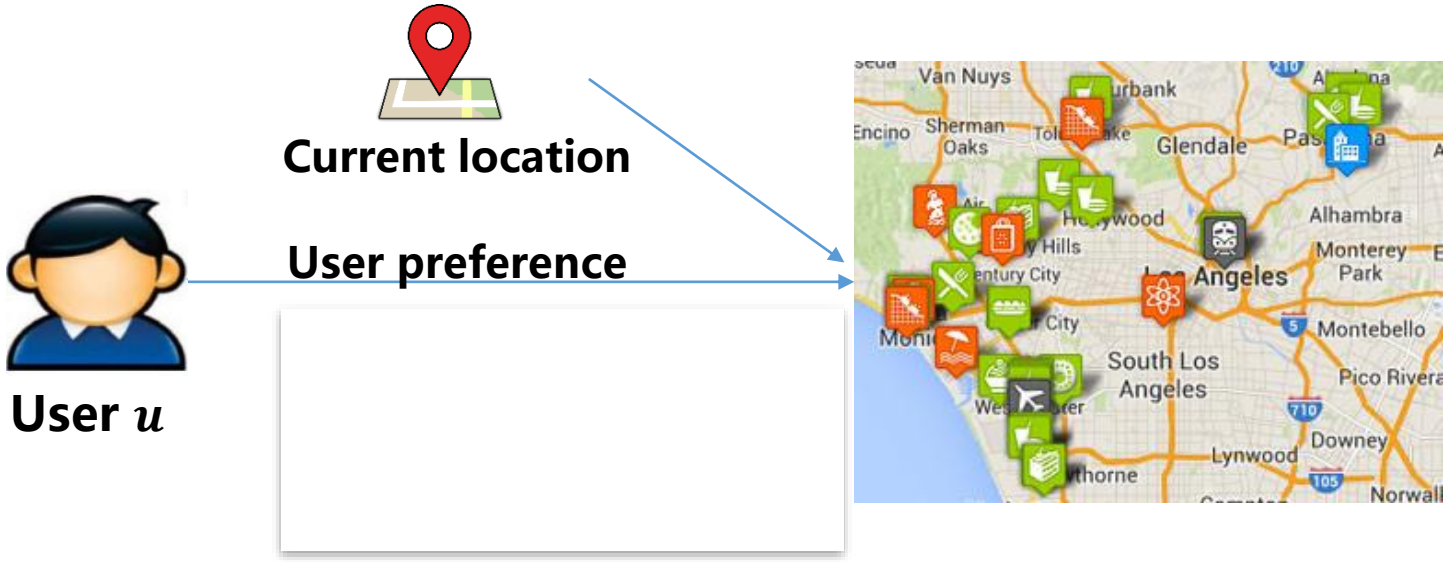
Why Starting with Spatial Databases?

- One of the most widely used multidimensional databases
  - The most basic multidimensional data
  - Foundation of understanding other high-D data
- Many important application domains have **spatial data** and **queries**. Some Examples:
  - **Insurance Risk Manager**: Which homes are most likely to be affected in the next great flood on the Brisbane river?
  - **Location-based Recommendation**: A location-based recommendation is an information filtering service, which selectively returns items (e.g., venues, events, travel routes) to a user with the consideration of relevant spatial information (e.g., current/historical locations) and the personal preferences

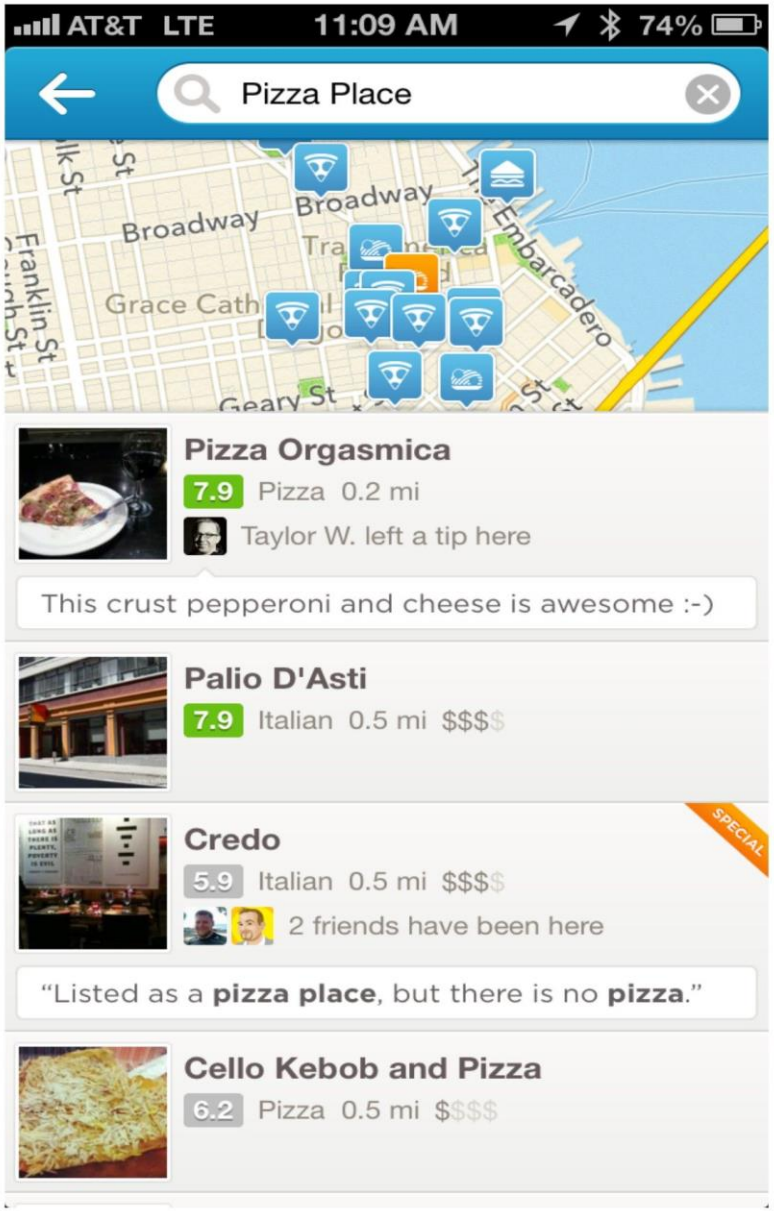


Location-based Recommendation

- **Given a user *u* with his/her current location *l*, recommend top-*k* spatial items that *u* would be interested in.**

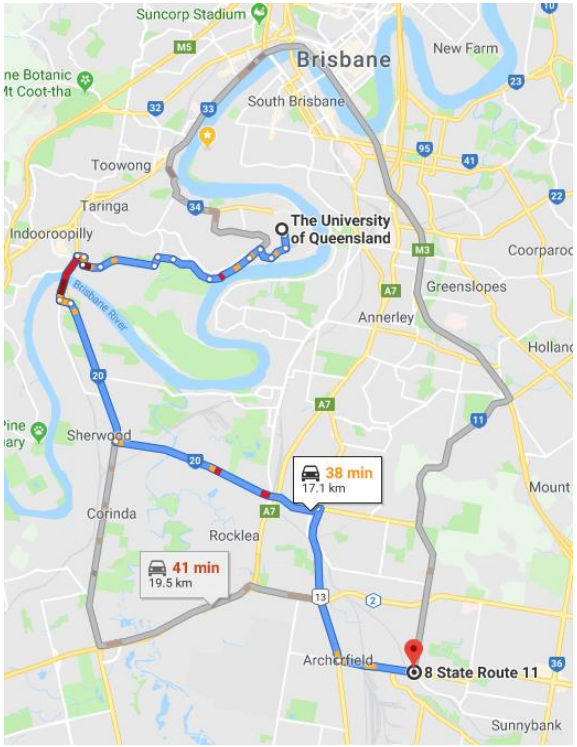


+ Location-based Search



+ Route Planning in Road Network

1. Shortest path from UQ to Coopers Plain
2. Fastest path from UQ to Coopers Plain when I depart at 5pm
3. Fastest path from UQ to Coopers Plain when I depart between 5pm and 7pm
4. Paths from Indooroopilly, City and Woolloongaba to UQ such that everybody can arrive by 10am
5. Fastest path from West End to Airport with toll fee under $6 (Go Between? M7?)
6. Earliest to arrive Coopers Plain by bus when I depart at 5pm with exchange times under 3
7. I want to spend at least 3 hours at UQ, 1 hour at Garden City, 1 hour at Southbank Parkland and arrive home by 9pm, what's the schedule to waste the least time for transportation?



How About Relational DBMS?

- Queries
  - Where is Building 78?
  - Which courses are delivered in GP Building?

Non-spatial

- Which buildings are adjacent to *the UQ* lake?
- Which buildings are adjacent to *a* lake?

spatial

What's Special about Spatial Queries?

- Retrieval & update of spatial data is based on the **spatial location** of a data object
  - *(vs. alphanumeric attributes in RDBMS)*
- Fast execution of **geometric operations like** the intersection, union, and difference of spatial objects
  - *(vs. simple comparison (=, >, <) in RDBMS)*
- Fast execution of **complex spatial queries**
  - Spatial join, knn queries, skyline queries
- RDBMS Obvious limitations
  - Limited data types – no support for multidimensional data!
  - Limited query types

Spatial DBMS (SDBMS)

- A SDBMS is a software module that
  - Can work with an underlying DBMS
  - Supports spatial data models, spatial abstract data types (ADTs)
  - Supports a spatial query language
  - Supports spatial indexing, efficient algorithms for processing spatial operations, and domain specific query optimization
- Examples:
  - Oracle Spatial and Graph in 12c (formally Oracle Spatial Extension of Oracle 8 in 1997), now separately licensed
  - Microsoft SQL Server supports spatial types since 2008
  - PostgreSQL DBMS uses the spatial extension PostGIS
  - Various of Spatial extensions on NoSQL systems

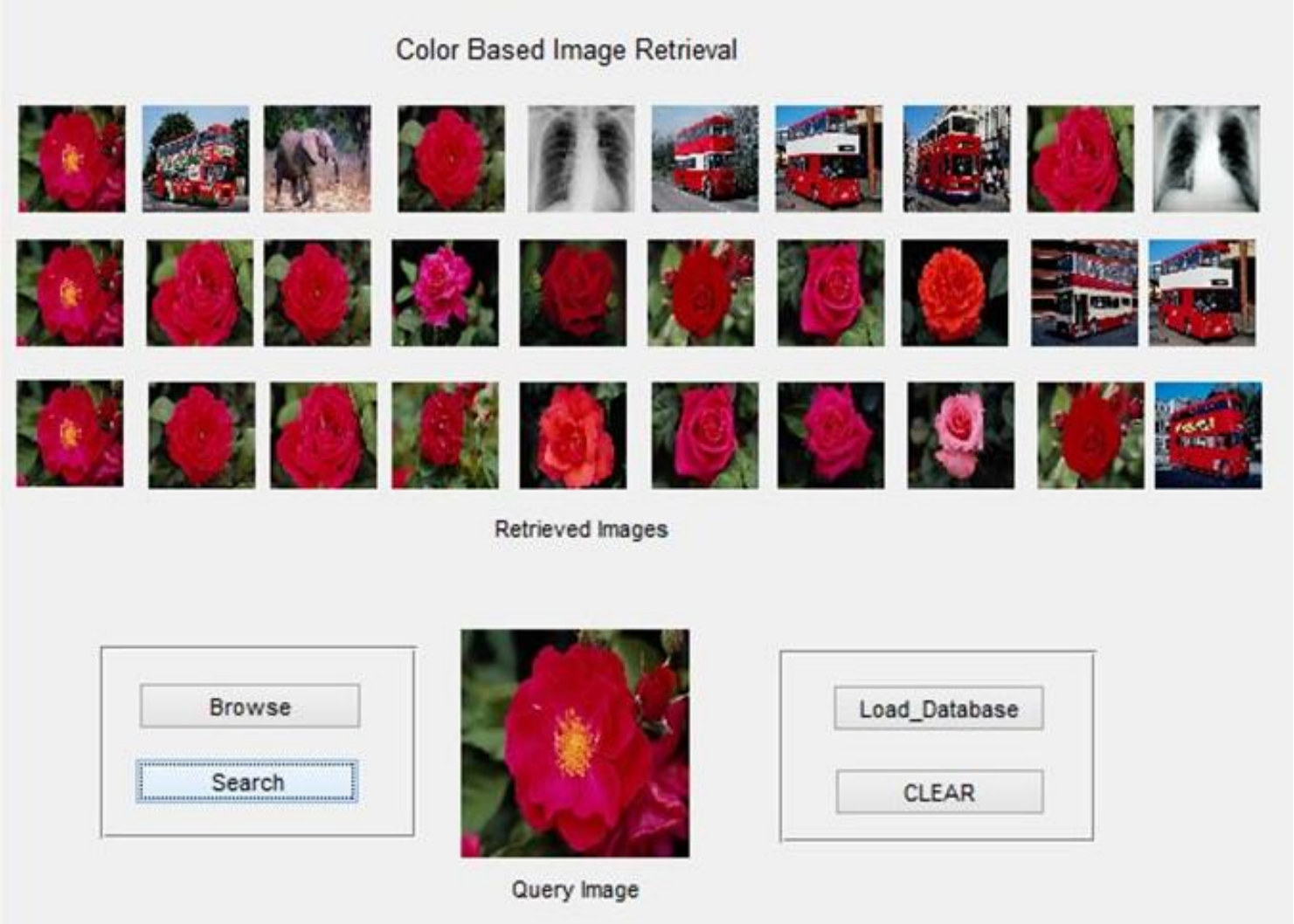
+ MM Databases

- Multimedia is a much more powerful communication tool than traditional data in our daily life
  - Image showcase, short videos, graphic design, TV commercial, speech, movie, mobile phone multimedia message, etc
- We need to organize, manage and search these new multimedia data
  - RDBMS are no longer suitable for complex multimedia data
  - Need for robust systems which can manage and search multimedia data in a reliable and efficient way

From Spatial to Multimedia

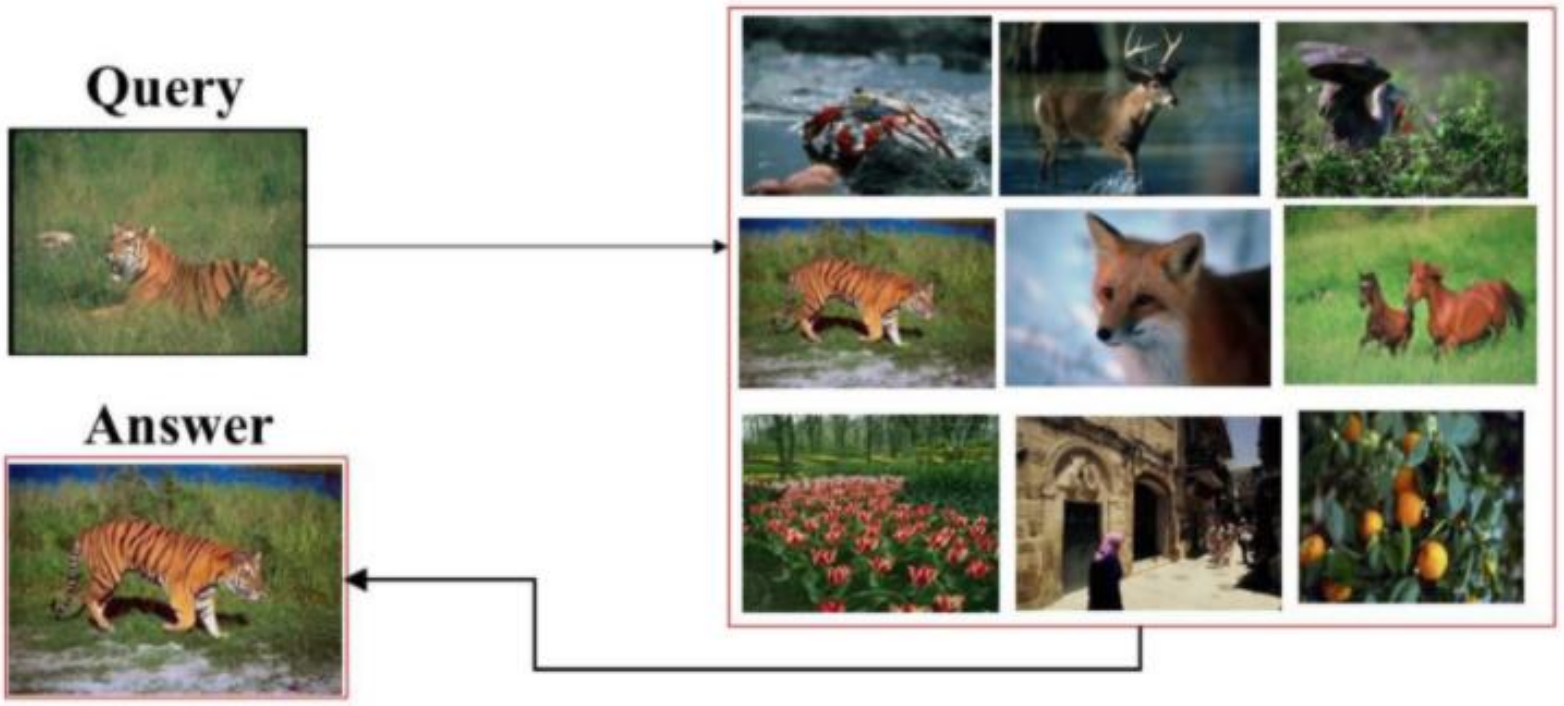
- Very different on the surface, but many similarities fundamentally
  - Data represented as **multidimensional vectors**
- Keywords-based vs Content-based search
  - **Keywords-based:** using text annotations
  - **Content-based:** using automatically extracted features such as colors, textures and shapes, CNN features
  - Both have advantages and disadvantages
- Applications of content-based multimedia search
  - knn query in a high-dimensional vector space

+ Colour-based image retrieval



+ Content-based image retrieval

Given a query image, try to find visually similar images from an image database



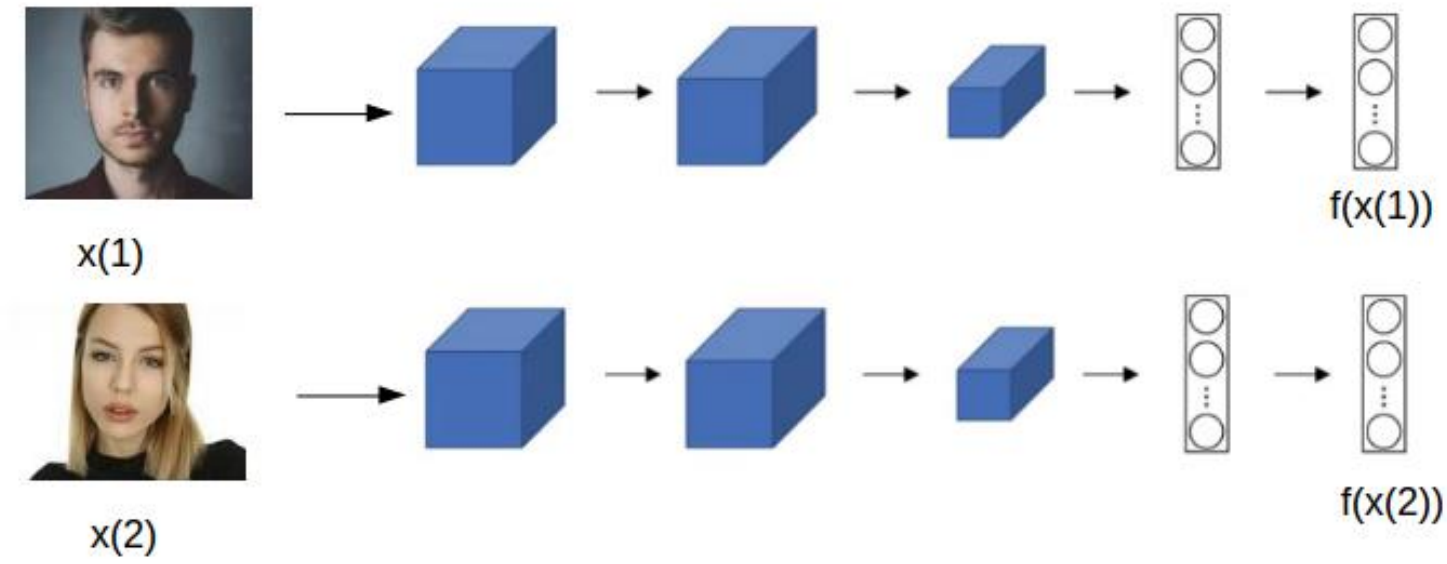
+ Person Reidentification

Do you recognize
this person on the right?

+ Person Reidentification

Content-based Search

k-nearest neighbour search in a high-dimensional space.



Here, f(x(1)) and f(x(2)) are the encodings of images x(1) and x(2) respectively. So,

$$d(x(1),x(2)) = || f(x(1)) - f(x(2)) ||^2_2$$

+ Managing high-dimensional Data

- Time Cost factors
  - I/O cost: Disk page accesses
  - CPU cost: Computation of similarity/distance
- As dimensionality increases, the portion of CPU cost in total response time increases too
  - This is different from traditional databases, where only I/O costs are considered
  - Especially similarity defined/computed by deep neural networks
- Indexing high-dimensional feature vectors
- Dimensionality Curse
  - The performance of an index degrades rapidly as dimensionality increases, and eventually underperforms linear scan!

+ Why Study This Course?

- For an IT graduate from UQ
  - What you have learnt about relational DB may be not applicable to some applications, but…
  - What you have learnt about machine learning may be not applicable to the real big data applications
    - Memory-based machine learning (both training data and learned models)
- For prospect researchers
  - Spatial and multimedia databases are still a major area with many open problems, and there are many other types of high dimensional data
- For job seekers
  - Complex data are typically dimensional
  - High-dimensional data management and processing evolve rapidly from a specialised area to a “commodity” skill with great demand