DATA2001 – Data Science, Big Data, and Data Diversity

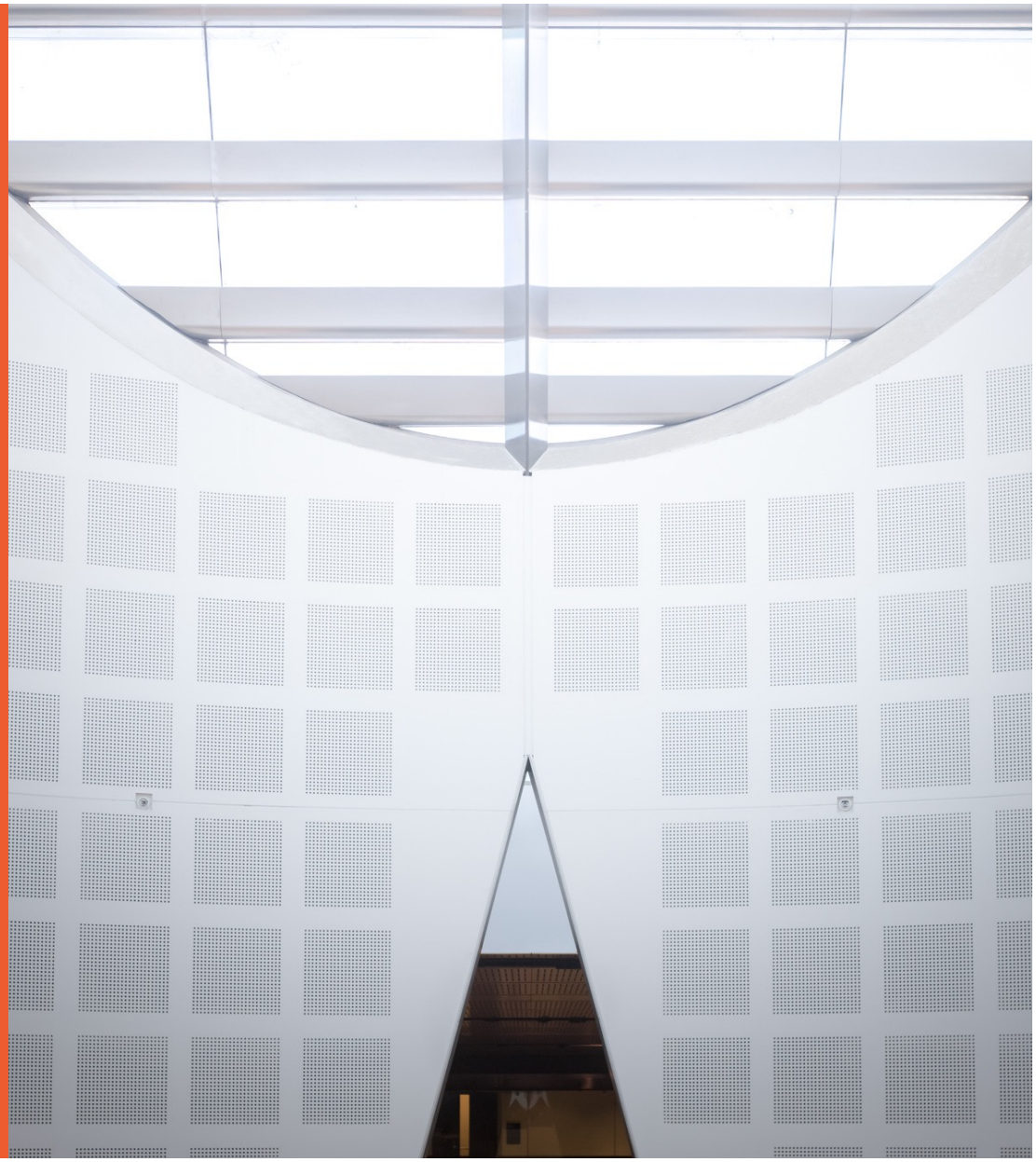
Assignment Announcement

Presented by

A/Prof Uwe Roehm

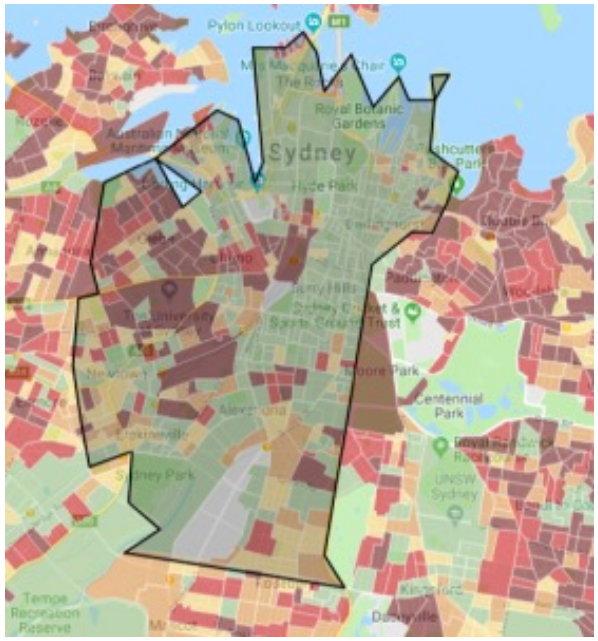
School of Computer Science

THE UNIVERSITY OF SYDNEY



Practical Assignment: Sydney Liveability Analysis

- Assignment specification available in Canvas (Canvas: Modules –> Assignment)
  - **Worth 20% of the final grade in DATA2001/DATA2901**
  - **Due on Friday of Week 12**
    - **Python/SQL notebook; brief report; team demo in tutorials of Week 12/13**
- Main idea:
  - Calculate a 'liveability score' per SA2 suburb in Greater Sydney
    - Based on ABS data about population and school and crime data in NSW
    - Also extend score with own data specifically for City of Sydney
  - Visualise and correlate with income data



[SA3 vs. SA1 https://communityinsightaustralia.org/what-are-sas/]

Practical Assignment: Sydney Liveability Analysis

- Goal: Practical experience with data variety, data analysis, and presentation
  - Technologies as covered in this course: Python, Jupyter notebooks, Web APIs, and SQL
- Three tasks:
  - Data import, integration and database generation
    - We provide census data and spatial data from NSW government and BOCSAR
    - Needs to be loaded into database and combined, eg. via spatial join
    - Extend with own datasets from "City of Sydney Open Data Hub"
    - **Milestone 1: Propose stakeholder and extra datasets by Week 11 tutes**
  - Liveability Analysis (Jupyter Notebook)
    - Computation of risk score per neighbuorhood; example formula is provided
    - When adding other datasets, feel free to adjust formula
    - Correlation analysis to affluency of neighborhoods
  - Documentation and (brief) report, including stakeholder pitch
  - Additional ML task for teams in advanced DATA2901

Provided Datasets (cf. Canvas)

- ABS Data
  - Census data on *neighbourhoods* (SA2-level areas) in Greater Sydney such as population, land area, number of dwellings
  - Business statistics per SA2-area
  - Income and rent statistics to check for correlation with
- school_catchment Data
  - shape data for primary, secondary and future Government schools catchments
- break_and_enter Data
  - shape data of theft 'hotspots' in NSW as determined by BOCSAR
- Note that SA2-level data from the ABS does not always match suburbs; neither the ABS neighbourhoods nor the BPFL data contain actual shapes
  - cf. tutorial this week on how to retrieve boundary data for neighbourhoods
- Adding more datasets from your side is explicitly encouraged (and gives points).
  - Try different types and forms, not just CSV…

Assignment Rules

- Groupwork
  - teams of 2 or 3 (unless odd-size class or other good reasons)
  - All team members should be in the same tutorial
- Deliverables: **Jupyter notebook** with source code and a short **report (PDF)**
  - **See page 4 of the assignment handout**
- Due on Friday of Week 12
  - Submission page and marking rubric will be published in Canvas
  - Only one member per team needs to submit for the whole group; they should submit both a ZIP archive under "Sydney Liveability Analysis Assignment" and also the PDF of your report in the separate "TurnItIn Dropbox – Sydney Liveability Analysis"
  - Late submissions: -5% of available marks per day late; 0 after more than 5 days
- **Demo in Weeks 12 and 13**
  - There will be a short demo during the tutorials of the last two weeks to the tutors
  - Individual grades can be scaled based on participation in project or demo

Tip: PostGIS

[http://postgis.net/documentation/]

- Spatial database extension for PostgreSQL supporting geographic objects (OGC)
  - **Geometry types** for Points, LineStrings, Polygons, MultiPoints, etc.
    - including import/export from standard formats such as GeoJSON or KML
  - Support for **spatial reference systems** and transformations between
  - **Spatial predicates** on geometries using the 3x3 nine-intersection model
  - Spatial operators for determining **geospatial measurements** like area, distance, length and perimeter, and **geospatial set operations**, like union, difference etc.
  - **R-Tree indexing** (over GiST)
- Example:

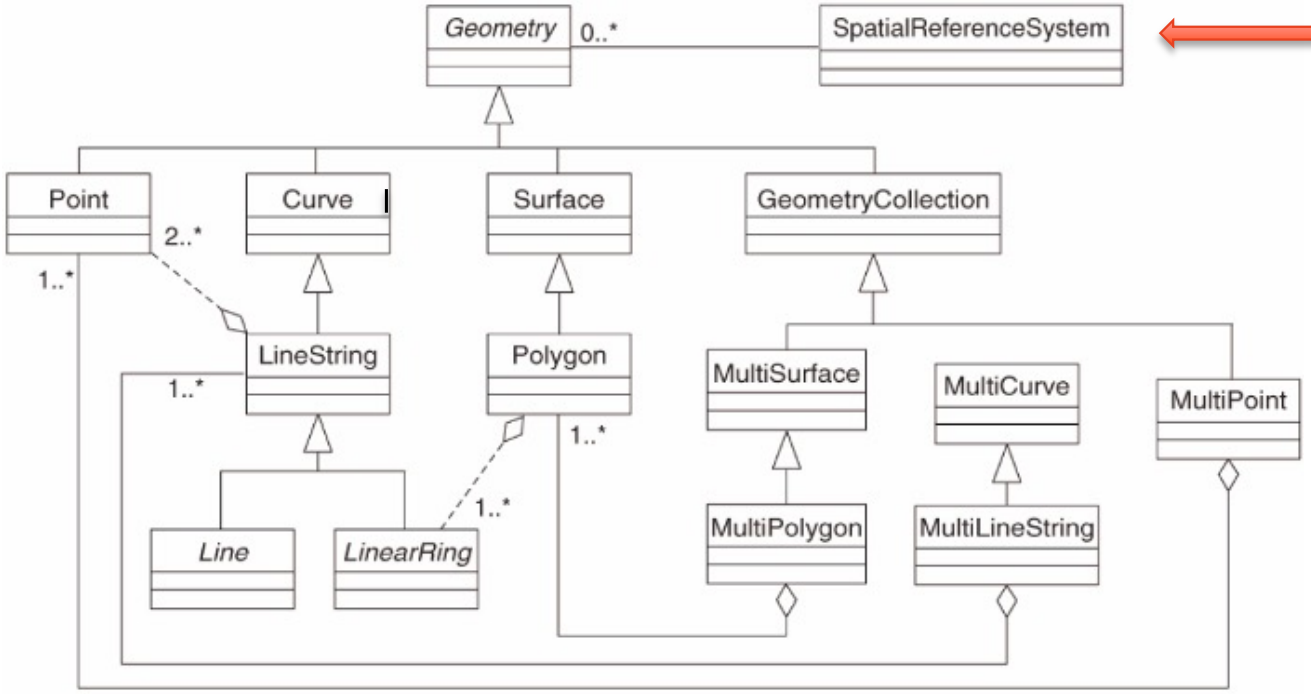
```
INSERT INTO superhero VALUES ('Catwoman', ST_SetSRID(ST_MakePoint(41.87,-87.634), 4326);
SELECT superhero.name
  FROM city, superhero
 WHERE ST_Contains(city.geom, superhero.location)
   AND city.name = 'Gotham';
```

WGS84 versus Australian GDA94

- WGS84 is used by the GPS system
- The official geodetic datum (coordinate system) for Australia is GDA94 ("Geocentric Datum of Australia")
  - Based on IERS Terrestrial Reference Frame (ITRF), but fixed to a number of reference points in Australia.
  - ABS data will use GDA94
- Difference between WGS84 and GDA94:
  - "The spheroids used for WGS84 and GDA84 are also almost identical, and both systems are geocentric. Thus for most mapping, exploration and GIS uses, WGS84 and GDA94 coordinates will be the same. […] For precise surveys, however, the difference between WGS84 and GDA94 may be significant, and changes slowly over time. […] The difference between GDA94 and WGS84 is approximately 45cms in 2000."

    [http://www.geoproject.com.au/gda.faq.html]

OpenGIS Consortium (OGC) Data Model



SRID
- part of every geometry
- needs to match for spatial predicates

[Source: OGC Simple Features, 2016]