COMP9321 Data Services Engineering

Term1, 2023

Week 10: Final Wrap-up

Morty Al-Banna, CSE UNSW Australia

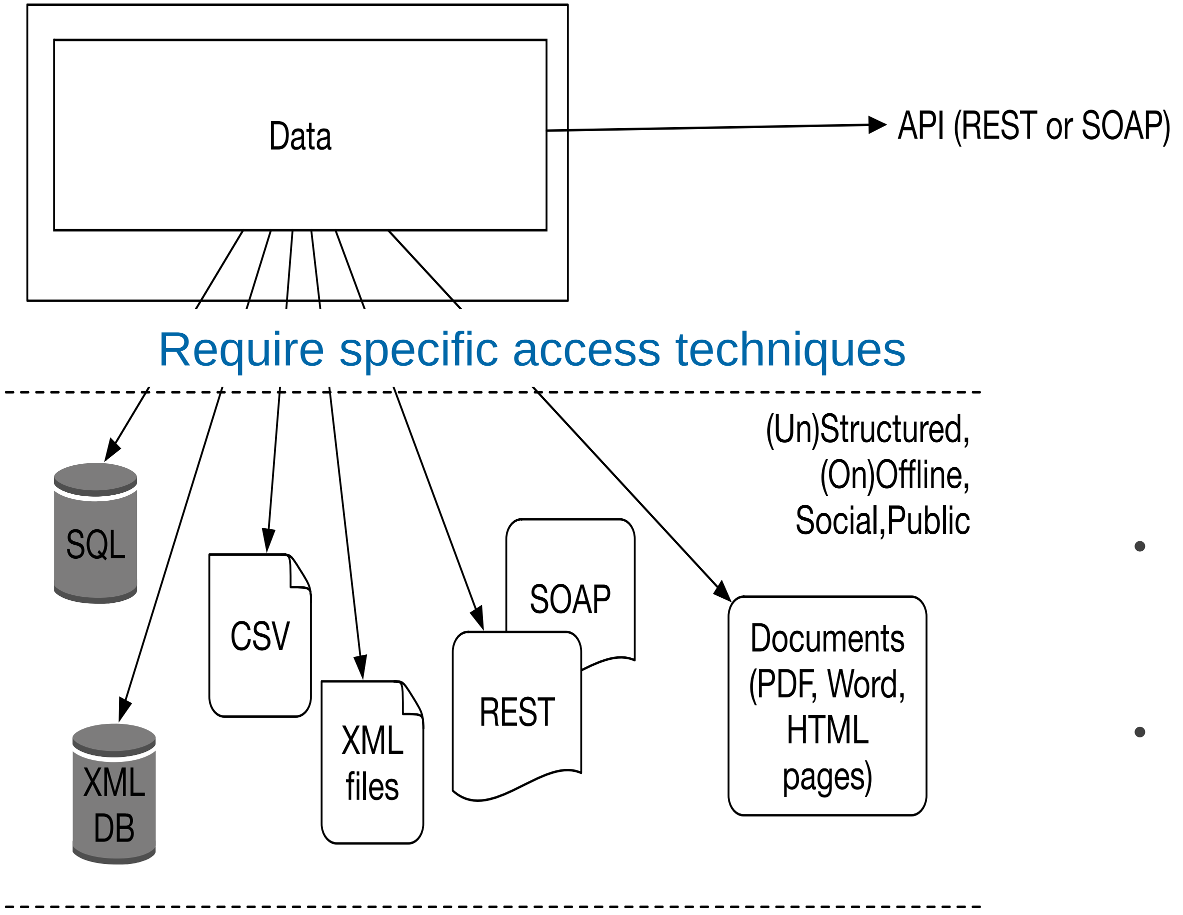
Course Aims

This course aims to introduce the student to core concepts and practical skills for engineering the data in service-oriented data-driven applications. Specifically, the course aims to answer these questions:

- *How to access and ingest data from various external sources?*
- *How to process and store the data?*
- *How to curate (e.g. Extract, Transform, Correct, Aggregate, and Merge/Split) and publish the data?*
- *How to apply available analytics to the data?*
- *How to visualize the data to communicate effectively*

Data Services – what is it about?

Two sides of a coin:



- Data integration/aggregation from multiple sources (data prep)
- Data publication for consumer access (API)

Understanding the Data (ask the right Questions)

- What is this dataset?
- What should I expect within this dataset?
- Basic concepts (e.g., domain knowledge)
- What are the questions that I need to answer? (Are there any questions?)
- Does the dataset have some sort of a schema? (utilize domain knowledge)

Obtaining Data

Useful data can be found in many places

- on the Web, possibly via an API
- in documents in a file system
- in spreadsheets
- in videos….etc. etc. etc.

and in a variety of formats

- Unformatted text (in files)
- PDF documents (in files)
- HTML documents (web pages)
- XML documents (via web APIs)
- JSON data (often via web APIs)
- CSV data files (spreadsheets)

Relational Model vs. “NoSQL” Models

Relational Model (more or so synonymous with SQL)

- The best known, probably the most successful data model which has proven itself in many aspects to be the data model of choice in many applications
- Data is organised into relations (table) where each relation holds an unordered collection of tuples (rows)

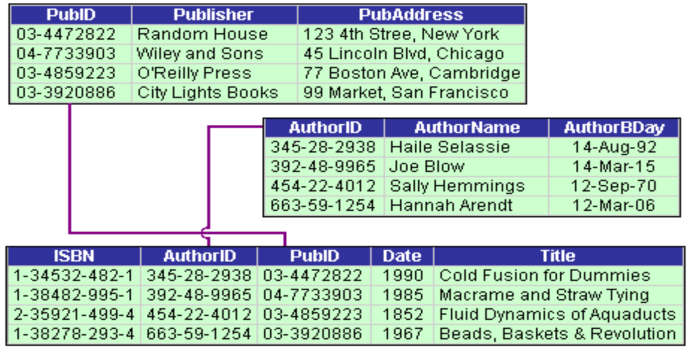
Based on solid theory and well engineered implementation -> many competing models have been proposed, but never managed to take over SQL

Built for business data processing

- Typical business transactions (airline reservations, stock keeping, etc.)
- Batch processing (invoicing, payroll, reporting, etc.)

Turned out it was still generically applicable to many modern Web applications too

Hypothetical Relational Database Model



PubID	Publisher	PubAddress
03-4472822	Random House	123 4th Stree, New York
04-7733903	Wiley and Sons	45 Lincoln Blvd, Chicago
03-4859223	O'Reilly Press	77 Boston Ave, Cambridge
03-3920886	City Lights Books	99 Market, San Francisco

AuthorID	AuthorName	AuthorBDay
345-28-2938	Haile Selassie	14-Aug-92
392-48-9965	Joe Blow	14-Mar-15
454-22-4012	Sally Hemmings	12-Sep-70
663-59-1254	Hannah Arendt	12-Mar-06

ISBN	AuthorID	PubID	Date	Title
1-34532-482-1	345-28-2938	03-4472822	1990	Cold Fusion for Dummies
1-38482-995-1	392-48-9965	04-7733903	1985	Macrame and Straw Tying
2-35921-499-4	454-22-4012	03-4859223	1852	Fluid Dynamics of Aquaducts
1-38278-293-4	663-59-1254	03-3920886	1967	Beads, Baskets & Revolution

Image: http://gnosis.cx/publish/programming/xml_matters_8.html

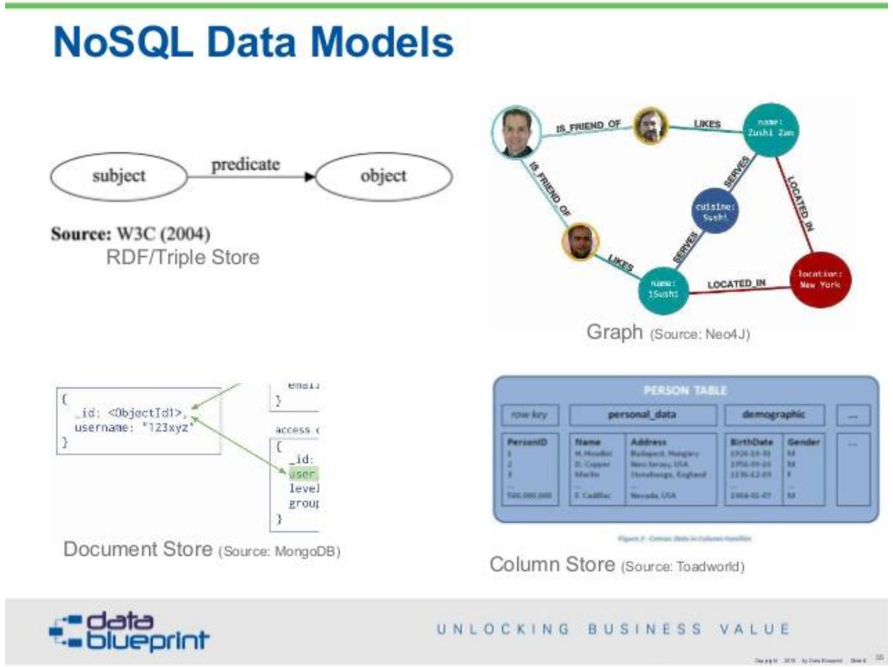
Relational Model vs. “NoSQL” Models

The rise of NoSQL … (since 2010 or so)

- Refers to a host of technologies that implement distributed, “non-relational” databases

Why NoSQL?

- A need for greater scalability – very large datasets or very high ‘write’ throughput
- A need for more expressive and dynamic data model
- Usually do not require a fixed table schema nor do they use the concept of joins
- All NoSQL offerings relax one or more of the ACID properties



NoSQL Data Models

Source: W3C (2004)
RDF/Triple Store

Graph (Source: Neo4J)

Document Store (Source: MongoDB)

Column Store (Source: Toadworld)

data blueprint

UNLOCKING BUSINESS VALUE

Designing Data-Intensive Applications, by Martin Kleppmann

Image: https://www.slideshare.net/Dataversity/trends-in-data-modeling

Conventional Definition of Data Quality

Accuracy

- The data was recorded correctly.

Completeness

- All relevant data was recorded.

Uniqueness

- Entities are recorded once.

Timeliness

- The data is kept up to date.
  - Special problems in federated data: time consistency.

Consistency

- The data agrees with itself.

Data Cleansing

- Dealing with Missing Data
- Removing Unnecessary Data (rows, or columns)
- Formatting data

Manipulating the data

- Merging Data
- Applying a function to data
- Pivot tables
- Change the index of a dataframe
- Groupby

Data Visualisation

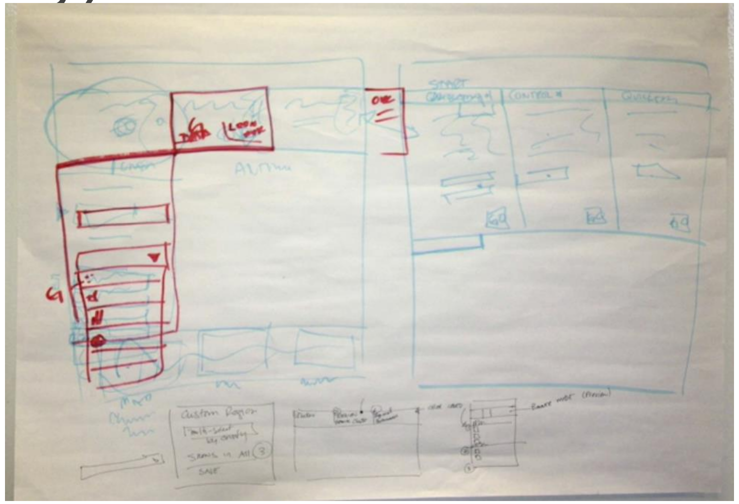
Referring to any visual representation of data that is:

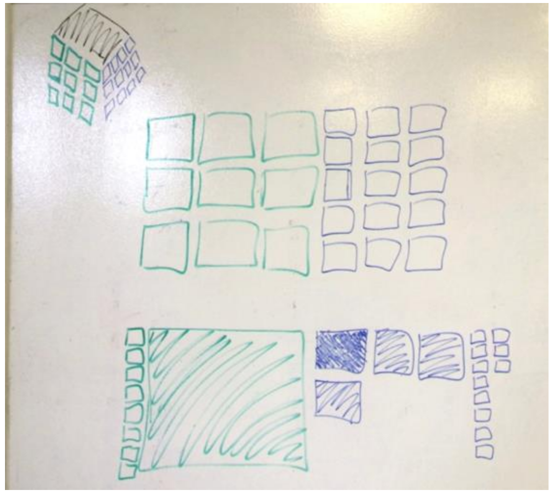
- algorithmically drawn (may have custom touches but is largely rendered with the help of computerized methods);
- easy to regenerate with different data (the same form may be repurposed to represent different datasets with similar dimensions or characteristics);
- often aesthetically simple (data is not decorated); and
- relatively data-rich (large volumes of data are welcome and viable, in contrast to infographics).

https://www.safaribooksonline.com/library/view/designing-data-visualizations/9781449314774/ch01.html

What You Need to Cater for Data Visualisation

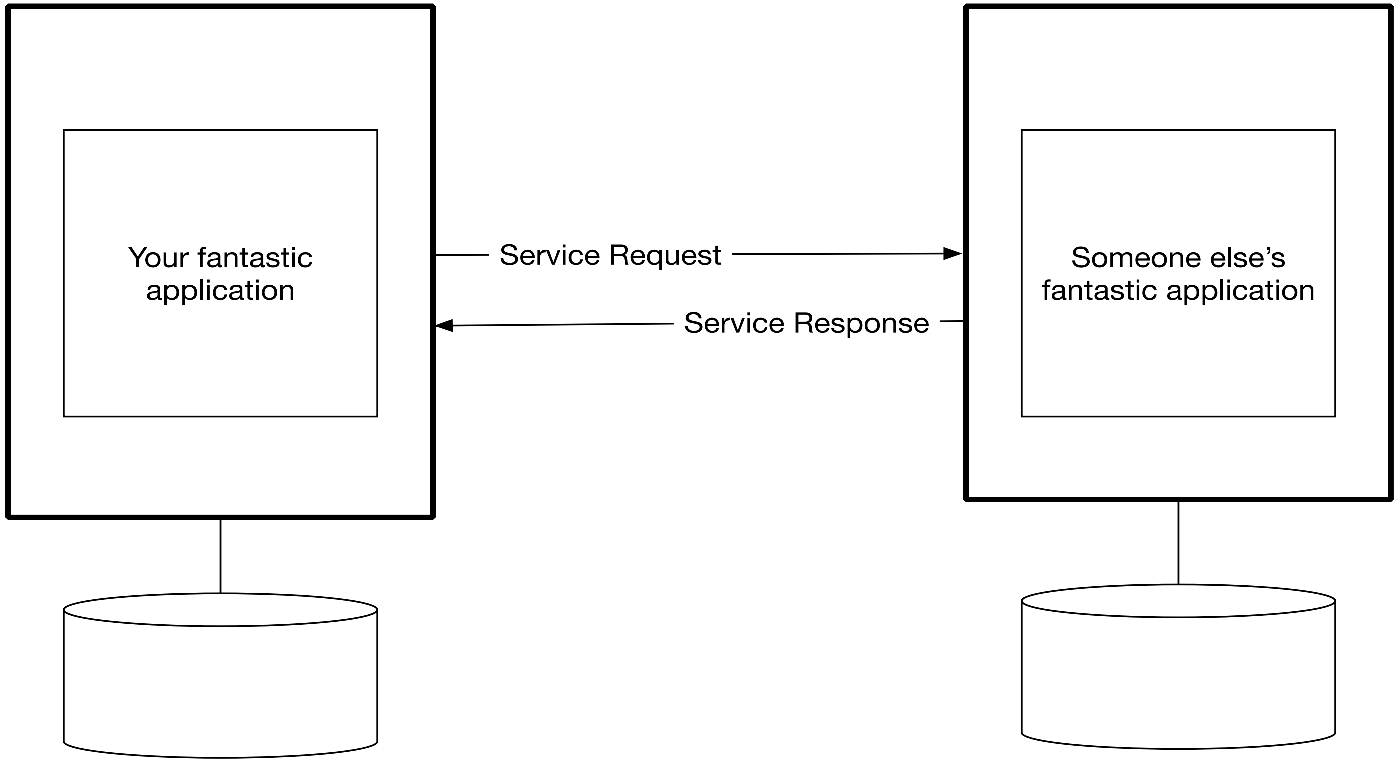
- Accuracy is important, having a clear story to tell is important
- You need to be ready to do some basic data prep and pre analysis before visualisation
- Knowing the right paradigm (form) to use for the story
- Aware of your own limitation as ‘non-expert’ (visualisation is not easy)





Actually, a lot of experts recommend “sketching the idea out” with pen and paper.

API – Application Programming Interface


Your fantastic application
Service Request
Service Response
Someone else's fantastic application

- The interface is not meant for human interactions – there is another program on the other side → implication of this: you must have a clear contract (e.g., IOU Alice Bob 100)
- These days companies use APIs internally (private APIs) as well as exposing them externally (public APIs)

What is and Why a RESTful Service

REST is an architectural style of building networked systems - a set of architectural constraints in a protocol built in that style.

The protocol in REST is HTTP (the core technology that drives the Web)

Popular form of API ... It is popularised as a guide to build modern distributed applications on the Web – let’s work with the components that the Web itself is built in.

REST itself is not an official standard specification or even a recommendation. It is just a “design guideline” for building a system (or a service in our context) on the Web

Architectural Constraints of REST

1. Client-Server
2. Uniform Interface
3. Statelessness
4. Caching
5. Layered System
6. Code on demand (optional)

Designing RESTful APIs

A well-designed API should make it easy for the clients to understand your service without having to “study” the API documents in-depth.

self-describing, self-documenting as much as possible

the clients are developers like yourself, so probably they would like to have an API that is easy to pick up and go

The RESTful service principles actually give us a straightforward guideline for designing the Web API ...

“Clean, Clear, Consistent” are the key

REST API Security

- REST API security is Important, Why?
- It matter enough that OWASP included many instances in their web security Top ten related to APIs and they have the REST Security cheat sheet.
- REST relies on the elements of the Web for security too (Check OWASP top 10)
- HTTPS (SSL) “Strong” server authentication, confidentiality and integrity protection The only feasible way to secure against man-in-the-middle attacks
- Any security sensitive information in REST API should use SSL
- Other things to remember (Input Validation, Methods restriction, logging)

https://www.owasp.org/images/7/72/OWASP_Top_10-2017_%28en%29.pdf.pdf
https://www.owasp.org/index.php/REST_Security_Cheat_Sheet

REST APIs and Security (Cont'd)

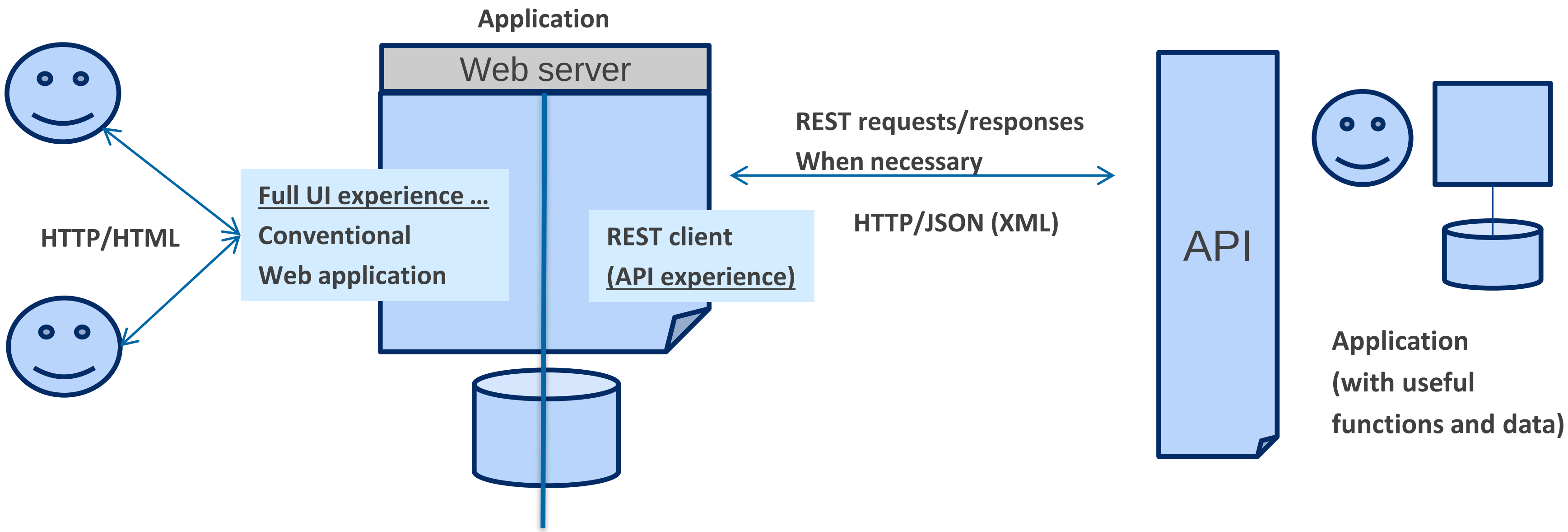
API developers at least must deal with authentication and authorisation:
Authentication (401 Unauthorized) vs. Authorisation (403 Forbidden):

Common API authentication options:

- HTTP Basic (and Digest) Authentication: IETF RFC 2617
- Token-based Authentication
- API Key [+ Signature]
- OAuth (Open Authorisation) Protocol - strictly uses HTTP protocol elements only

https://www.owasp.org/index.php/Transport_Layer_Protection_Cheat_Sheet

RESTful Service Client



Application

Web server

HTTP/HTML

Full UI experience …

Conventional

Web application

REST client

(API experience)

REST requests/responses

When necessary

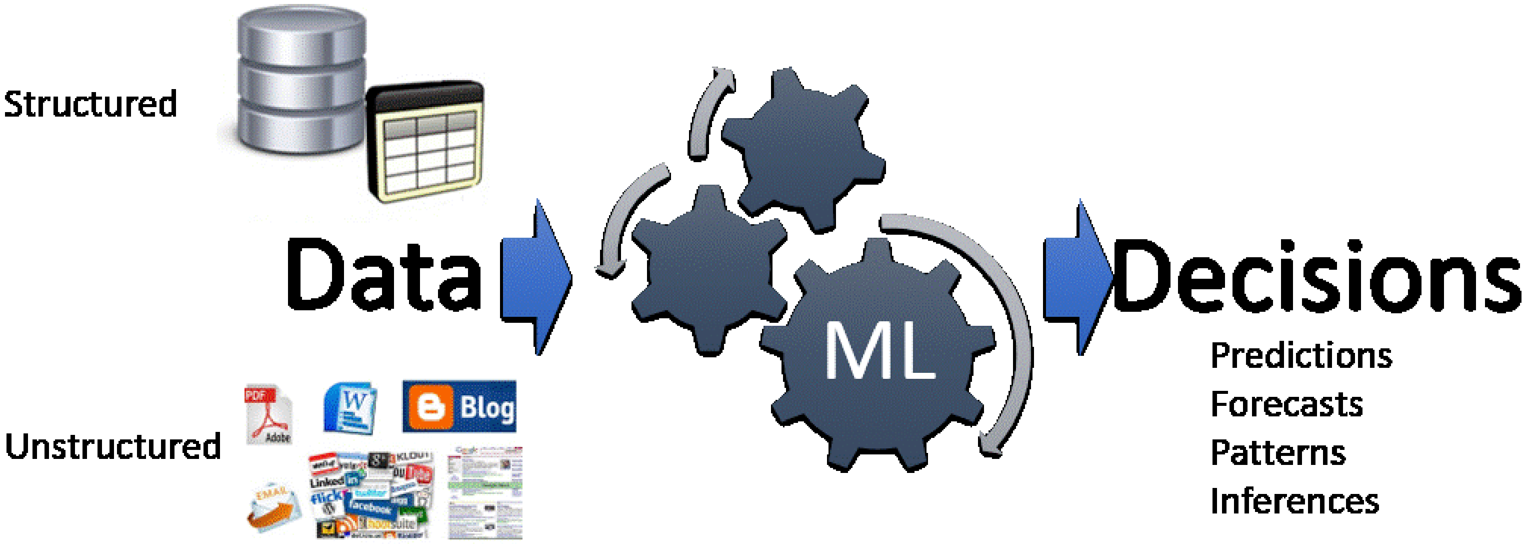
HTTP/JSON (XML)

Application

(with useful

functions and data)

Machine Learning for Data Analytics

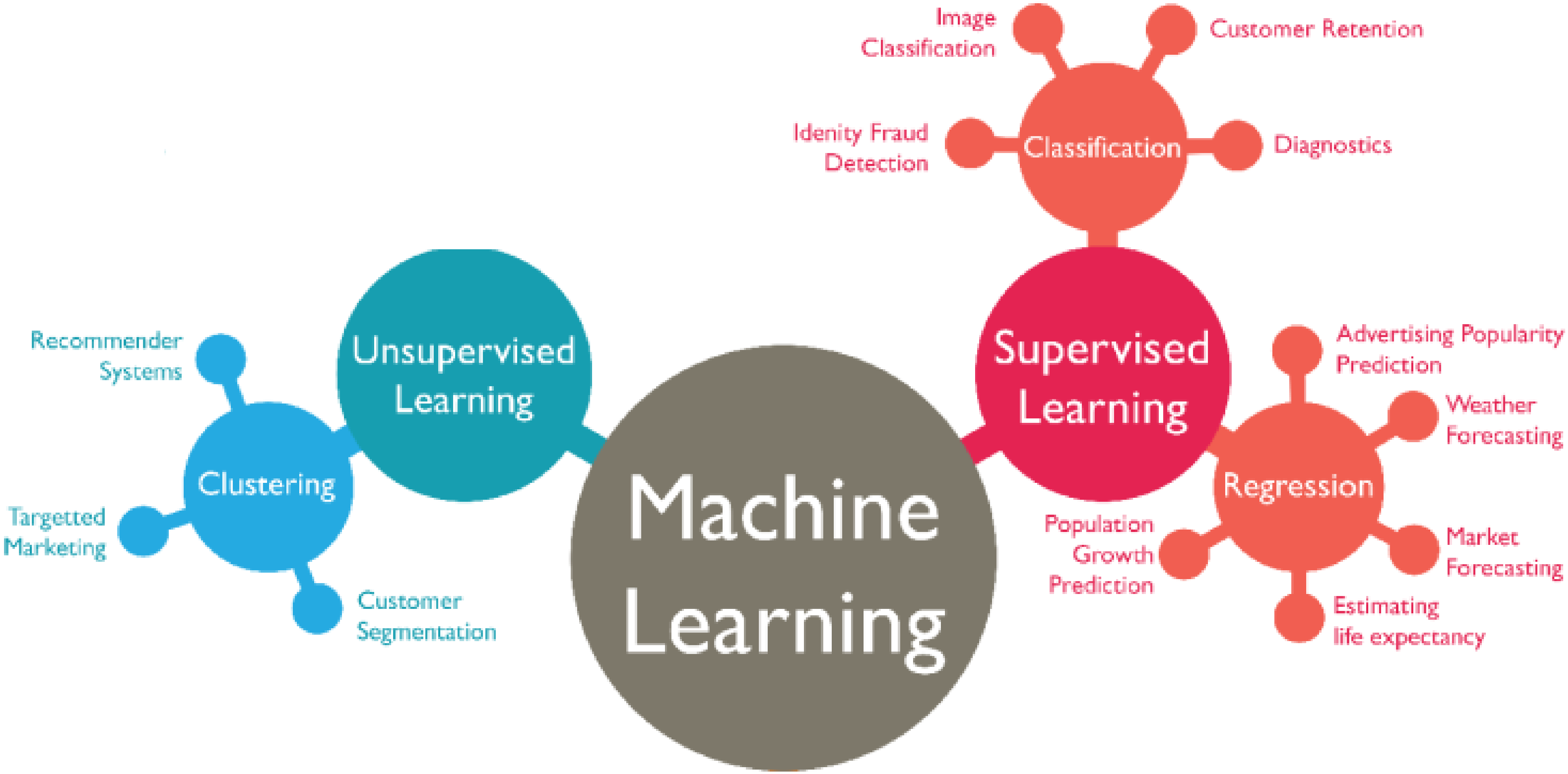


https://www.datasciencecentral.com/profiles/blogs/decoding-machine-learning-methods

Machine Learning for Data Analytics

1. **Define** and **Initialize** a Model
2. **Train** your Model (using your training dataset)
3. **Validate** the Model (by prediction using your test dataset)
4. Use it: **Explore** or **Deploy** as a web service
5. **Update** and **Revalidate**

https://www.slideshare.net/SetuChokshi/machine-learning-101-74017810



https://www.datasciencecentral.com/profiles/blogs/decoding-machine-learning-methods

Recommender Systems

Given:

- User model (e.g. ratings, preferences, demographics, situational context)
- Items (with or without description of item characteristics)

Find:

- Relevance score. Used for ranking.

Deep Learning

Neural networks define functions of the inputs (hidden features), computed by neurons

Artificial neurons are called units

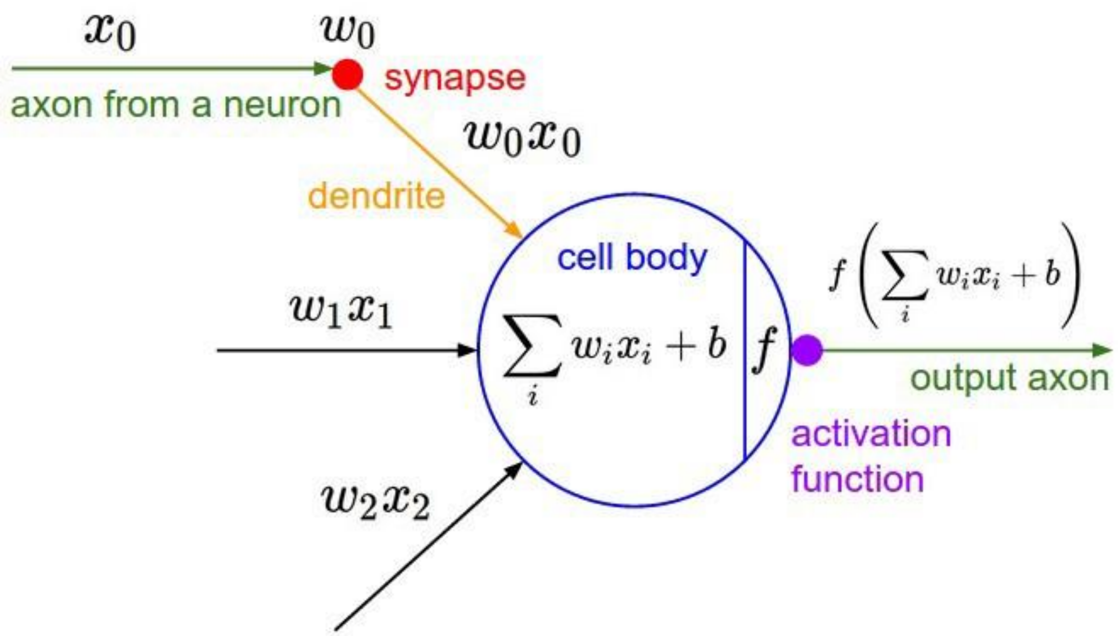


Figure : A mathematical model of the neuron in a neural network

[Pic credit: http://cs231n.github.io/neural-networks-1/]

Wrap-up Example

- **Task:** I want to find candidate security researchers to discover vulnerabilities in my systems

- **Possible Data sets**:
    - Hackerone full disclosure
    - OpenBugBounty
    - ExploitDB

Example Cont’d

- Working with the Data
  - Understand the Data: visit the platforms offering the data and start browsing or inspecting the few records. Understand what is the type of the data and the schema and the meta-data.
  - Acquire the Data: Download the data file, query the data source, or dump the database if possible. Store the data in your selected mode of storage.
  - Clean the Data: drop the useless fields or the ones with missing or corrupted data.
  - Augment the Data: it might be useful to combine two data sources to acquire better insight.

Example Cont’d

- Design the Solution
    - What is the problem that you need to solve. What are the features you need to provide.
    - What kind of Data analytics you need to perform
        - Visualize your Data
        - What Machine learning model
        - What are your constrains
    - Design the API endpoints to allow the consumption of your service
    - Document everything

Assessment

Assessment:

- 40% Take home exam: individual assessment.
- 50% on assignment work (including bonus marks)
  - Assignment1 on Working with Data 15%
  - Assignment2 on building a REST service 20%
  - Assignment3 Machine Learning for Data Analytics 15%
- 10% on 5 online quizzes (WebCMS-based quiz system, ‘open’ test)

Final Mark = quizzes + assignments(including the bonus) + exam

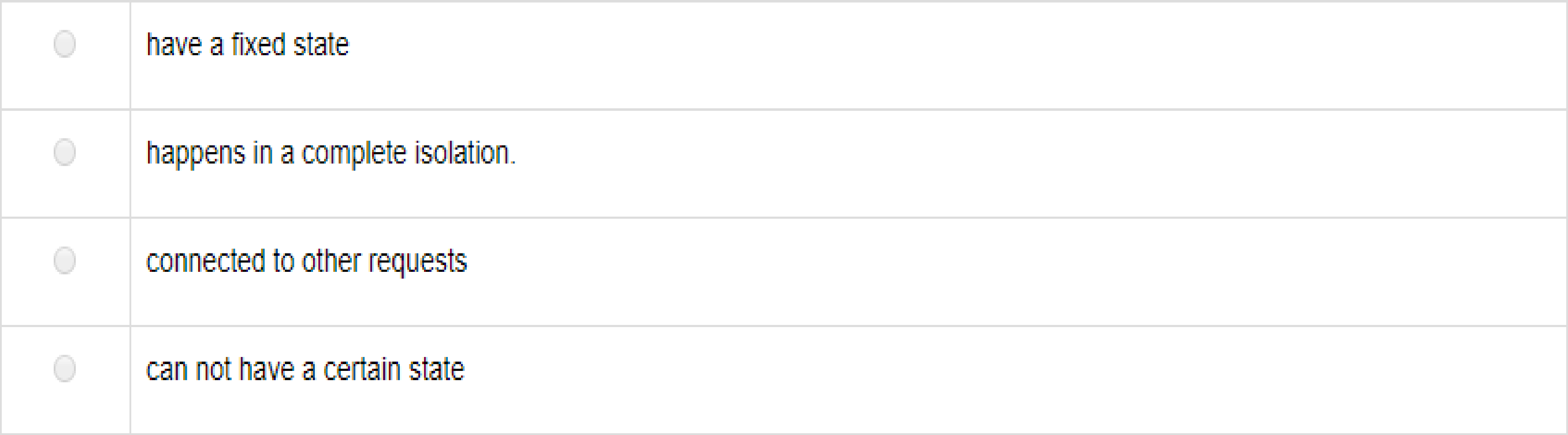
The Final Exam

- Duration: 2 hours + 15 minutes reading and submission
- Online Exam via Inspera
- Types of questions:
  - Multiple Choice using Moodle Quiz and Question Bank
  - Written Answers

The Final Exam (examples of Questions)

- Multiple Choice Question:

In REST Services, Stateless means that every HTTP request



○	have a fixed state
○	happens in a complete isolation.
○	connected to other requests
○	can not have a certain state

The Final Exam (examples of Questions)

Questions: Consider the following HTTP request invoking a POST method of a RESTful API:

```
POST /orders HTTP/1.1

Host: api.coffeehouse.com

Content-Type: application/xml

<order>

<drink>latte</drink>

</order>
```

Write down the content of the HTTP response that you would return as the result

The Final Exam (examples of Questions)

Question: Assume you are creating a Data Service to predict the Price of Car Fuel in NSW. You acquired the CSV file containing the Price of Fuel in NSW for the past 3 year and another CSV file containing the prices of Crude Oil for the same period (sample of both CSV provided in form of a table snippet).

1. Provide a brief description of how you are going to approach the problem
2. Explain what are the specific pre-processing steps that you need to perform to increase the value of the data provided.
3. Explain what is the machine learning model you are going to consider? Explain why you chose this Machine learning Model and why do you see it fit to solve the problem? Explain if there are any modification to the data you need to perform to be able to fit the model.

Supplementary Exam Policy

Supp Exam is only available to students who:

- DID NOT attend the final exam
- Have a good excuse for not attending
- Have documentation for the excuse

Submit special consideration within 72 hours (via myUNSW with supporting docs)

Everybody gets exactly one chance to pass the final exam. For CSE supplementary assessment policy, follow the link in the course outline.

Warning…
Warning

Huge Thanks Coming Your Way…

Thanks for the Teaching Team

- Course Administrator
  - Mohammad Ali Yaghub Zade Fard

- Tutors
  - Mohammad Ali Yaghub Zade Fard
  - Dylan Sanusi-Goh
  - Tim Arney

That's all Folks!