Data on the Web

Andre Freitas
andre.freitas@manchester.ac.uk

September, 2019
The Web

- A network of interlinked computers, protocols, software, representation and data.

- A socio-cultural phenomenon.

- *The* new medium.

- Changed society over the last 20 years.
What Happens in an Internet Minute?

- 639,800 GB of global IP data transferred
- 204 million Emails sent
- 47,000 App downloads
- 583,000 In sales
- 20 New victims of identity theft
- 61,141 Hours of music
- 3 million Photo uploads
- 20 million Photo views
- 320+ New Twitter accounts
- 100,000 New hexlets
- 6 New Wikipedia articles published
- 135 New mobile users
- 100+ New LinkedIn accounts
- 277,000 Logins
- 6 million Facebook views
- 2+ million Search queries
- 1.3 million Video views
- 30 Hours of video uploaded

And Future Growth is Staggering

Today, the number of networked devices = the global population
By 2015, the number of networked devices = 2x the global population
In 2015, it would take you 5 years to view all video crossing IP networks each second
Web Facts (as of 2018)

- “~3.95 billion Internet users.”
- “2.62 billion social media users in 2018.”
- “~1.89 billion websites in the world.”
- “Netflix Bandwidth Usage Climbs to Nearly 37% of Internet Traffic at Peak Hours.”
Technological Strata

• Backend
  – From simple web pages to large complex programs!

• Formats and representations
  – Documents: HTML and CSS
  – Data: CSV, XML, JSON, RDF
  – Knowledge: OWL
On Data & Representation

Algorithms and data structures for accessing, transforming and analyzing data efficiently

- **Data:** CSV, XML, JSON, RDF
- **Tables**
- **Trees**
- **Graphs**
- **SQL**
- **XPath/XQuery**
- **SPARQL**
- **Query**
- **Abstractions**
- **Formats**
On Data & Representation

Algorithms and data structures for accessing, transforming and analyzing data efficiently

Data: CSV, XML, JSON, RDF

SQL, XPath/XQuery, SPARQL

Tables, Trees, Graphs

Querying, Modelling

COMP62421, COMP60411
Dialogue between Representations and Algorithms
ACS Theme: Data on the Web

This theme aims to provide students with core concepts, formalisms, and techniques to develop data-intensive applications on the Web.

- We examine:
  - web oriented & inspired technologies
  - design choices & their impact on relevant properties

- You get:
  - theoretical and conceptual understanding
  - appreciation of the design trade-offs
  - understanding of the main data architectures on the Web
Course Units in this Theme

• COMP60411: Modelling Data on the Web
  – Data models:
    ▪ which data model & format to use for what kind of data?
    ▪ how to manipulate it?
    ▪ …in a robust, suitable way

• COMP62421: Querying Data on the Web
  – Query Processing:
    ▪ how query answering systems are built
    ▪ how query processing is made efficient
    ▪ …for different data models
Pre-Requisites

• For both: a background in Databases
  – a good UG module “Fundamentals of Databases”
  – remember tables, SQL queries, Joins,…

• For both: being happy to
  – think things through
  – analyse pros & cons
  – understand technically challenging concepts
Modelling Data on the Web
COMP60411: Modelling Data on the Web

- Is about data models:

<table>
<thead>
<tr>
<th></th>
<th>Week 1</th>
<th>Week 2-4</th>
<th>Week 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>data model</td>
<td>Tables</td>
<td>Trees</td>
<td>Graphs</td>
</tr>
<tr>
<td>formalisms</td>
<td>SQL</td>
<td>XML, PSVI, JSON</td>
<td>RDF</td>
</tr>
<tr>
<td>schema languages</td>
<td>SQL</td>
<td>RelaxNG, XML Schema, Schematron</td>
<td>RDFS</td>
</tr>
<tr>
<td>manipulation</td>
<td>SQL</td>
<td>SAX, DOM, XQuery</td>
<td>SPARQL</td>
</tr>
</tbody>
</table>

- When to use which data model/formalism/…
- How to use them **well:**
  - robustness
  - affordances
  - error handling
COMP60411: Modelling Data on the Web

- Starts next week
- Taught by Bijan Parsia and Uli Sattler

**Assessment:** 50% coursework, 50% exam

- Weekly coursework:
  - quiz
  - short essay
  - programming/data manipulation task
  - modelling task

- We expect you to read
Querying Data on the Web
COMP62421: Querying Data on the Web

- This course is about **Query processing** over different data models (tables, trees, graphs).
- It also covers some **emerging paradigms in databases** (e.g. NOSQL, stream architectures).

<table>
<thead>
<tr>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>Week 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relational Query Processing (Fundamentals)</td>
<td>Query Execution</td>
<td>Massively-Parallel Schemes</td>
<td>Graph Databases</td>
<td>Streaming &amp; Event Centric Platforms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Challenges &amp; Approaches</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query Optimization</td>
<td>Query Processing Using XQuery</td>
<td>NOSQL Databases</td>
<td>Query Processing using SPARQL</td>
<td>Contemporary Data-Intensive Architectures</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
COMP62421: Querying Data on the Web

- Starts after Reading week
- Taught by Andre Freitas and Bijan Parsia

**Assessment**: 50% coursework, 50% exam

- Weekly coursework:
  - quiz
  - short essays / explanations
  - problem solving questions
  - some programming

- We expect you to read
Research Related to this Theme

- UoM is a Leading Research Centre in Ontology Engineering, Language Design and Semantic Applications
- Instrumental in W3C Standardization
  - Web Ontology Language OWL
  - Simple Knowledge Organisation Systems SKOS
  - SPARQL Query language
- Tool Development
  - OWL API, Protégé, FaCT++, SWOOP
- Applications
  - (Semantic) Web-based information systems in Life Sciences & Biohealth
Data on the Web fits nicely with

1. Semester: Learning from Data
   or other themes

2. Semester: Ontology Engineering & Automated Reasoning
   or other themes
Barack Obama / Height

1.85 m

People also search for

Donald Trump 1.9 m
Michelle Obama 1.8 m
Bill Clinton 1.88 m
Employability Skills

• Understanding the interplay between data, representation and algorithms at scale.
  – Common need in Big Data and AI scenarios

• You will learn and use tools, but the main goal is to understand the principles.
Is it for me?

• This theme **is not** for those who:
  – just want to build web sites
  – don’t like to “get their hands dirty”
  – don’t like to read around the subject

• This theme **is** for those who are:
  – seeking a deeper understanding of the principles & concepts and technologies used to support the continuing evolution of the Web
  – like understanding & discussing trade-offs and modelling/design choices