



University of Dublin
Trinity College

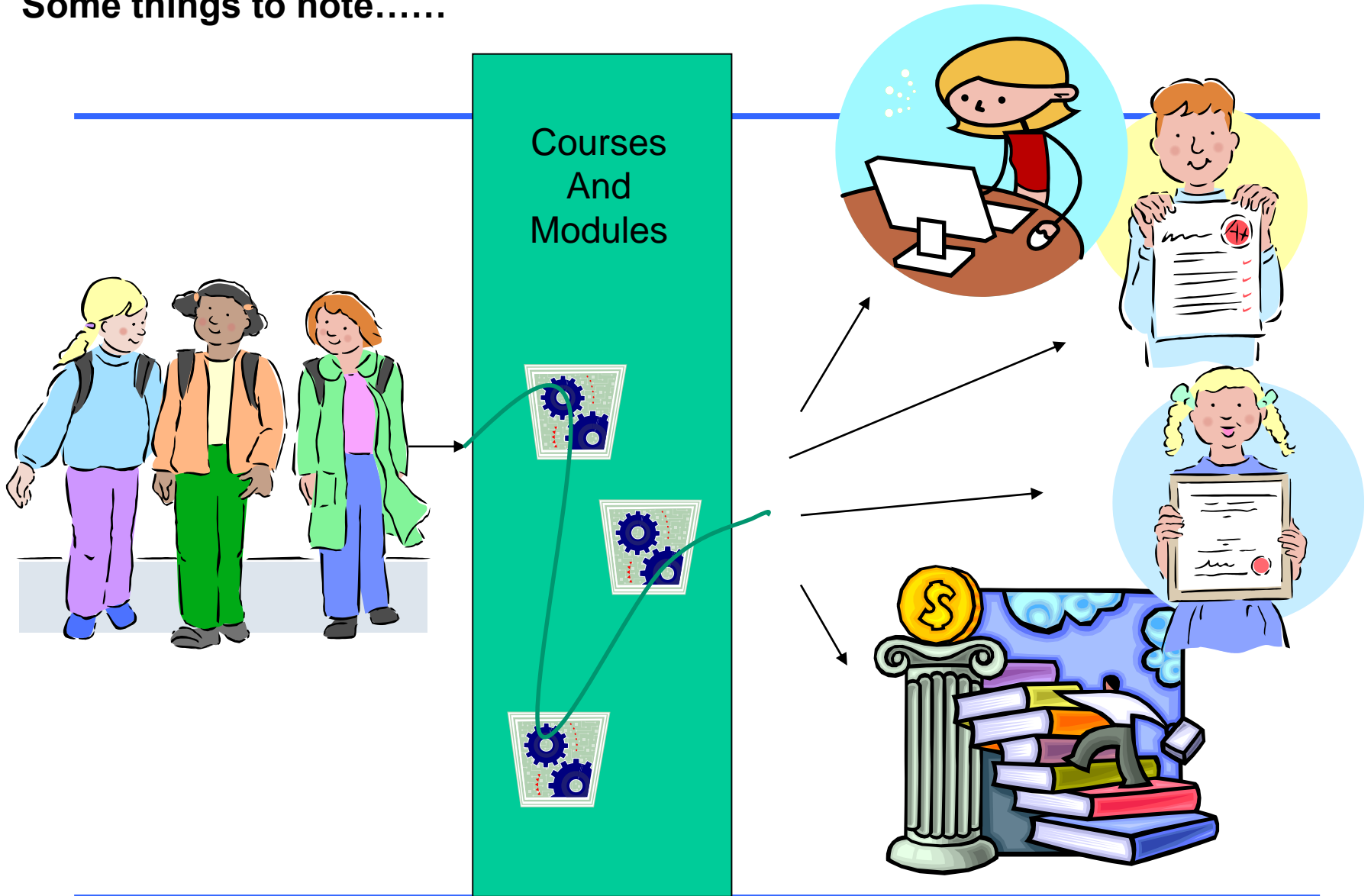


CS2041: Information Management I

An Introduction to the module
2018-19

Gaye Stephens gaye.stephens@tcd.ie

Some things to note.....



Expectations

**Bring Paper and Pen
to every session!**

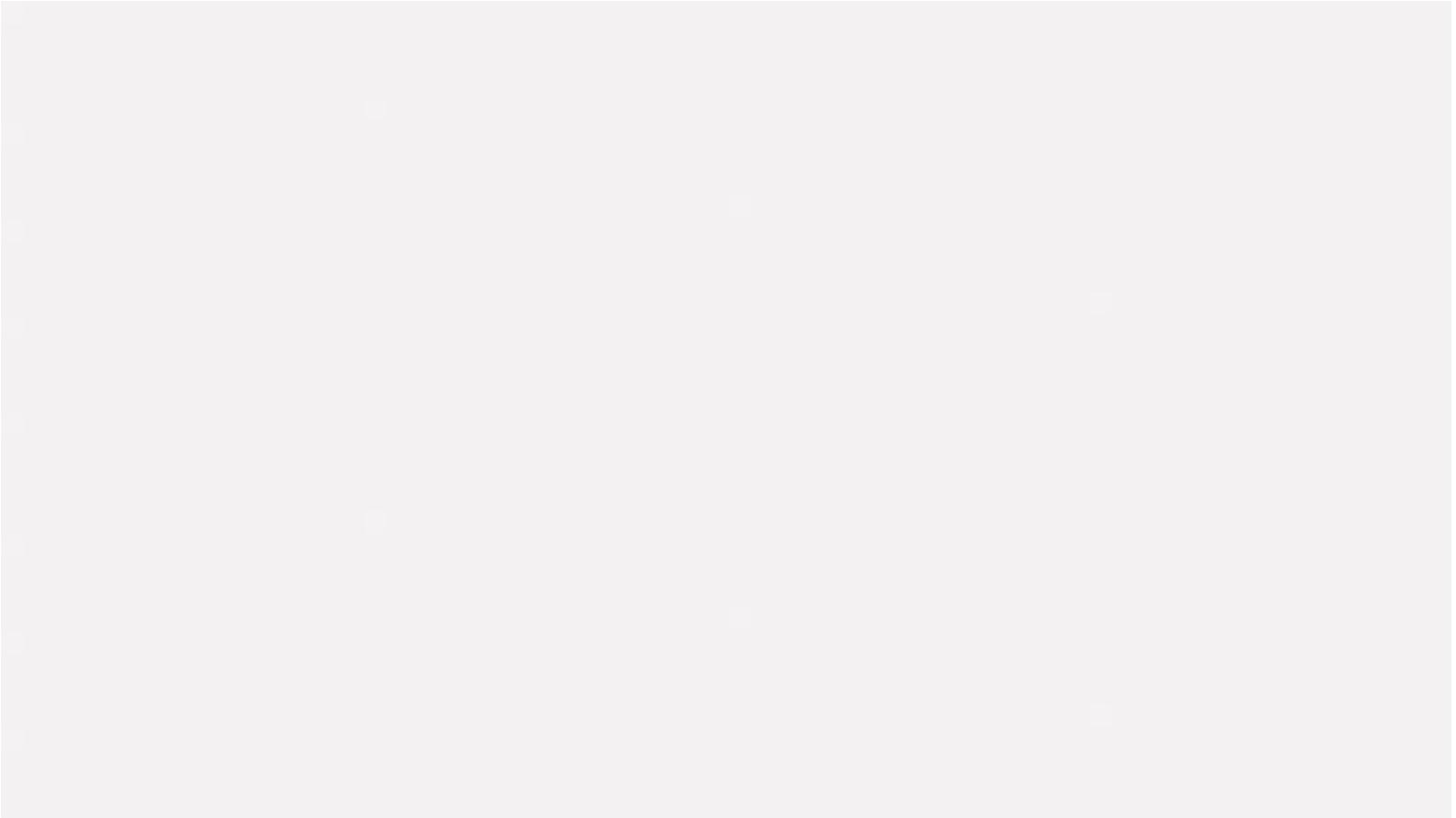


This is a respectful, supportive and constructively critical environment for people to learn.

**Shut your laptop during Exercises,
Student Presentations and
Tutorials**



What is information?- One Perspective



What is the difference between Data, Information and Knowledge?

Data:

- Raw; building blocks of information
- Unprocessed information

90 **Rehab**
Smith

Information:

- Data associated together to convey some meaning
- Basic Unit of Communication

Heart **Surname** **Location**
Beats per
minute

Knowledge:

- Interrelating and “understanding” information

Normal- If **Heart** **Gym**
Male and
HBPM **Patient**
<=70
>=50

Core Concepts

ORGANISATION

How data represented/associated

METADATA

Data about what the data is

ACCESS

How get at the data efficiently

Data Storage

Solid
State
- Chip
based



Organisation: Series of Bytes
Metadata: Allocated/Unallocated space
Access: Block transfers, Buffering etc.

Optical Discs – Laser based



Hard Disc
- Magnetism based

Organisation: File

Metadata: What parts file stored where

Access: Read/Write APIs for bytes

Operating System

File system treats each file as series of bytes

File Manager creates/maintains this view

File represented by a sequence of blocks together with a **file access table**

Application/Developer does not need to know the physical location on storage, just the logical name.

Block Number	Disk Address (Cylinder, Track, Sector)	Range of Bytes
0	1200, 1, 98	0–1023
1	1200, 1, 100	1024–2047
2	1200, 1, 102	2048–3071
3	1200, 2, 56	3072–4095
4	490, 0, 0	4096–5119
5	490, 3, 8	5120–6143

So, what software do you know that manages data?

So, what software do you know that manages data?

All applications

- File formats inherently organise data for particular applications: .xls, .doc, .mp4, .jpg, .eps, .exe etc.

Specialist data management applications

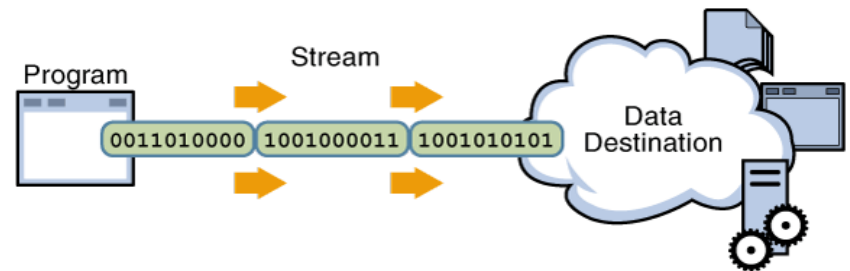
Your applications!

Maintaining structure in your own data file

Files just represent data as a series of bytes and will **lose the structure** that you might have imposed either logically or physically (e.g. as an object/field or record) unless you do something about it

Take an example:

```
Public class Movie {  
    // members  
    String title; int movieId; String genre;  
  
    // constructor  
    public Movie (String t, int i, String g) {  
        title = t; movieId = i; genre=g;  
    }  
};
```



So how can we encode the structure we want in the file itself?

Take a second..

How maintain structure within a file?

Maintaining structure in your own data file

There are many ways of adding structure to files, for example

- Choose a special character/delimiter that will not appear as a legitimate character within the information field and then insert that character into the file after writing each field... called **delimited-text field**
 - Use a fixed length for each information field (the size depending on field in question) and pad out when length of actual data value is less than the fixed length... called **fixed-length field**
 - Write the length of the value (in bytes) of the information field followed by the value in exactly that number of bytes... called **length-based field**
 - Write the name of the information field and then value both represented as delimited-text fields... called **identified field**
-

Turning Data into Information

Two distinct approaches

1. Deliberately associate data together to turn into information... to serve a range of known information needs and carefully manage. Let's call this Structured approach.
e.g. excel, databases, datawarehouses
2. Bring loosely managed data together to serve a specific information need, using information retrieval techniques. Let's call this Unstructured approach.
e.g. search engines

Representation: Structured vs Unstructured

Name	Gender	Salary	Date of Birth
String	Char	Int	Date
Kima Greggs	F	\$25,000	11/03/1978
Jimmy McNulty	M	\$20,000	18/07/1976
Cedric Daniels	M	\$50,000	23/10/1973

“James Joyce was born in Dublin in 1882. His works include Ulysses and Finnegans Wake. He died in 1941 in Switzerland.”

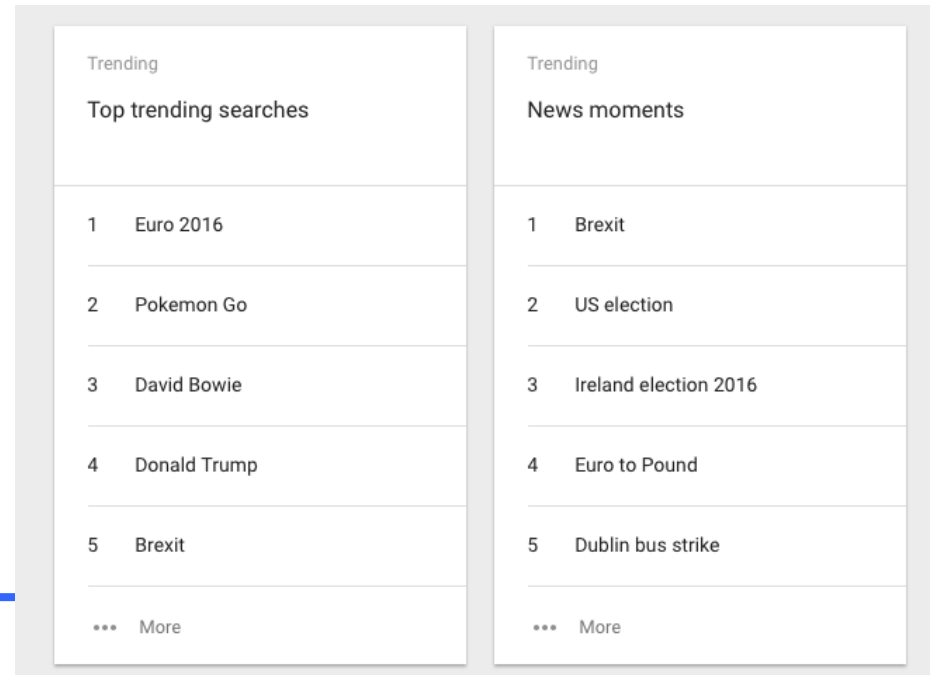
Nature of Querying: Structured vs Unstructured

Artificial Language
Known Data Types
Exact Criteria

Keyword based,
increasingly phrase
based

Google Zeitgeist 2016 for Ireland

SQL (or Xquery)
SELECT Name
FROM Character
WHERE Salary
BETWEEN 40000 AND
60000



The image shows a screenshot of the Google Zeitgeist 2016 for Ireland page. It is divided into two main sections: 'Top trending searches' and 'News moments'. Each section lists five items, numbered 1 to 5, with a 'More' link at the bottom of each list.

Trending	
Top trending searches	
1	Euro 2016
2	Pokemon Go
3	David Bowie
4	Donald Trump
5	Brexit
... More	

Trending	
News moments	
1	Brexit
2	US election
3	Ireland election 2016
4	Euro to Pound
5	Dublin bus strike
... More	

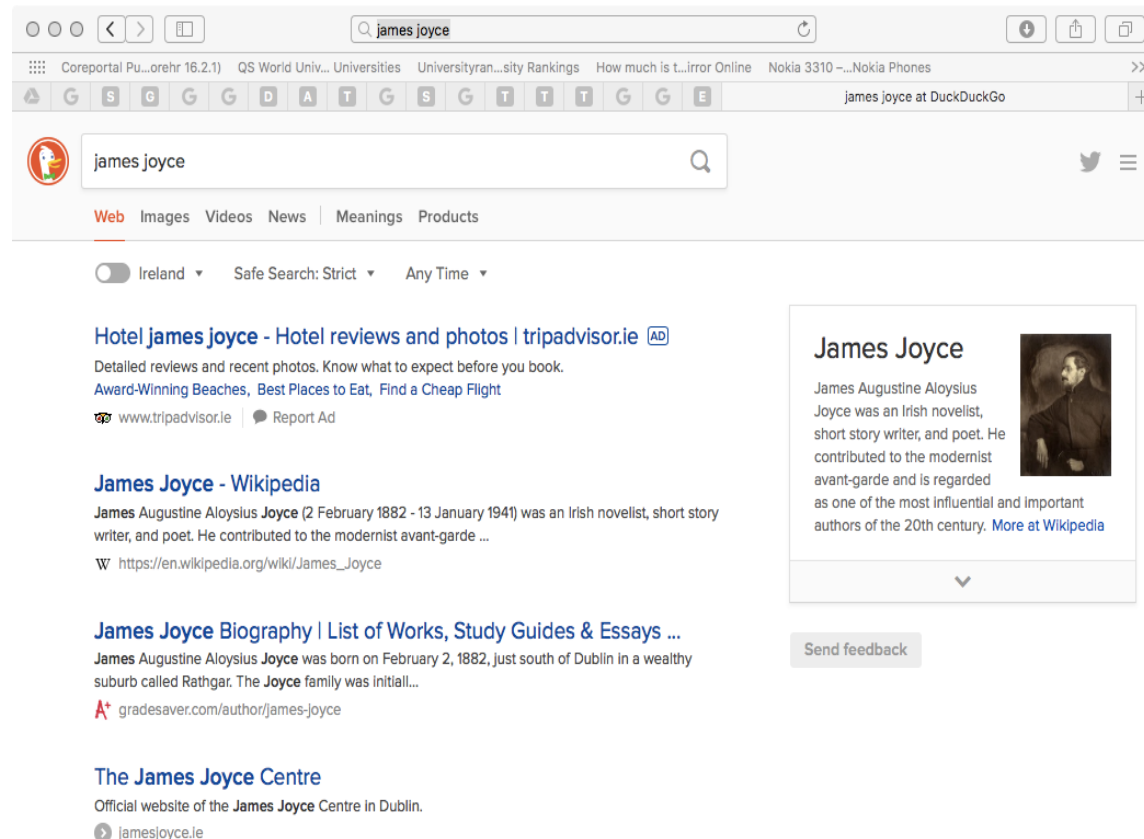
Nature of Results: Structured vs Unstructured

Structured

Definitive Results

Returns the Complete
Set of Data that
meets search
criteria

No estimation of
Relevancy



Structured Approach Specialist Software: Databases (DBs)

A combination of software and hardware
Optimised to reduce data to storage
transfer

Optimised to provide Transactional/ACID
properties upon the data

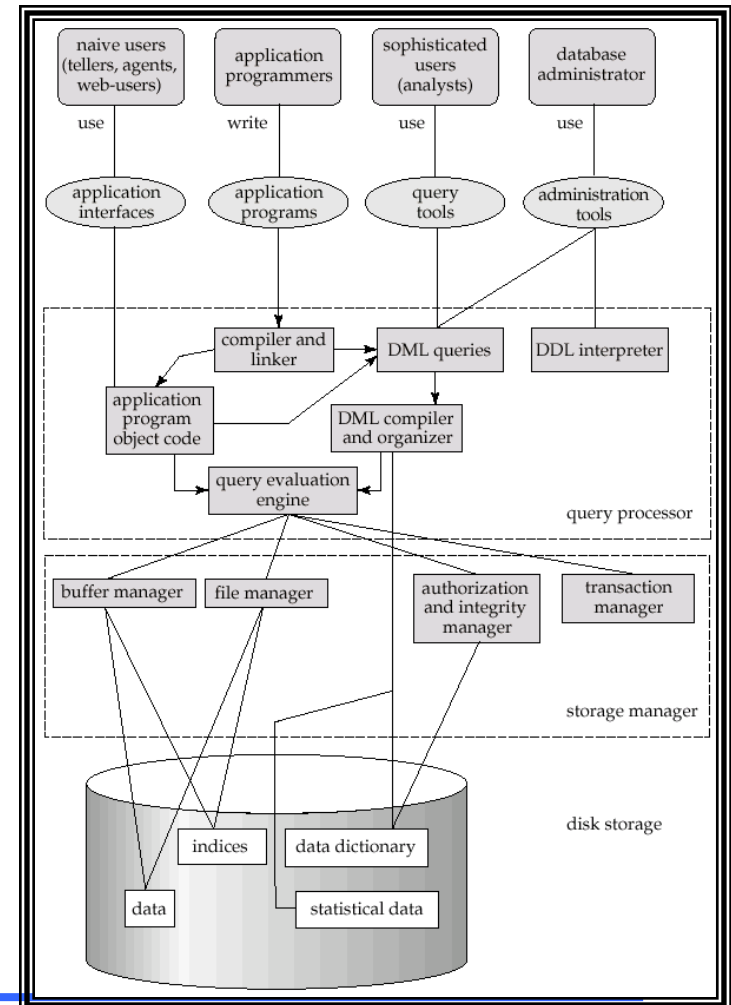
- **A**tomic, **C**onsistent, **I**solated, **D**urable

Designed to be administered and secure

Different Models

- Relational (by far the most popular)
- Networked (coming back in interest)
- Hierarchical (original model)
- Object-oriented

Primarily for operational purposes



Structured Approach Specialist Software: DataWarehouses (DWs)

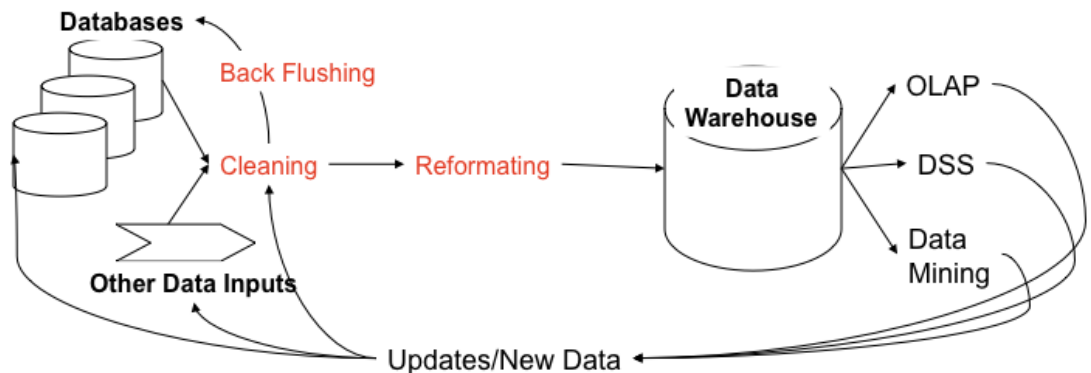
Data Warehouse is a subject oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions

Data Warehouse is a repository of data which is:

- Separate from operational systems and populated by data from these systems
- Provides a trend view of data
- Available entirely for the task of making data available to be interrogated by business users
- **Timestamped** and associated with defined periods of time, that is calendar periods or fiscal reporting periods
- Subject oriented around the high-level entities of the enterprise
- Accessible to users who have a limited knowledge of computer systems or data structures

Used for

- Data Mining
- Decision Support
- OLAP

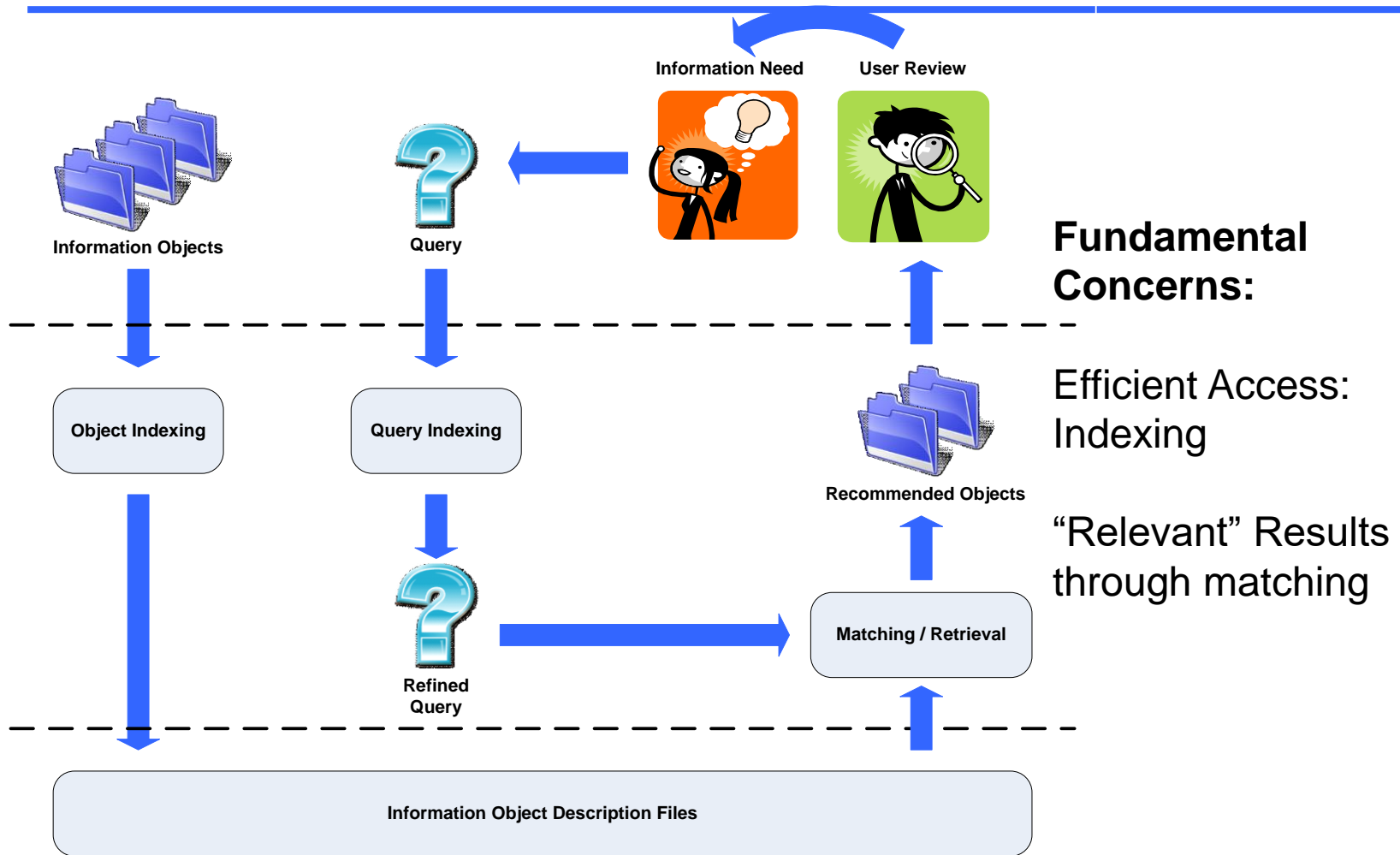


What does a Structured approach to data management involve?



- Architect
- Integrate
- Cleanse and Monitor
- Manage
- Associate
- Archive

Unstructured Approach: Information Retrieval



Common Challenges managing data for Enterprises and Individuals

Volume

Awash with data, consumers easily amassing terabytes and enterprises even petabytes of information.

Velocity

Often time-sensitive, data must be processed as it is streaming in order to maximize its value

Validity

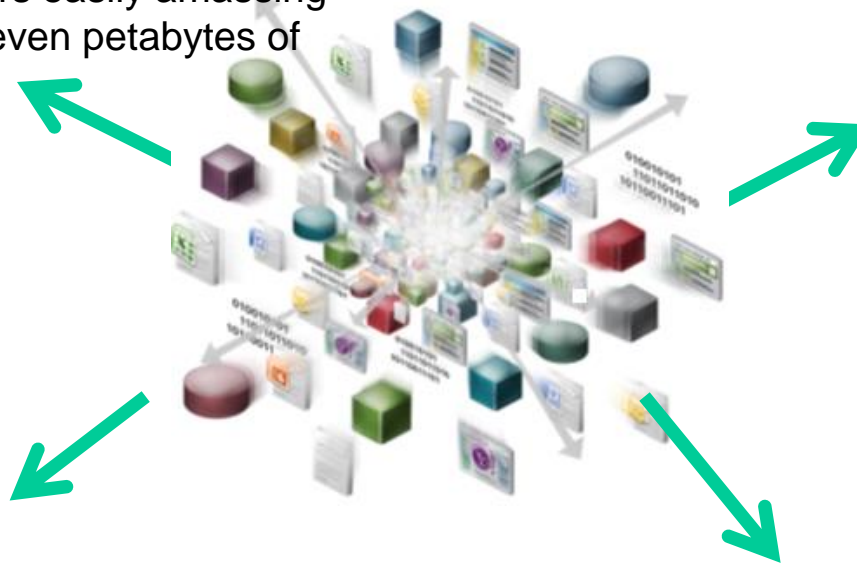
Data protection – consent and compliance;

Data privacy – what data an individual willing to share;

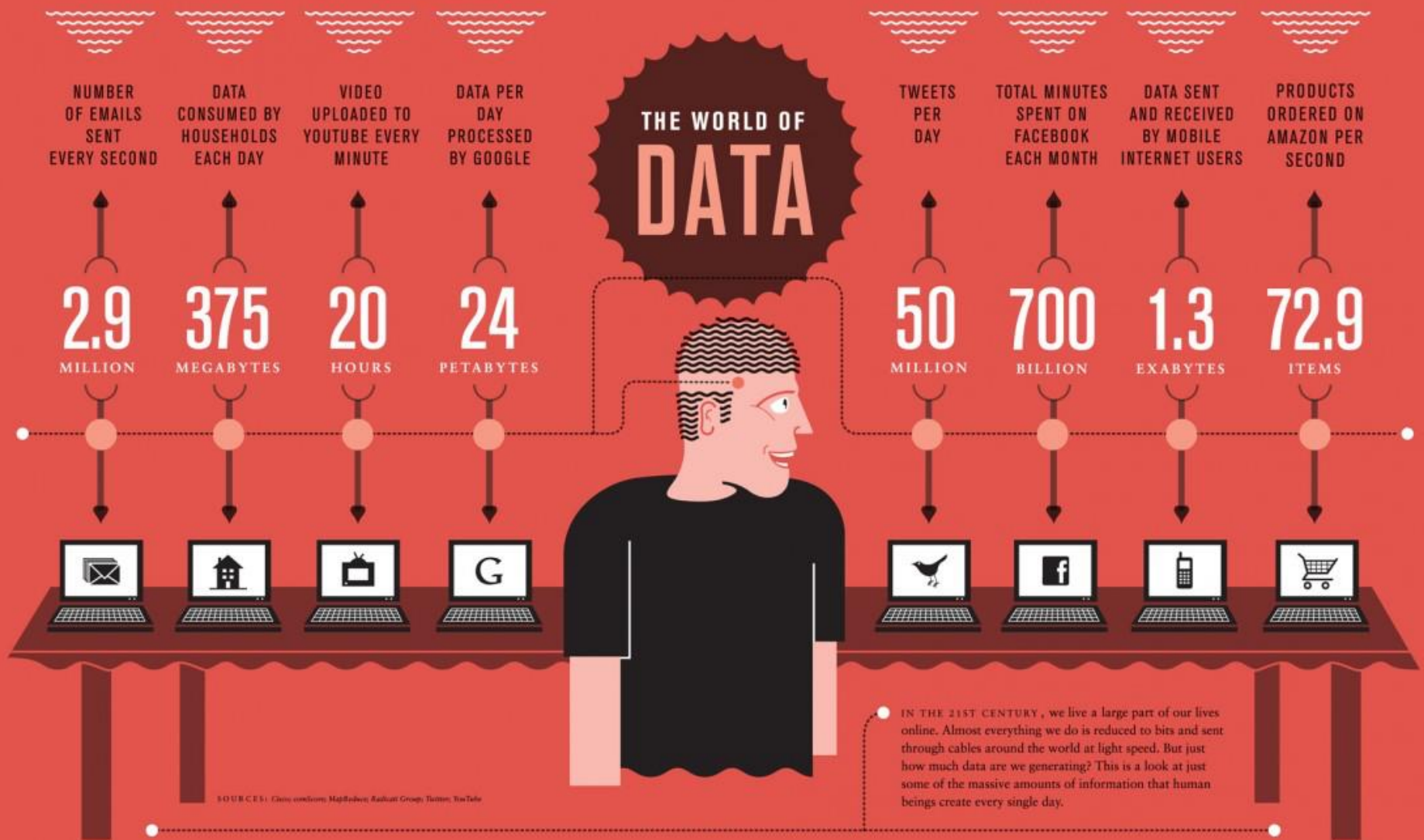
Data ethics – consideration of ethical issues when processing data.

Variety

Data extends beyond structured data, including semi-structured and unstructured data of all varieties: text, audio, video, click streams, log files and more.



What trends have you heard about to
deal with the challenges of
Volume, Velocity, Variety, Validity?
And
Which do you think is the biggest
challenge?



A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

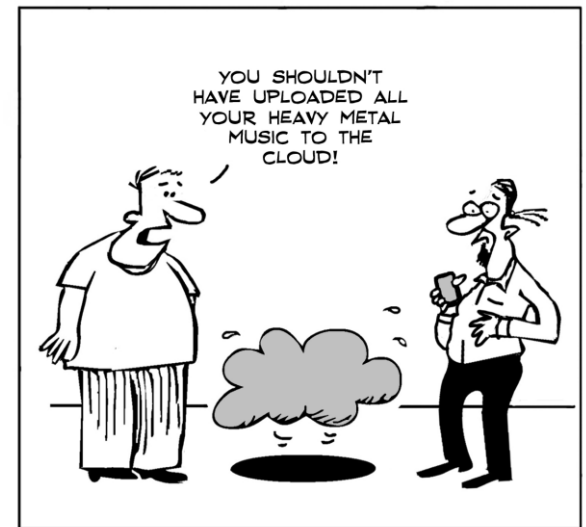
IN PARTNERSHIP WITH **IBM**

Solution Trend for coping with Volume: “The Cloud”

Desire to “out source” information management and technologies to massively distributed computing resources



By David Fletcher Of CloudTweaks



John Mc Carthy (1927-2012)

In 1958, McCarthy invented the computer programming language LISP, the second oldest programming language after FORTRAN. LISP is still used today and is the programming language of choice for artificial intelligence.

He also developed the concept of computer time-sharing in the late 1950s and early 1960s, an advance that greatly improved the efficiency of distributed computing and predated the era of cloud computing by decades.



<http://news.stanford.edu/news/2011/october/john-mccarthy-obit-102511.html>

Solution Trend for Velocity:

Solution Trend for Velocity: “Big Data”

Desire to examine and derive new insights from
information about:

- enterprise (organisation, customers, suppliers and
partners

- individuals (personalisation, recommendations etc.)

Realtime analytic techniques and technologies
increasingly key, **requires rapid data access**

BIG Data/Small data

Big Data sets that are so large that they are difficult to manage with current database management tools or traditional data processing applications.

- The data sets become large due to aggregation of data and then subsequent analysis.
- The challenges include- capture, maintaining the data, storage, search, sharing, analysis and visualisation.
- Data Analytics.....

Small data- data collected by a person. The idea of collecting everything even if there is no idea of what to use it for.

Made to measure Wearable sensors and smartphones are providing a flood of information and empowering population-wide studies, Neil Savage, Nature, Vol527,5 November 2015

Example Rapid Data Access approach: NoSQL approaches

Stands for **Non SQL**, also **Not Only SQL**

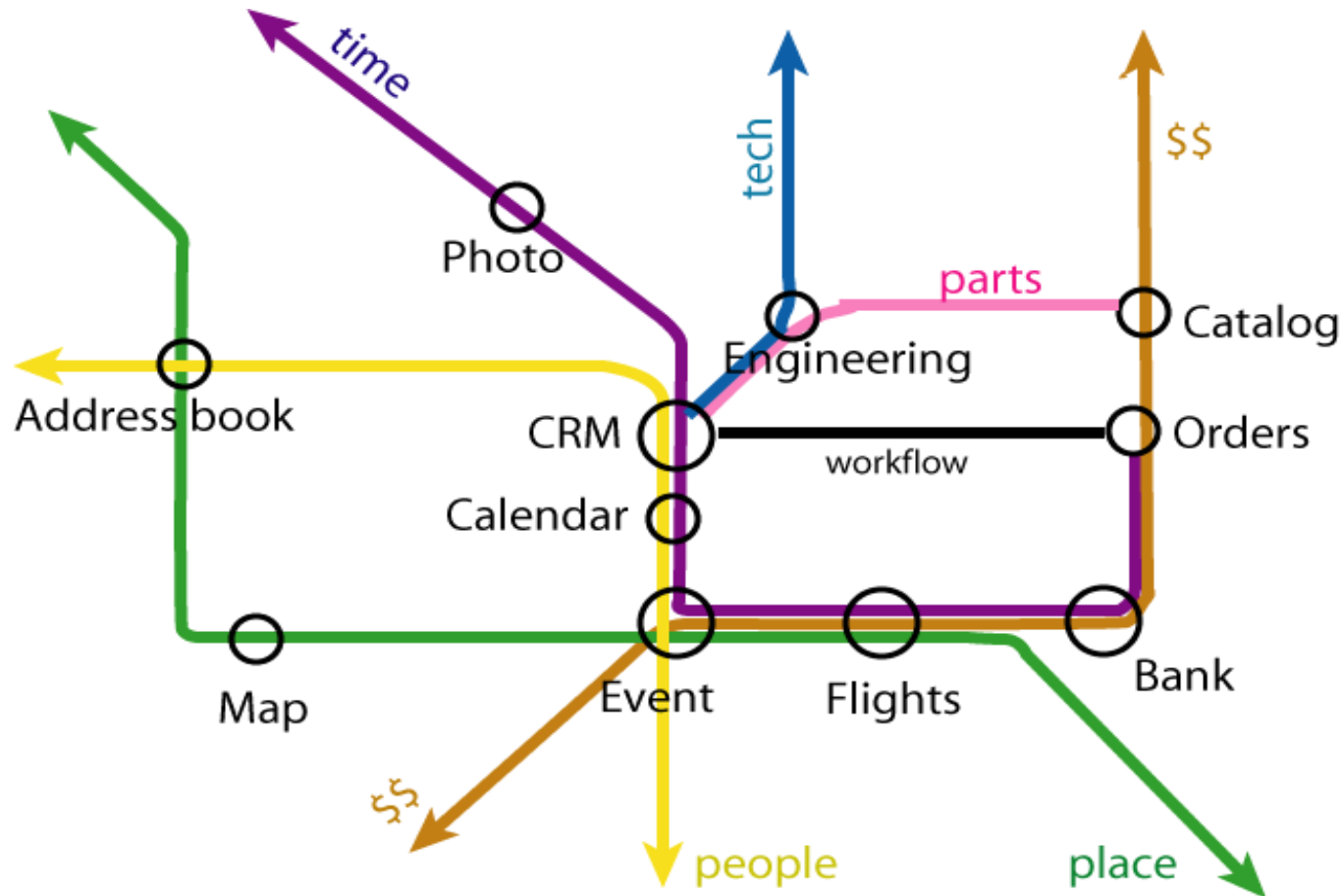
- Class of **non-relational** data storage systems
- Usually do not require a fixed table schema nor do they use the concept of joins
- **All NoSQL offerings relax one or more of the ACID properties**

Three major papers were the seeds of the NoSQL movement

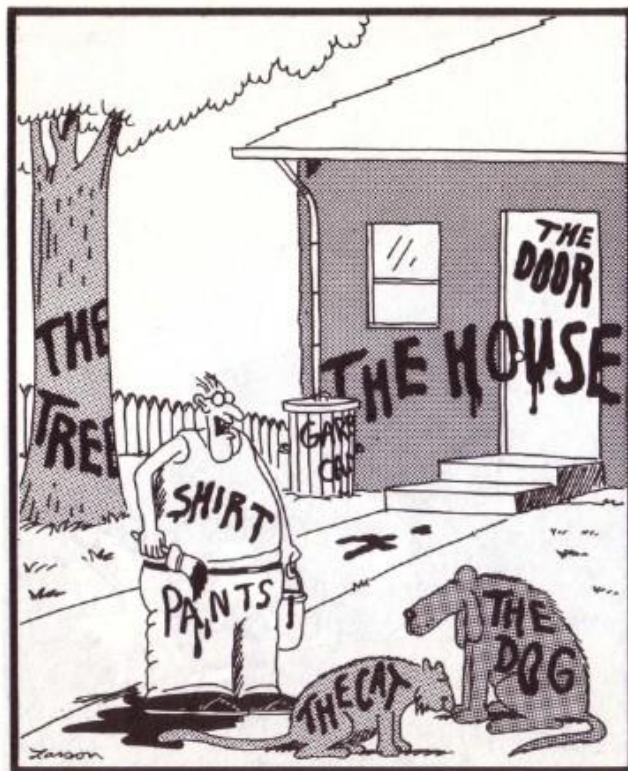
- BigTable (Google)
 - Dynamo (Amazon)
 - *Gossip protocol (discovery and error detection); Distributed key-value data store and Eventual consistency*
 - CAP Theorem
 - *Three properties of a system: consistency, availability and partitions*
 - *You can have at most two of these three properties for any shared-data system*
 - *To scale out, you have to partition. That leaves either consistency or availability to choose from*
 - *In almost all cases, you would choose availability over consistency*
-

Variety Challenge

Take advantage of data wherever it is

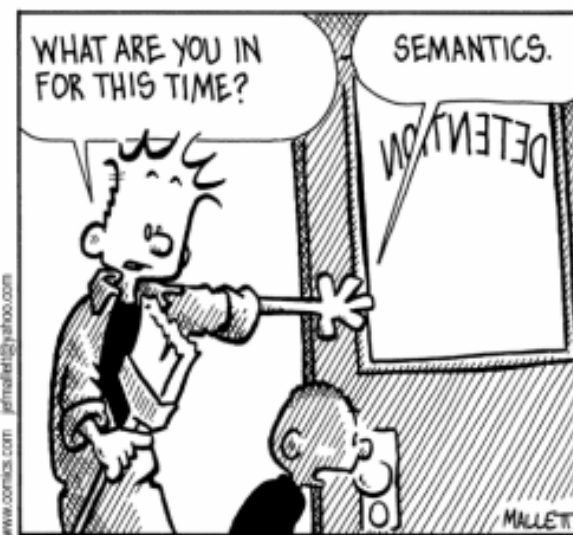
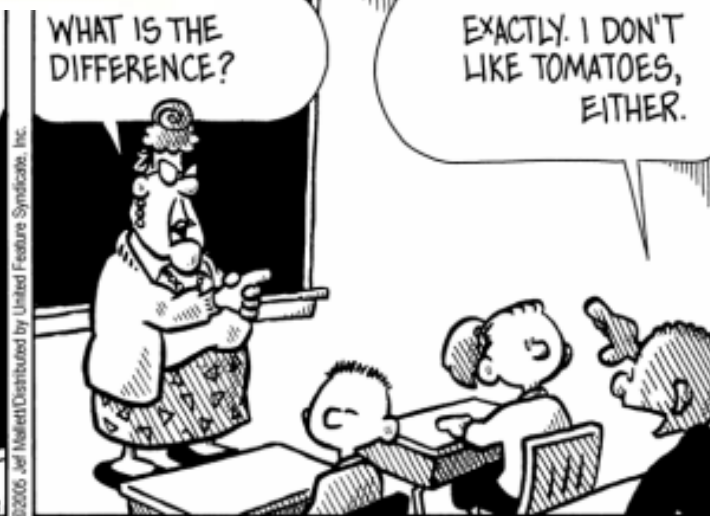


Solution Trend for coping with Variety:

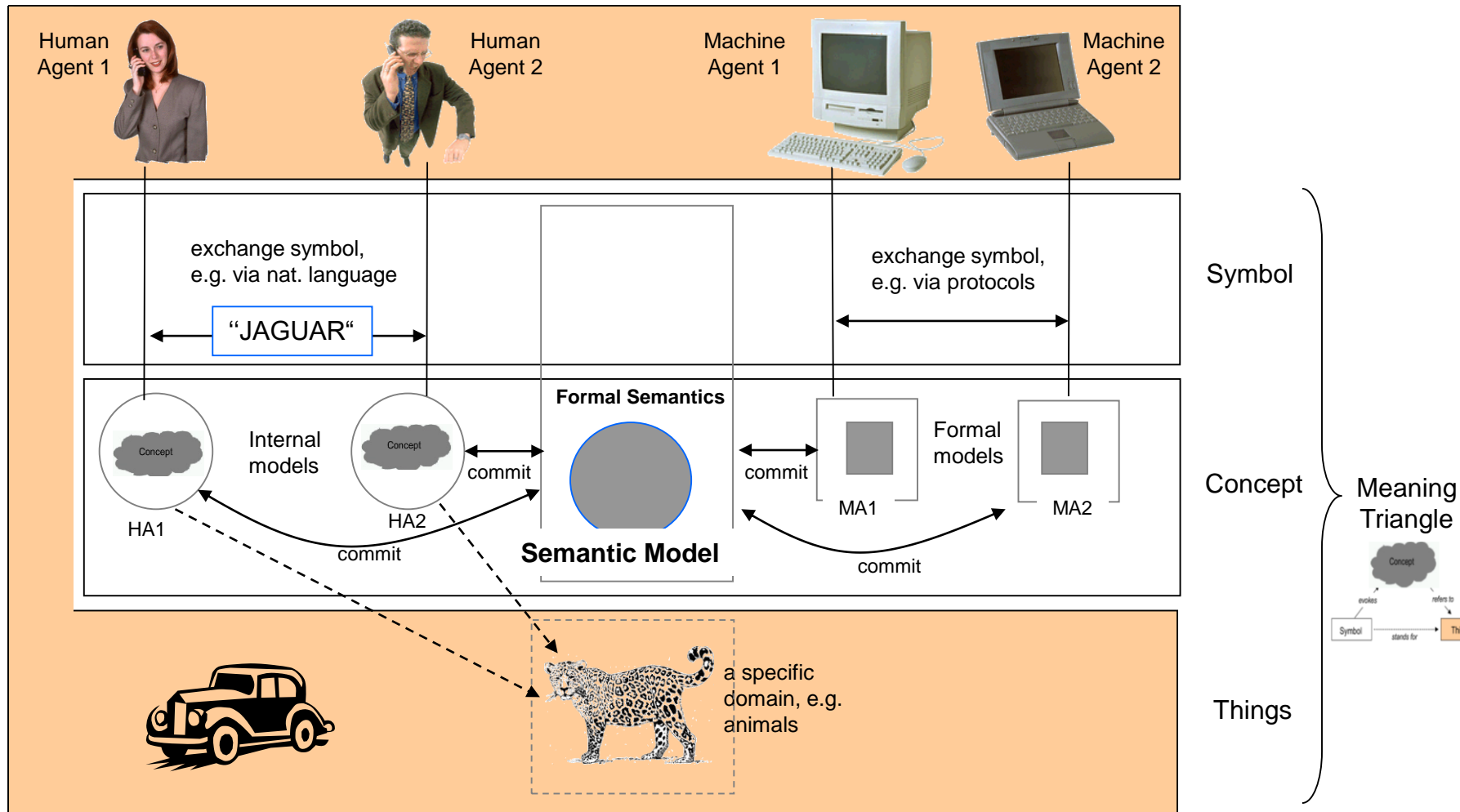


"Now! ... That should clear up
a few things around here!"

Solution Trend for coping with Variety: Natural Language Processing (NLP) and Semantic Web Technologies



Concept of Semantic Web in a nutshell



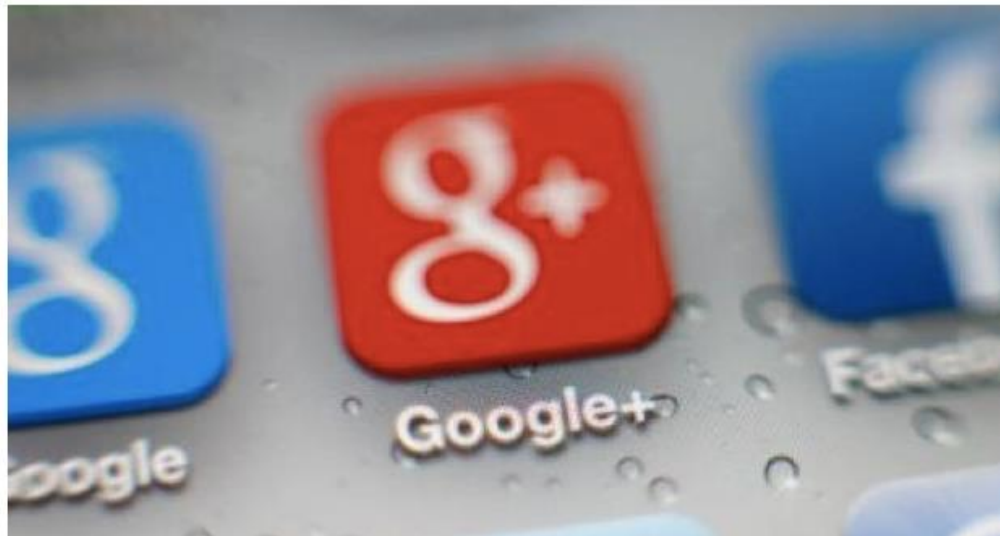
Validity: data access concern

<https://www.irishtimes.com/business/technology/data-power-could-make-1984-look-like-a-teddy-bear-s-picnic-1.3224435>

Data power could make 1984 'look like a Teddy bear's picnic'

Google and Facebook: How can we assess fairness of decisions algorithms make about us?

Marie Boran • about 11 hours ago



We don't search online, we "google", so it is no surprise that Google has over 77 per

Solution Trend for Validity: Data Protection, Data Privacy

Protection

In Europe GDPR (General Data Protection Regulations)

– challenges

explicit gathering and lifecycle management consent-
(check out Risk based approach, Notice and Choice based approach)

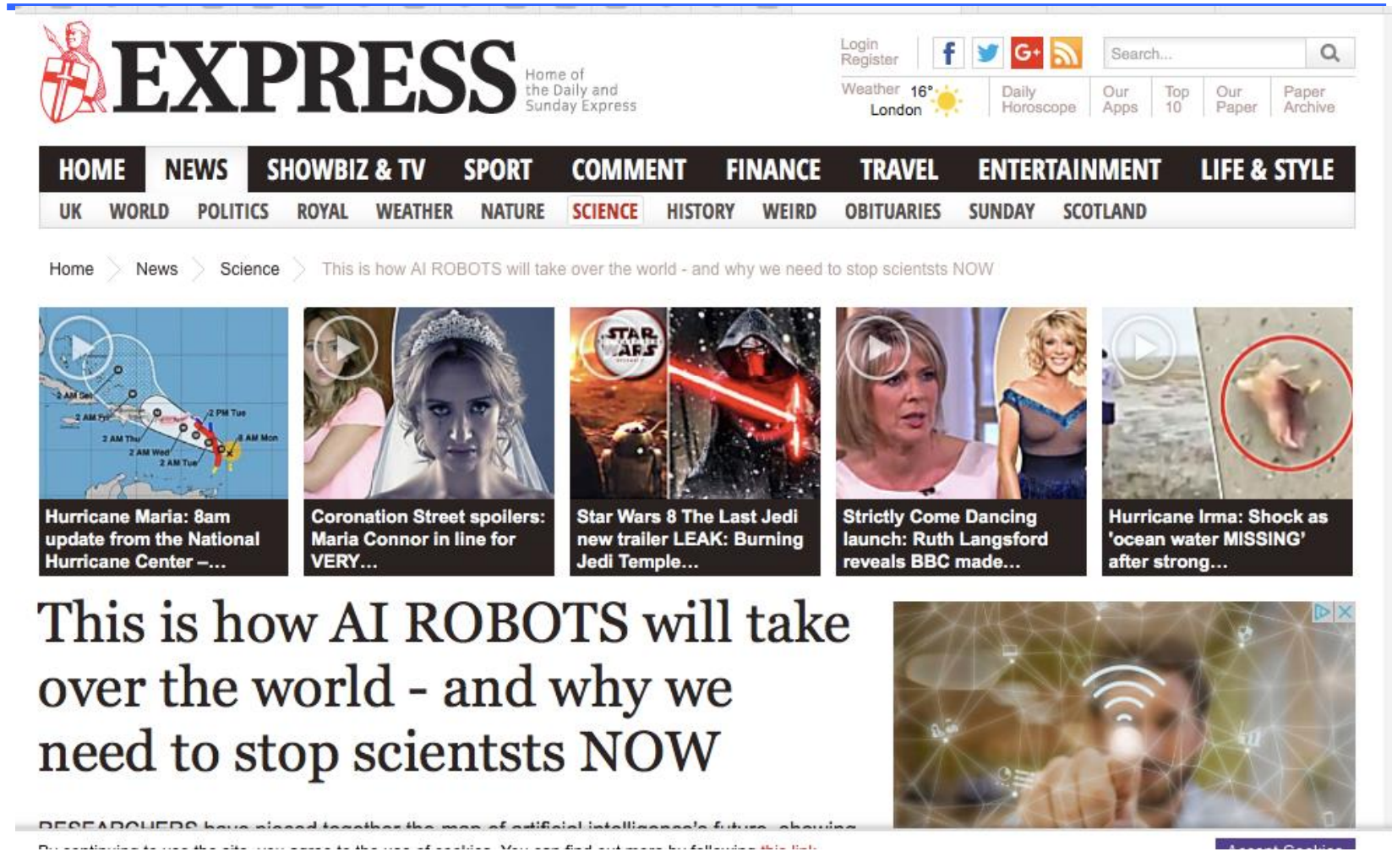
- Automatic compliance checking

Privacy

Raising awareness and providing tools for users to understand the “convenience vs privacy” tradeoff

- Check out your Digital Footprint <http://www.bigfoot.ie>

Validity: data processing concern



The screenshot shows the Daily Express website. The masthead features the 'EXPRESS' logo with the tagline 'Home of the Daily and Sunday Express'. Navigation links include 'HOME', 'NEWS', 'SHOWBIZ & TV', 'SPORT', 'COMMENT', 'FINANCE', 'TRAVEL', 'ENTERTAINMENT', and 'LIFE & STYLE'. A secondary row of links includes 'UK', 'WORLD', 'POLITICS', 'ROYAL', 'WEATHER', 'NATURE', 'SCIENCE', 'HISTORY', 'WEIRD', 'OBITUARIES', 'SUNDAY', and 'SCOTLAND'. The 'SCIENCE' link is highlighted. Below the navigation, a breadcrumb trail reads 'Home > News > Science > This is how AI ROBOTS will take over the world - and why we need to stop scientists NOW'. A row of five featured stories is displayed, each with a video player icon and a title: 'Hurricane Maria: 8am update from the National Hurricane Center -...', 'Coronation Street spoilers: Maria Connor in line for VERY...', 'Star Wars 8 The Last Jedi new trailer LEAK: Burning Jedi Temple...', 'Strictly Come Dancing launch: Ruth Langsford reveals BBC made...', and 'Hurricane Irma: Shock as 'ocean water MISSING' after strong...'. The main article headline is 'This is how AI ROBOTS will take over the world - and why we need to stop scientists NOW'. Below the headline, a video player shows a man pointing at a screen with a neural network overlay. The video player has a play button and a close button in the top right corner.

EXPRESS Home of the Daily and Sunday Express

Login Register | f | | G+ | | Search... |

Weather 16° London | Daily Horoscope | Our Apps | Top 10 | Our Paper | Paper Archive

HOME NEWS SHOWBIZ & TV SPORT COMMENT FINANCE TRAVEL ENTERTAINMENT LIFE & STYLE

UK WORLD POLITICS ROYAL WEATHER NATURE SCIENCE HISTORY WEIRD OBITUARIES SUNDAY SCOTLAND

Home > News > Science > This is how AI ROBOTS will take over the world - and why we need to stop scientists NOW

Hurricane Maria: 8am update from the National Hurricane Center -...

Coronation Street spoilers: Maria Connor in line for VERY...

Star Wars 8 The Last Jedi new trailer LEAK: Burning Jedi Temple...

Strictly Come Dancing launch: Ruth Langsford reveals BBC made...

Hurricane Irma: Shock as 'ocean water MISSING' after strong...

This is how AI ROBOTS will take over the world - and why we need to stop scientists NOW

RESEARCHERS have pieced together the map of artificial intelligence's future, showing...

Solution Trend for Validity: Data Ethics

Ethics

- Conversation just beginning on the ethics of processing data
- Being taken seriously at corporate level (e.g. IBM)
- Efforts ongoing to provide stakeholders to address ethics early in development lifecycle
 - Check out <http://ethicscanvas.org>

Homework Task #1

Read the article at:

<https://www.irishtimes.com/business/technology/data-power-could-make-1984-look-like-a-teddy-bear-s-picnic-1.3224435>

Prepare the following and Upload a pdf file(via Blackboard Task) by Wednesday 12th September for group work and discussion in class.

- a) 2 lines stating why you like or dislike about arguments made in the article
- b) 2 lines critique of the article
- c) 2 lines on whether you believe more attention needs to be placed on data ethics or not

So far..and Next

So Far

Context for the Information Modelling module

- Data, Information, Knowledge
- Structured vs Unstructured approaches to information
- Current challenges for data management :
 - Volume, Velocity, Variety, Validity
- **Homework Task to Complete- see slide above.**

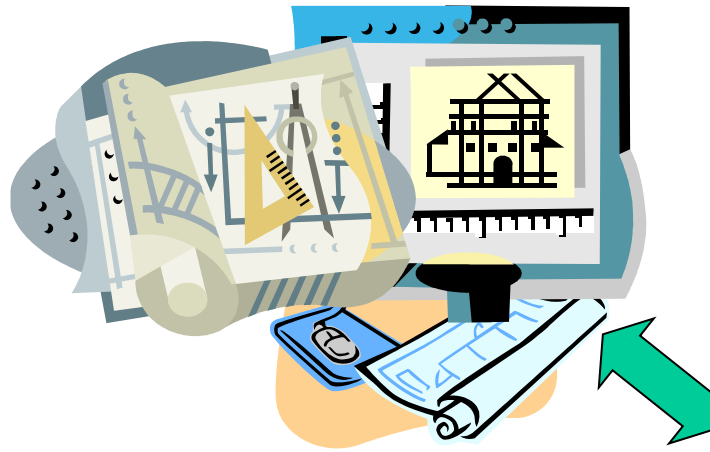
Next

- Modelling

Creating Buildings



Architect

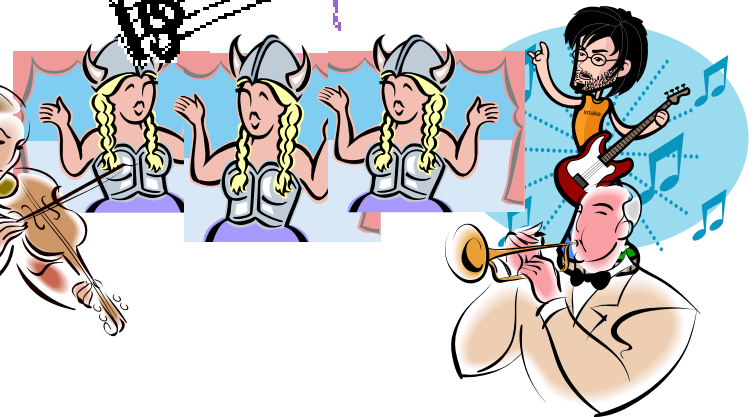
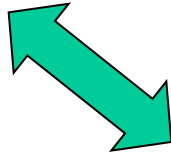


Engineers, builders
customer

Creating Musical Performances

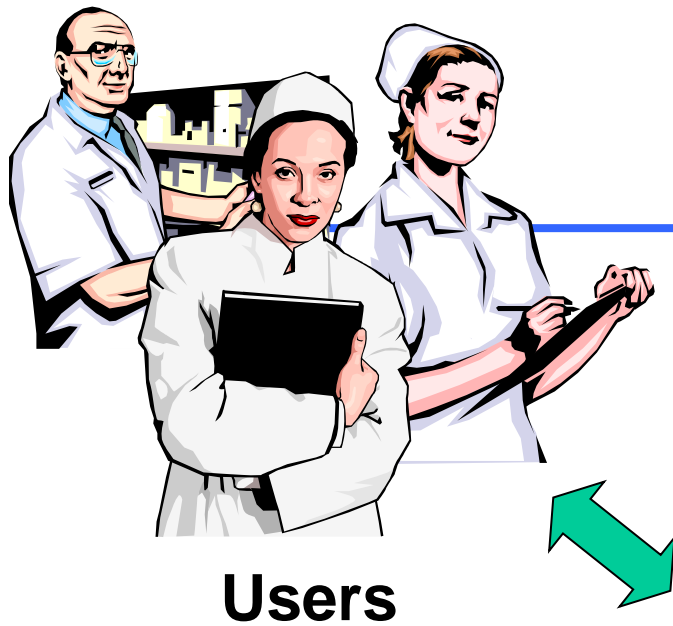


Composer



Conductor, Orchestra

Creating Information Systems



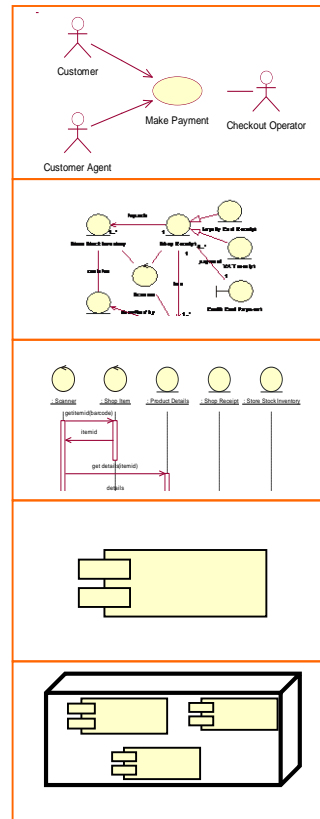
Designers & Developers

Creating Information Systems

UML Diagrams



Users



Designers & Developers

Modelling

Models are **abstractions** of the real world

Models are always less detailed than the real world

Many models can be created of any given physical object depending upon the level of detail and point of view selected- Think of maps.

There is no such thing as the most “correct” model; models are simply better or worse suited to accomplishing a particular task

Models and Systems

Modeling is required to **appropriately** automate these Systems
systems have behaviour which distinguishes them from other systems and
from its environment

- behaviour cannot generally be predicted by an examination of individual components – care must be taken to consider all components. The behaviour of the systems is not the same as the sum of the behaviour of the system as a whole.
- <https://hackaday.com/2015/10/26/killed-by-a-machine-the-therac-25/>
- systems are embedded in an environment
- e.g “superbugs”

systems have internal structure

- amount of internal detail is limitless e.g. description of cardiovascular system could descend to the cellular level
-

Learning Outcomes: Information Management I

- Describe and use Unified Modelling Language(UML) technologies for information modelling
- Describe and use XML technologies for data modeling and querying
- Describe the techniques used for exposing and retrieving information on the web using semantic web/linked data approaches

Marking

Coursework and Examination - Details will be given in class on Thursday 13th September.

Coursework will consist of a Group based project with two parts- Details of this project and groups will be given in class on Thursday 13th September.

Coursework Grading

- Marks for group project is based on **presentations; demos; reports and code**
 - *15% will be deducted from your percentage for every day late for a deliverable (e.g. presentation, code, report)*

Timetable

- (Monday(10-12, LB01) and Thursday(11-12, LB01)
- Sessions a mixture of lecture, discussion, exercises, tutorials and student presentations
- Periodic sessions replaced by Labs and Demos of project

Sign in sheets used

Notes distributed via Blackboard