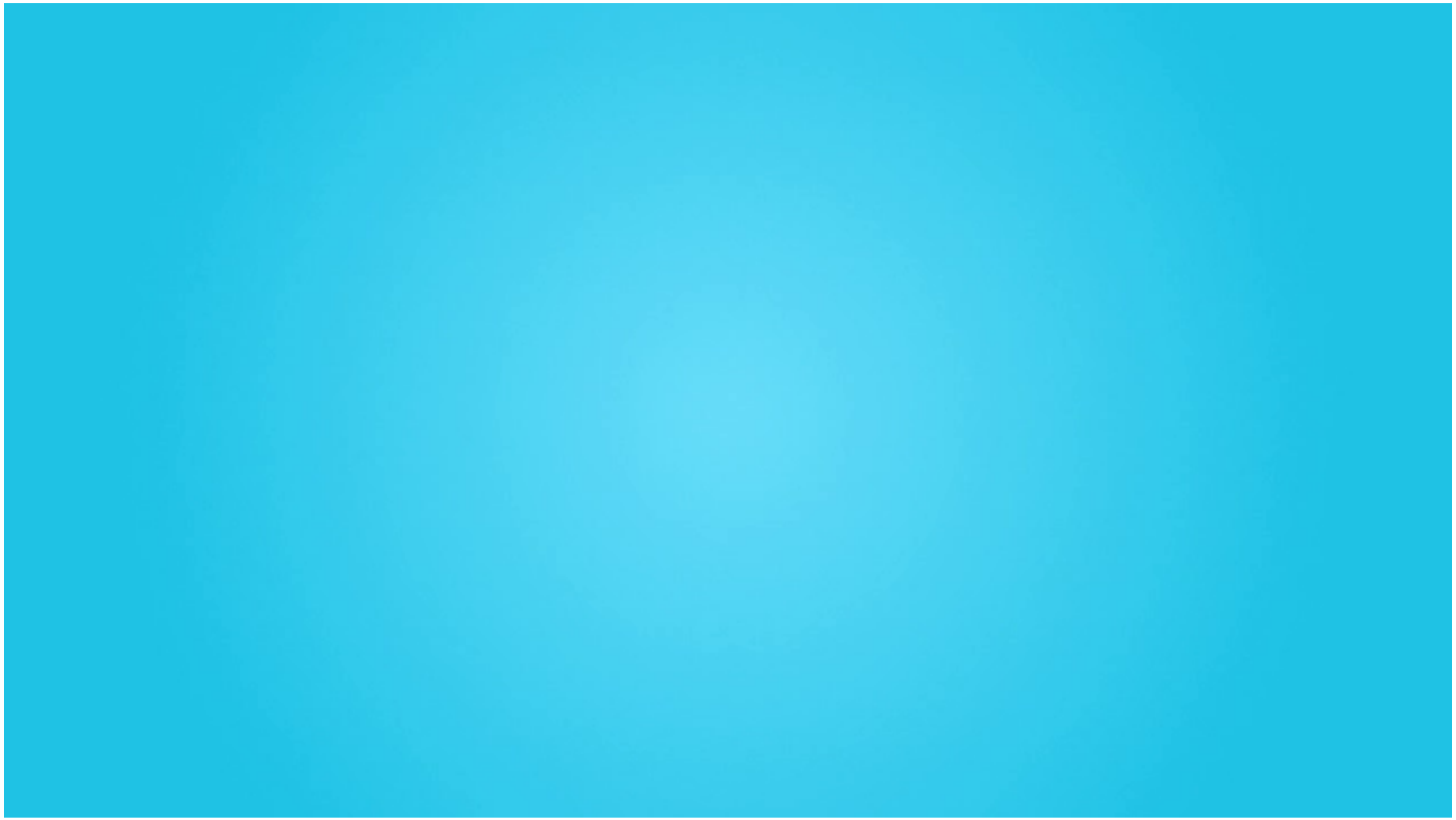
University of Dublin
Trinity College

CS2041: Information Management I

An Introduction to the module
2017-18

What is information?



What is the difference between Data, Information and Knowledge?

Data	Information	Knowledge
Raw	Data given meanng	Info relationship
Meaningless	Interpretation	Rules applied to info
No interpretation	Basic unit of communication	Apllication of info
Finite representation		

What is the difference between Data, Information and Knowledge?

Data:

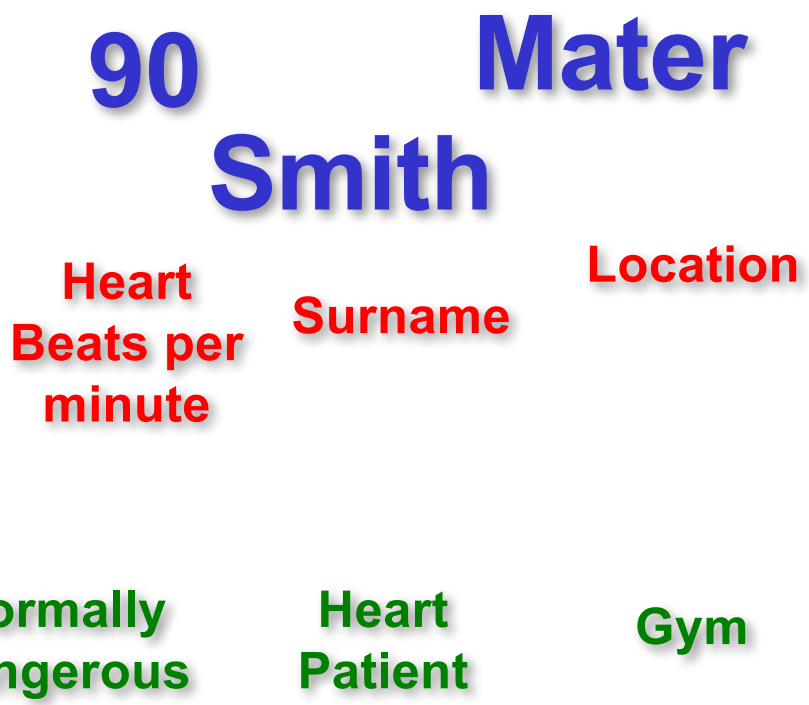
- Raw facts; building blocks of information
- Unprocessed information

Information:

- Data associated together to convey some meaning

Knowledge:

- Interrelating and "understanding"information



Core Concepts

ORGANISATION
How data represented/associated

METADATA
Data about what is the data

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 you know manages data?

So what software you know 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!

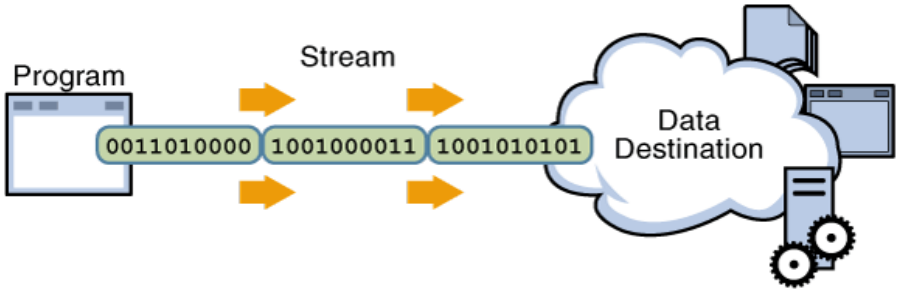
Maintaining structure in your own data file

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

SQL (or Xquery)

SELECT Name

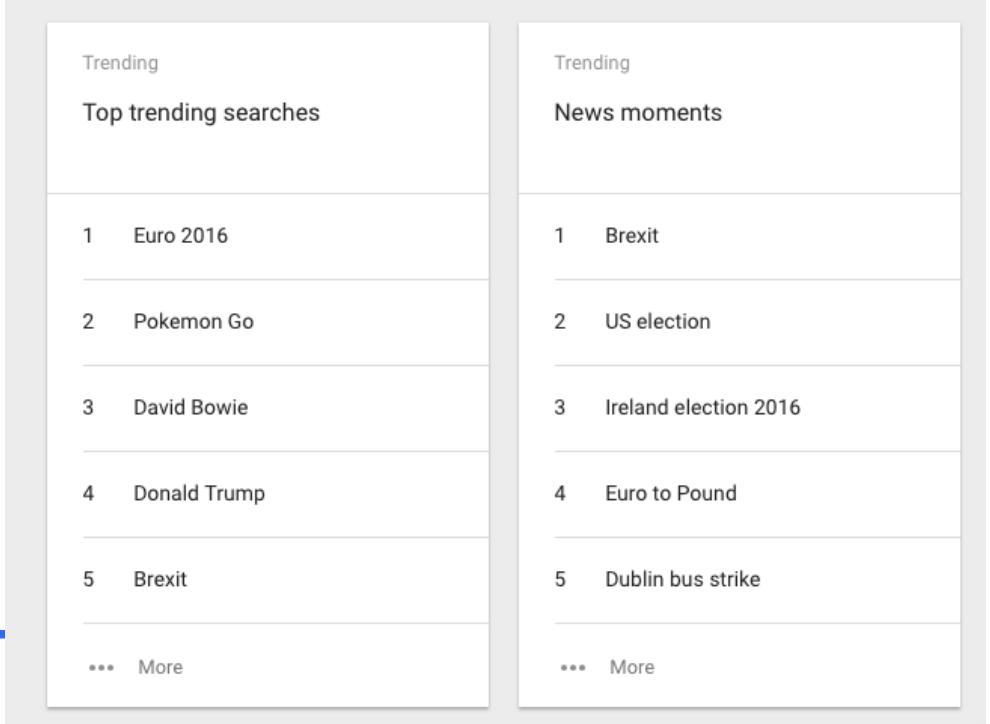
FROM Character

WHERE Salary

BETWEEN 40000 AND 60000

Keyword based, increasingly phrase based

Google Zeitgeist 2016 for Ireland



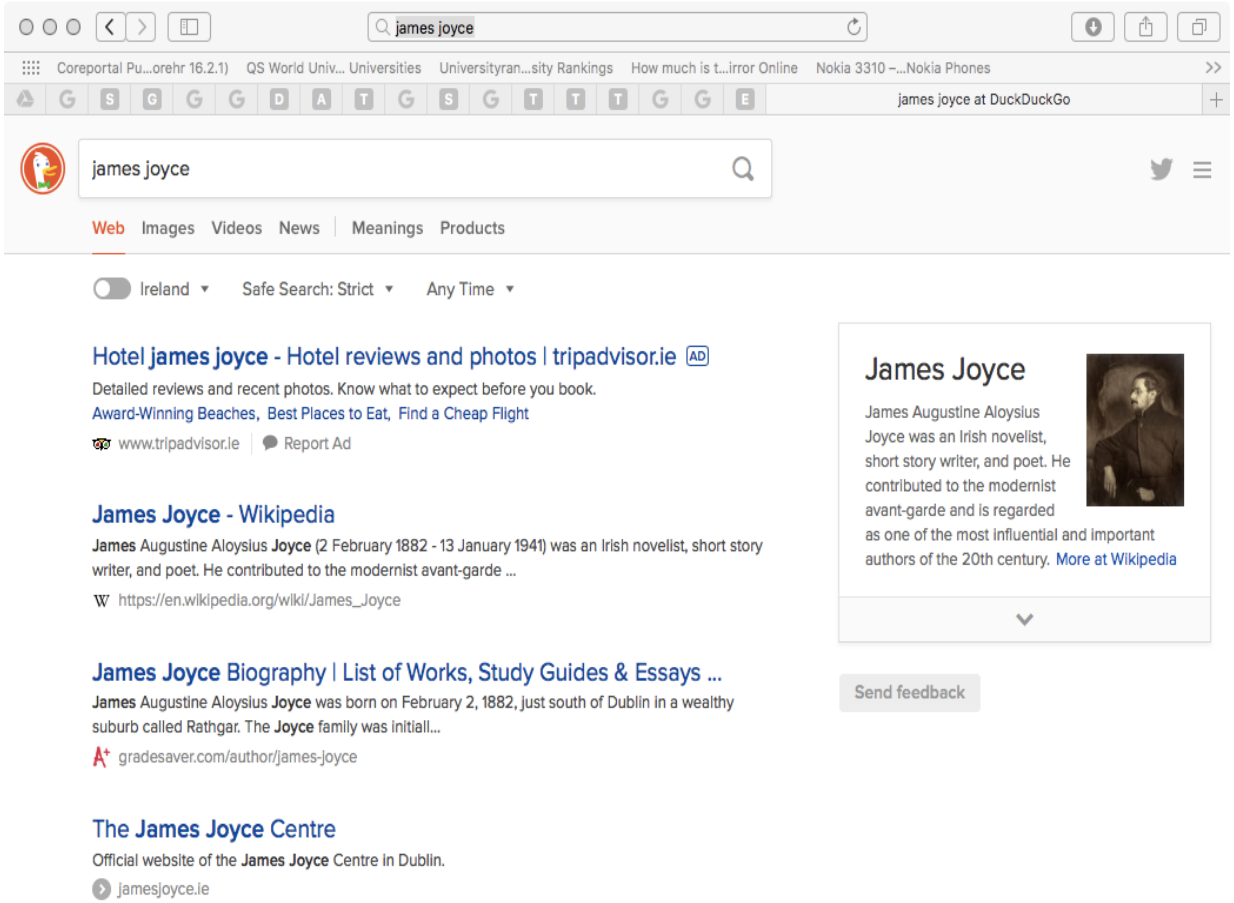
Nature of Results: Structured vs Unstructured

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

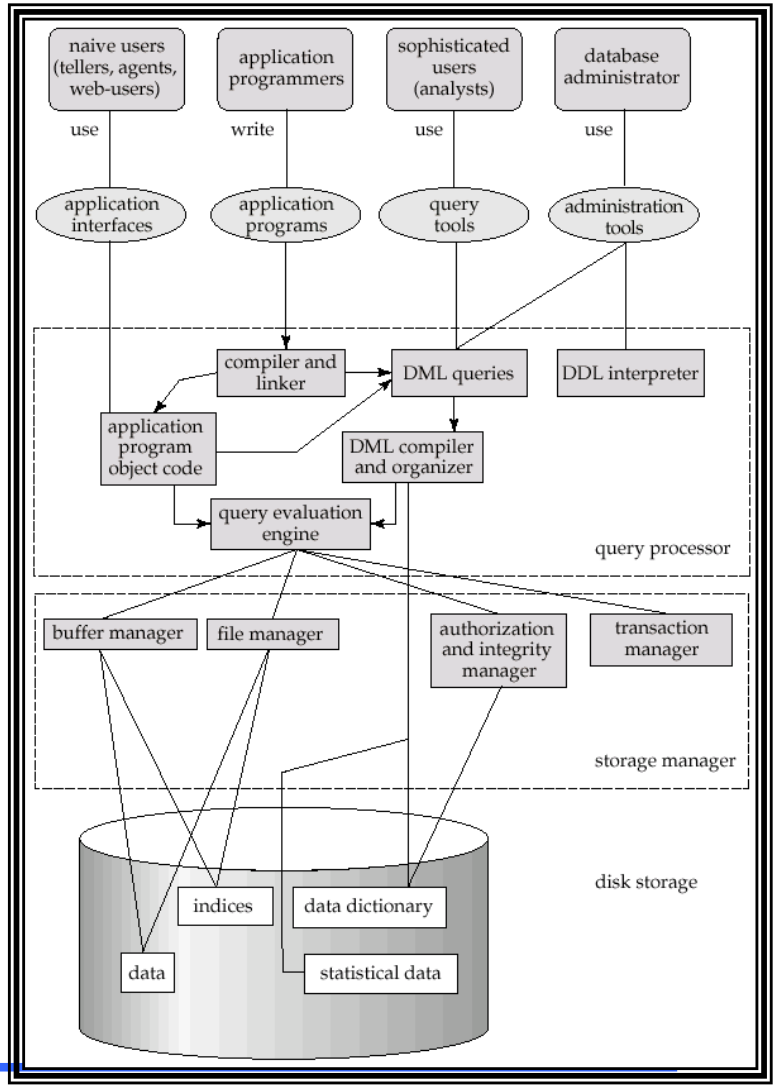
- **A**tomic, **C**onsistent, **I**solated, **D**urable

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)

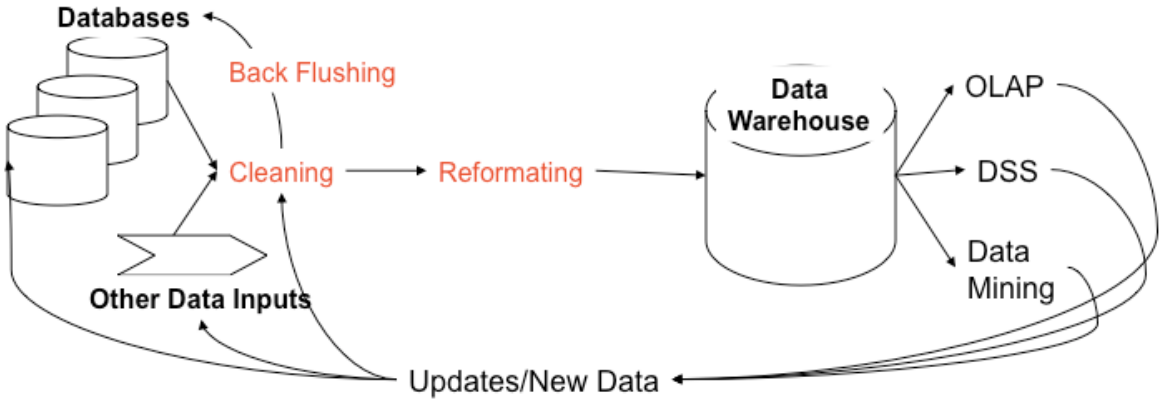
Data Warehouse is a subject oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions

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

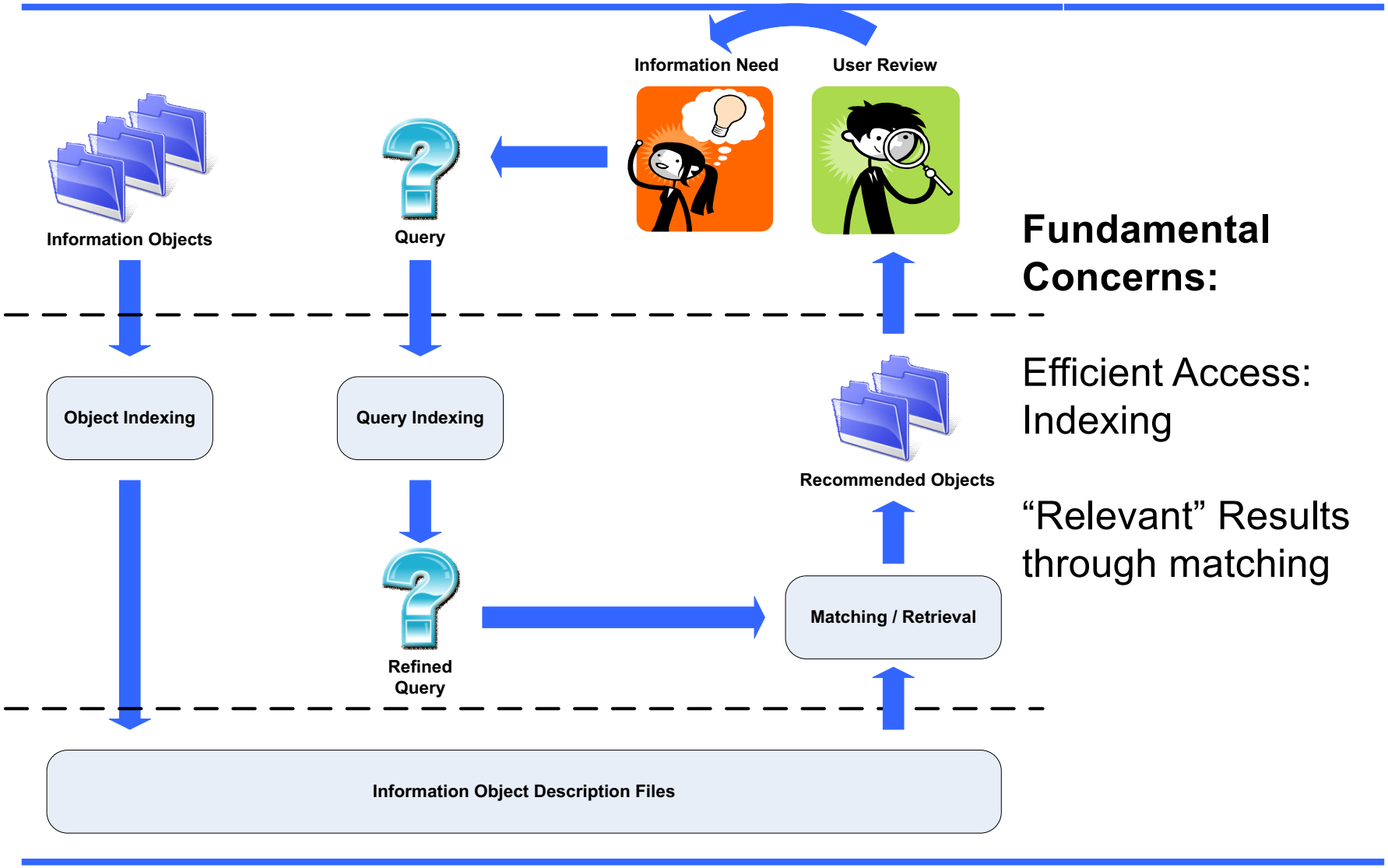
Used for

- Data Mining
- Decision Support
- OLAP



What does a Structured approach to data management involve?

Unstructured Approach: Information Retrieval

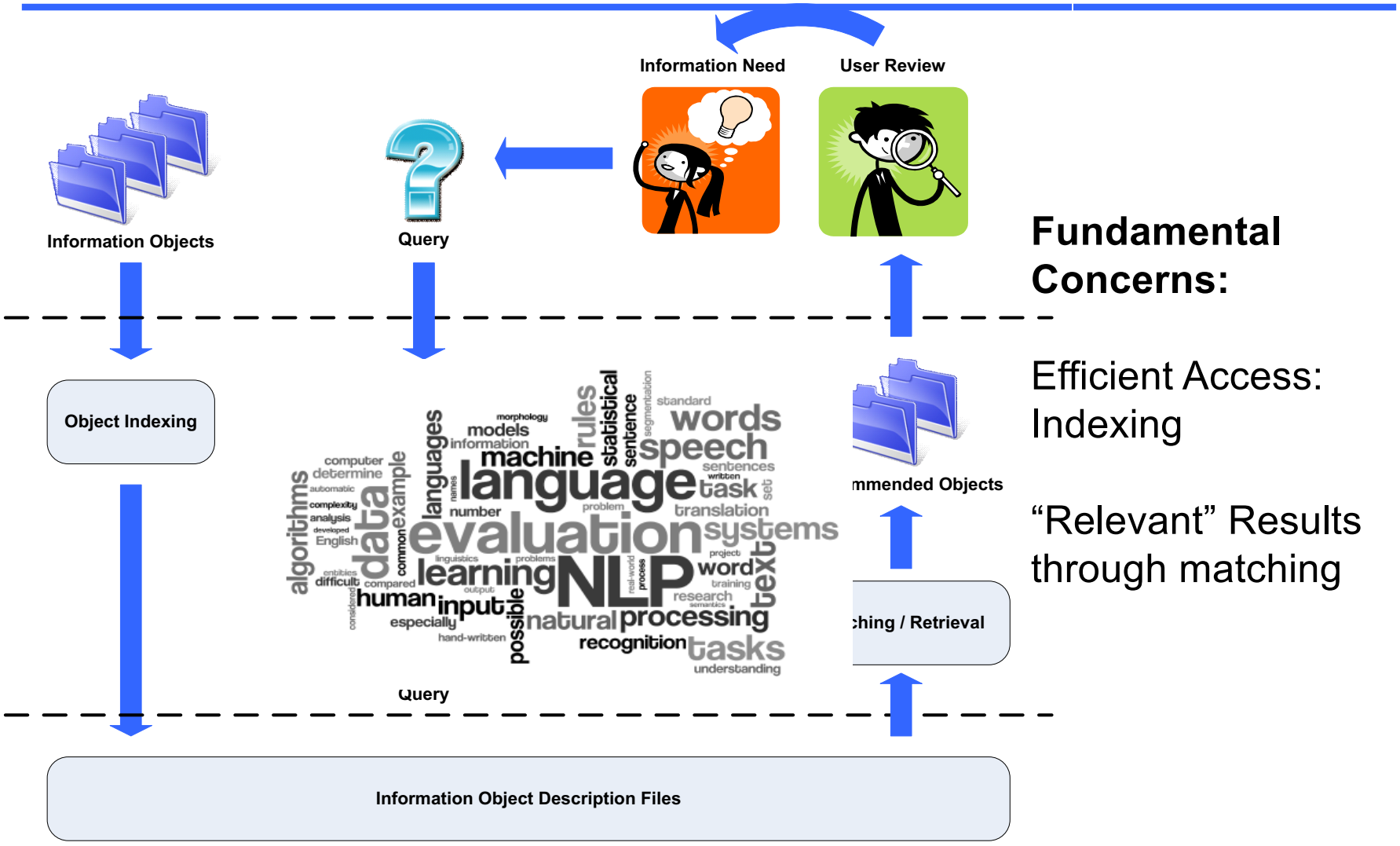


Fundamental Concerns:

Efficient Access: Indexing

"Relevant" Results through matching

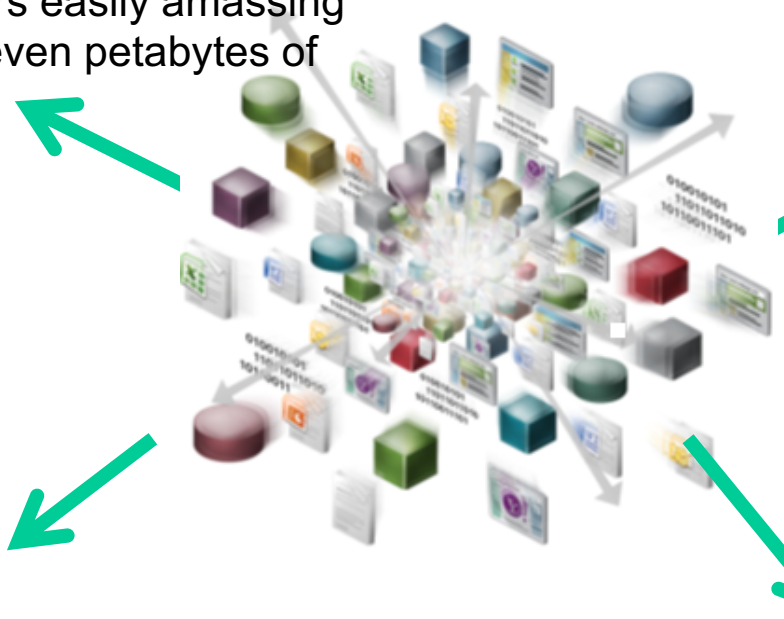
Unstructured Approach: Information Retrieval



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

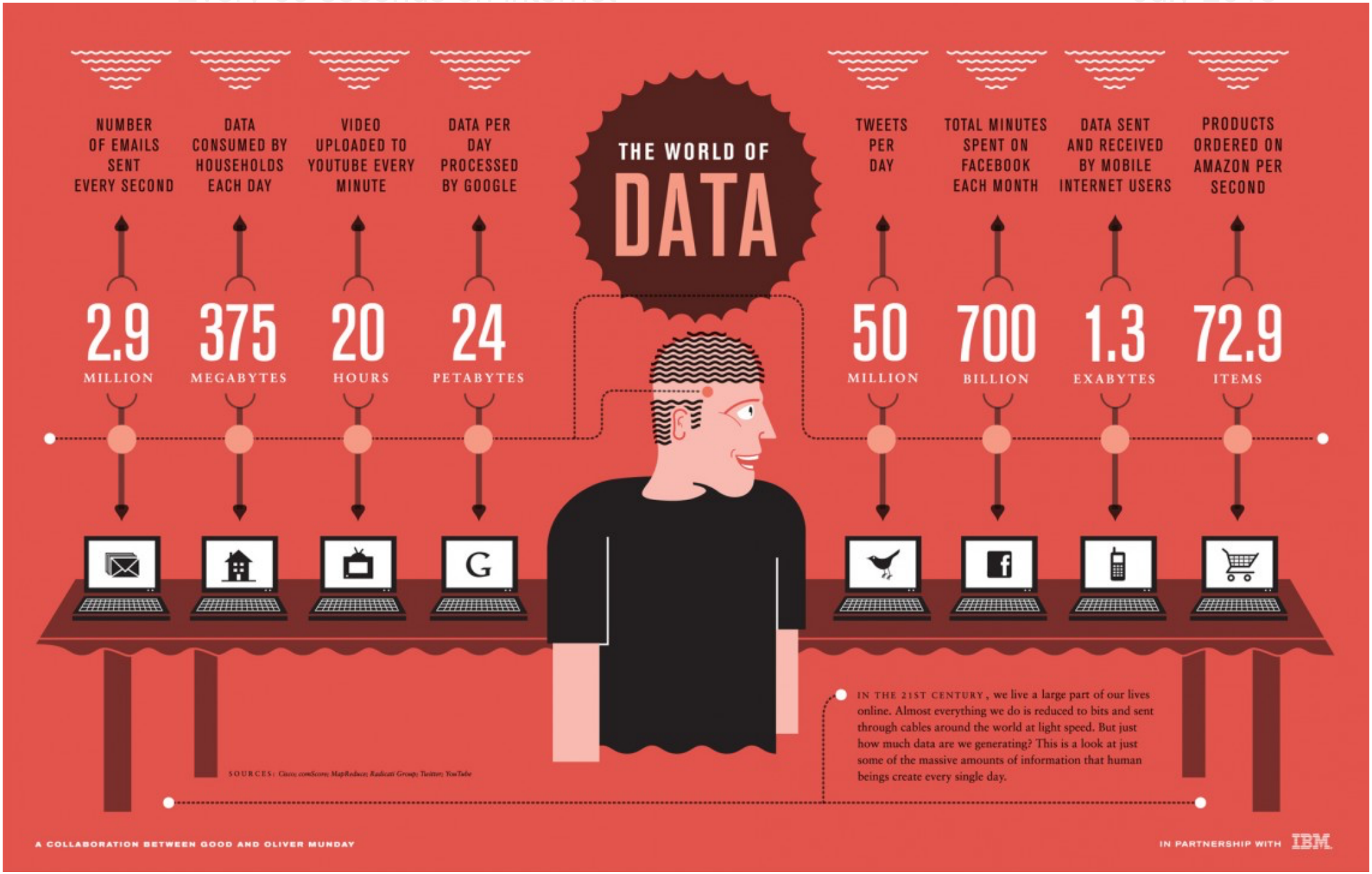
Data protection – consent and compliance;
Data privacy – what data an individual willing to share;
Data ethics – consideration of ethical issues when processing data.

Variety

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

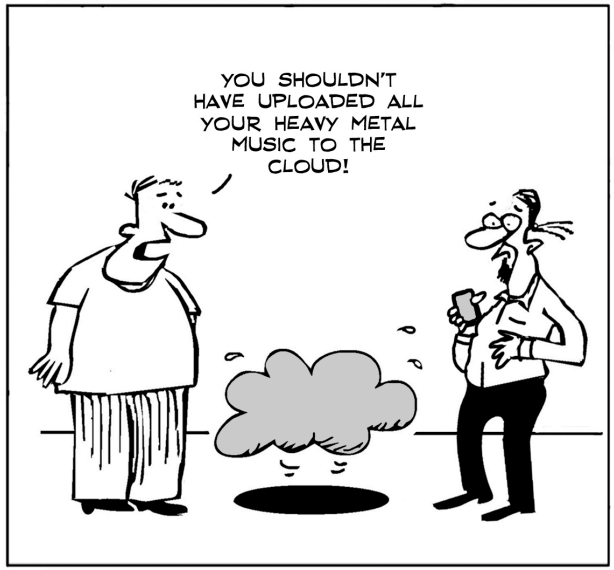
What trends have you heard about to deal with the challenges of Volume, Velocity, Variety, Validity?

Solution Trend for coping with Volume:


THE WORLD OF
DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9
MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375
MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20
HOURS
DATA PER DAY PROCESSED BY GOOGLE
24
PETABYTES
TWEETS PER DAY
50
MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700
BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3
EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9
ITEMS
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.
SOURCES: Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube
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



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 (personsalisation, recommendations etc.)

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

Example Rapid Data Access approach: NoSQL approaches

Stands for **N**on **SQL,** also **N**ot **O**nly **SQL**

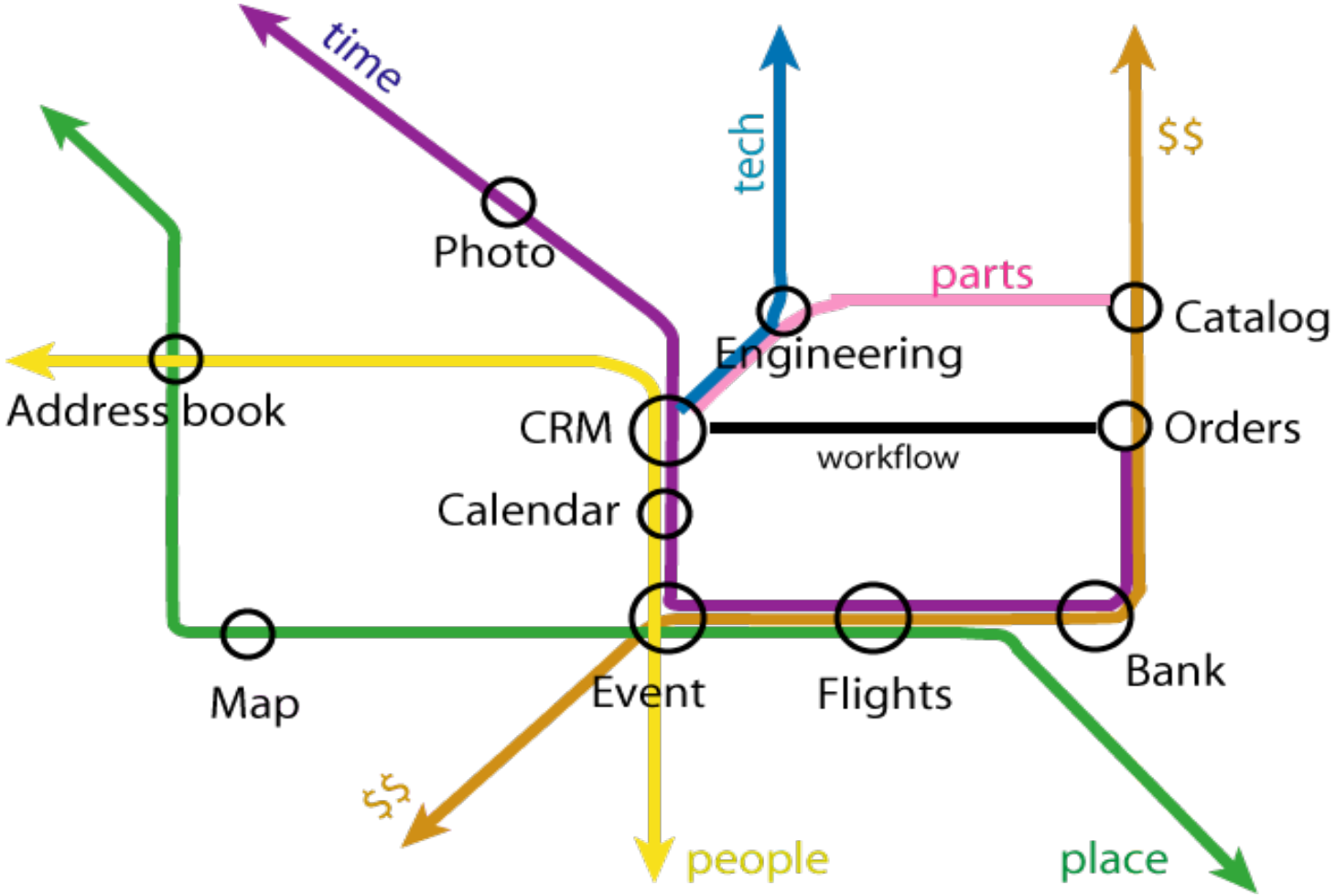
- Class of **non-relational** data storage systems
- Usually do not require a fixed table schema nor do they use the concept of joins
- **All NoSQL offerings relax one or more of the ACID properties**

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

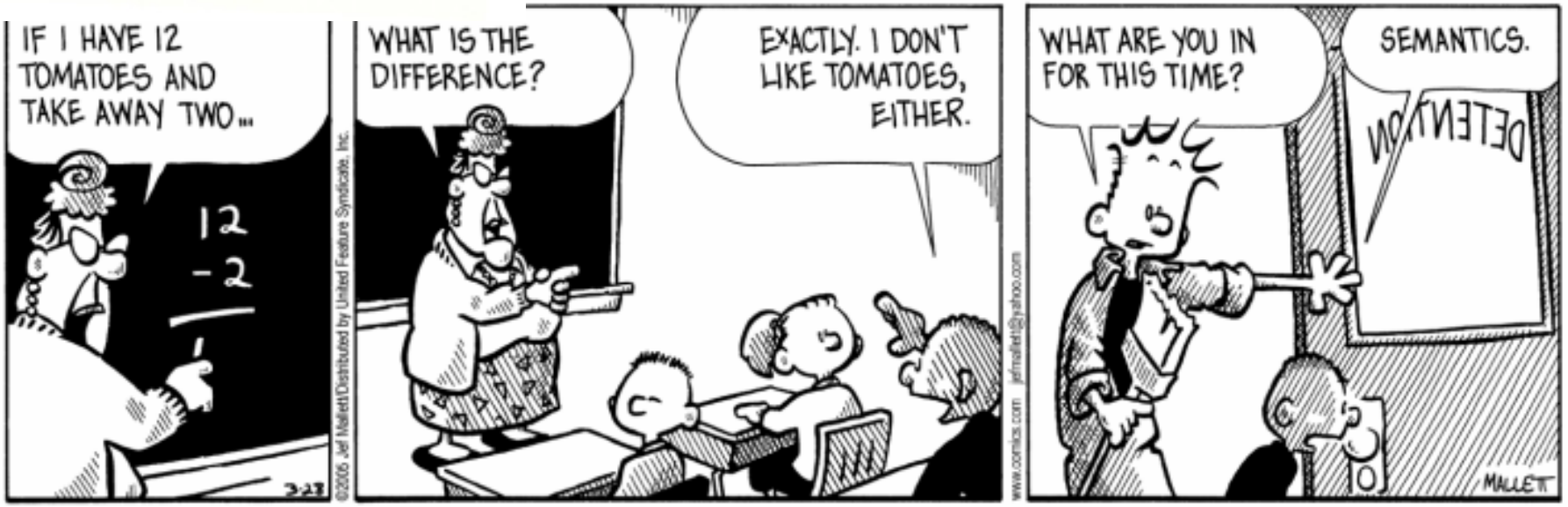
Take advantage of data wherever it is



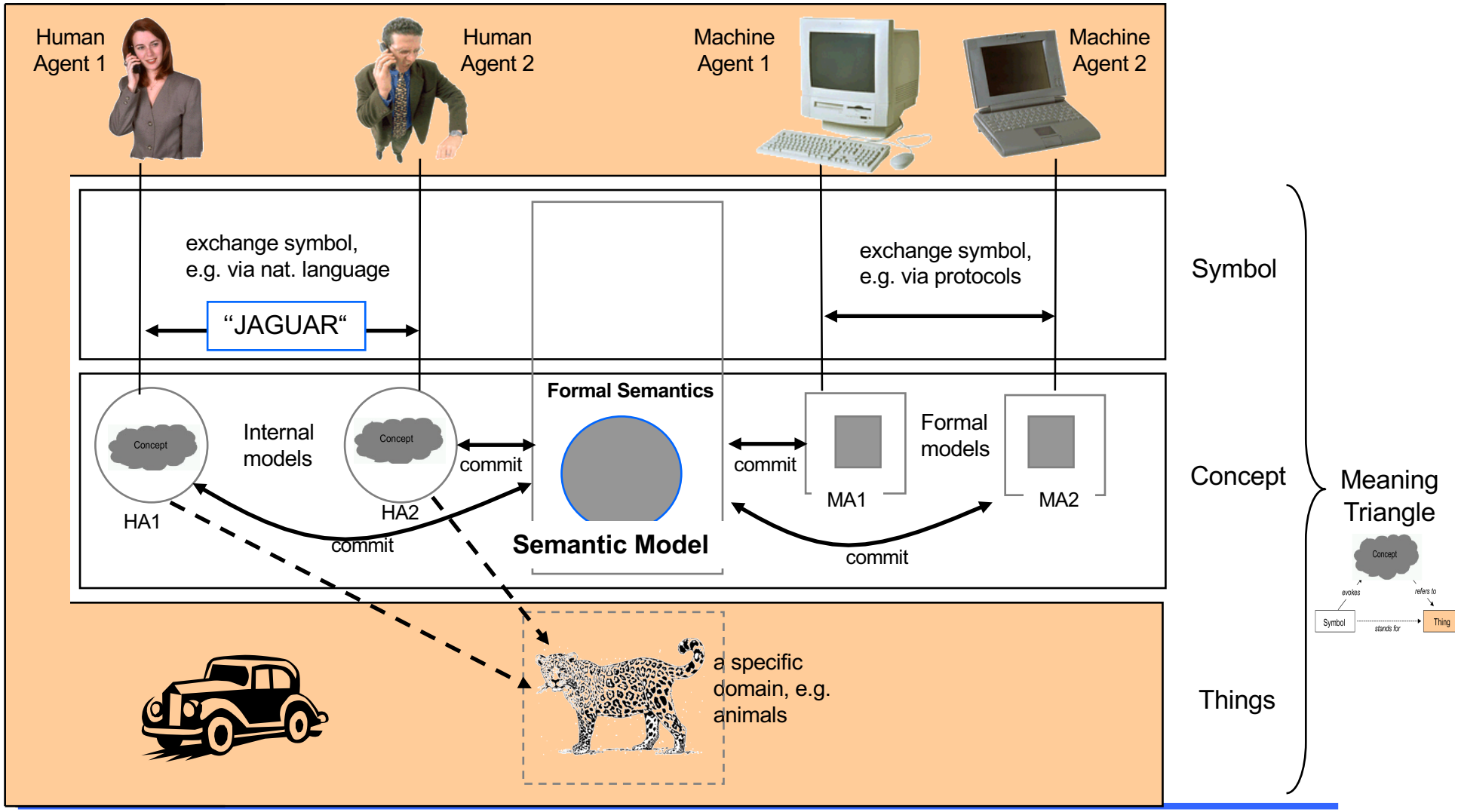
Solution Trend for coping with Variety:

"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



[Maedche et al., 2002]

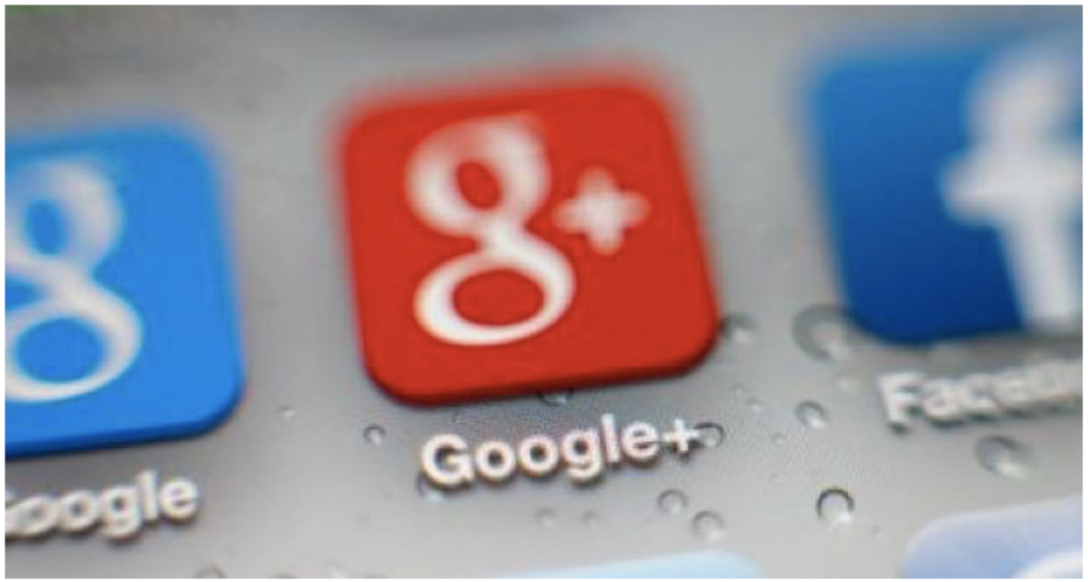
Validity: data access concern

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

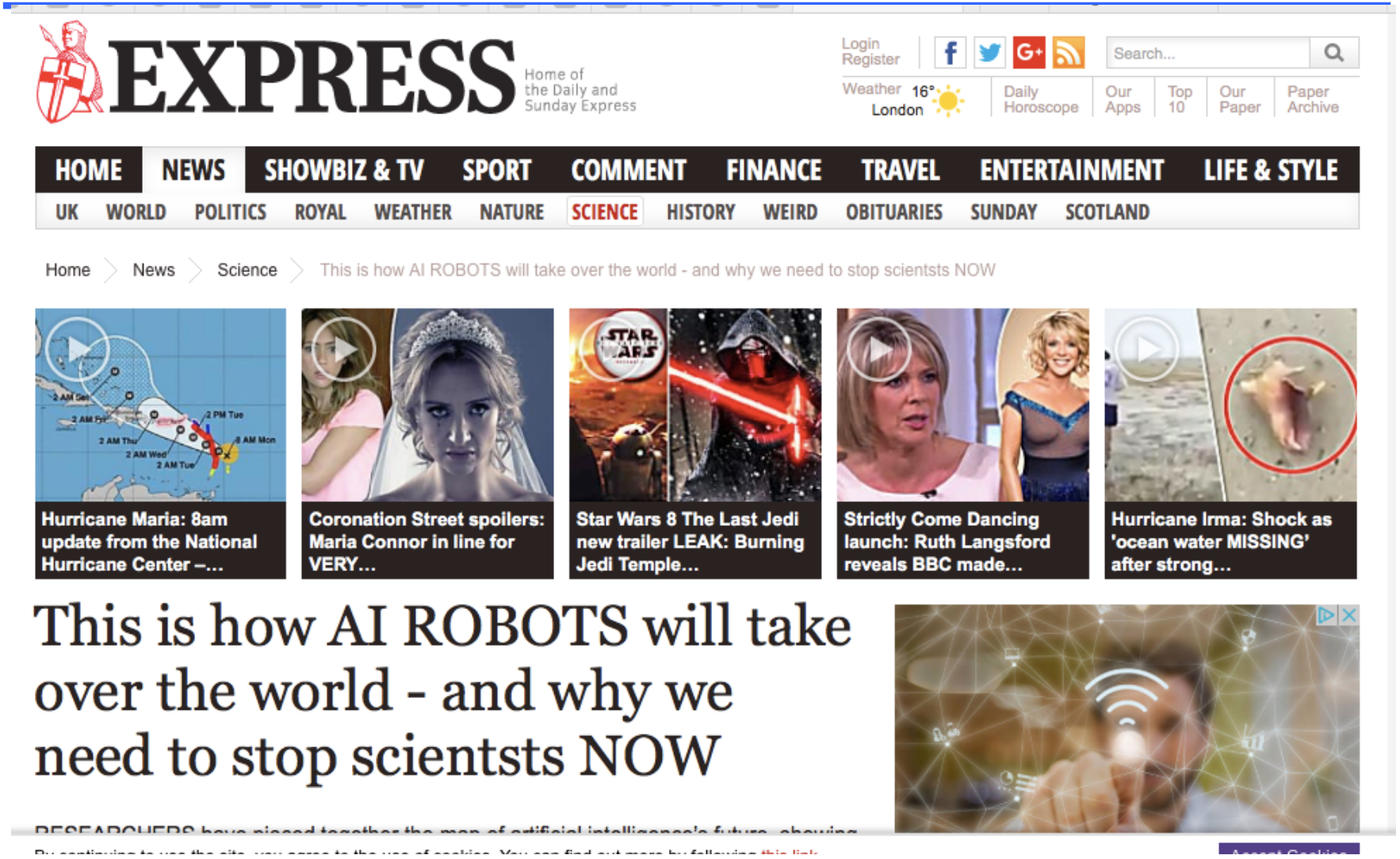
- explicit gathering and lifecycle management consent
- Automatic compliance checking

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



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

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

University of Dublin
Trinity College

CS2041: Information Management I

An Introduction to the module
2017-18

Learning Outcomes: Information Management I

- Describe and use 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

Logistics

Timetable

- 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 web only

https://www.scss.tcd.ie/CourseModules/CS2041/index.php

Expectations

Bring Paper and Pen to every session!

Shut your laptop during Exercises, Student Presentations and Tutorials

Marking

20% Coursework · 80% Examination

Coursework will consist of Group based project and occasional individual assignments

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)*

Individual Assignment #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

Email me with:

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

**** NB Include in Subject Header "Data Power Assignment"**

Email me BY THIS WEDNESDAY 27th BEFORE 7PM