Student Online Teaching Advice Notice

The materials and content presented within this session are intended solely for use in a context of teaching and learning at Trinity.

Any session recorded for subsequent review is made available solely for the purpose of enhancing student learning.

Students should not edit or modify the recording in any way, nor disseminate it for use outside of a context of teaching and learning at Trinity.

Please be mindful of your physical environment and conscious of what may be captured by the device camera and microphone during videoconferencing calls.

Recorded materials will be handled in compliance with Trinity's statutory duties under the Universities Act, 1997 and in accordance with the University's policies and procedures.

Further information on data protection and best practice when using videoconferencing software is available at https://www.tcd.ie/info_compliance/data-protection/.

University of Dublin
Trinity College

CSU22041: Information Management I

An Introduction to the module
2020-2021

Gaye Stephens gaye.stephens@tcd.ie

Some Core Concepts

ORGANISATION

How data represented/associated

METADATA

Data about what the data is

ACCESS

How get at the data efficiently

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

Knowledge:

- Interrelating and "understanding"information

Normal- If Male and HBPM <=70 >=50

Patient with a heart condition

Gym

More Context

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!

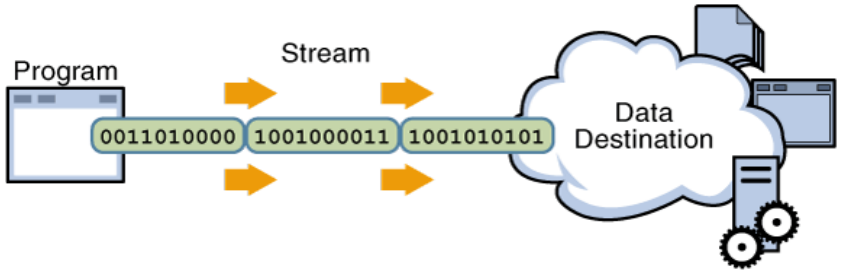
Maintaining structure in your own data file

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?

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

Use 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 Ireland top search terms of 2019

- Rugby World Cup
- Gay Byrne
- Storm Lorenzo
- Game of Thrones
- Brendan Grace
- Cameron Boyce
- EU election results
- Shane Lowry
- Joker
- Luke Perry

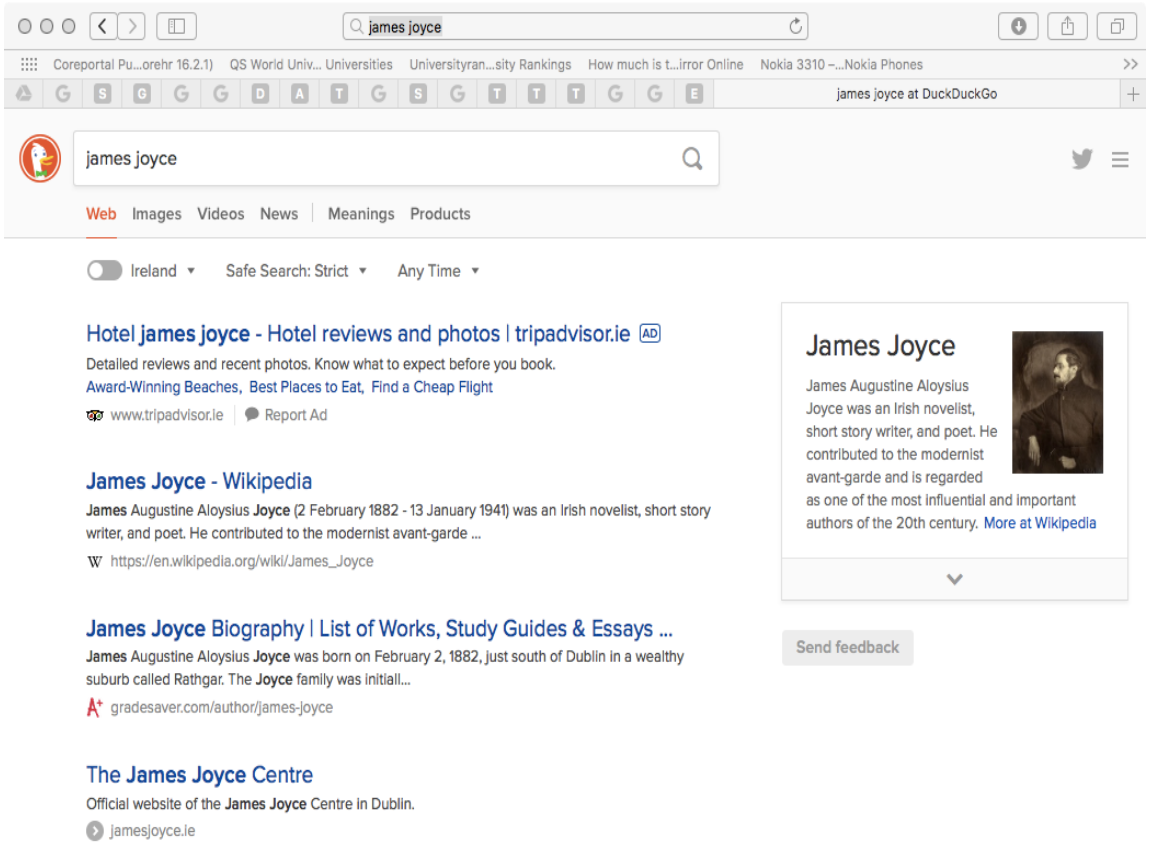
Nature of Results: Structured vs Unstructured

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

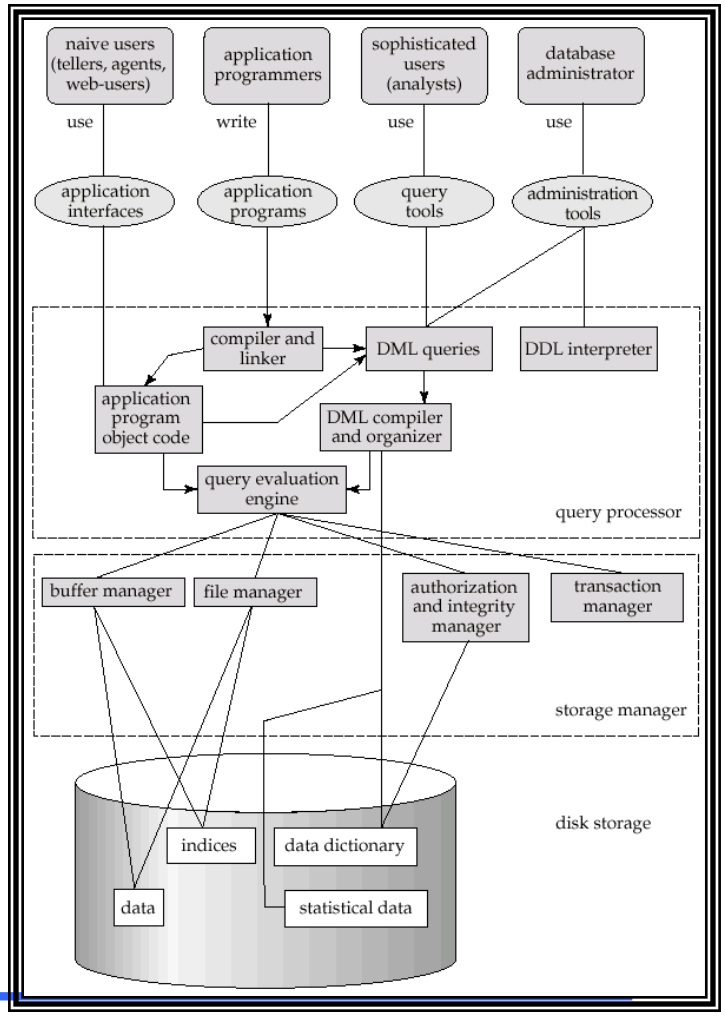
(Check the ACID terms out)

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)

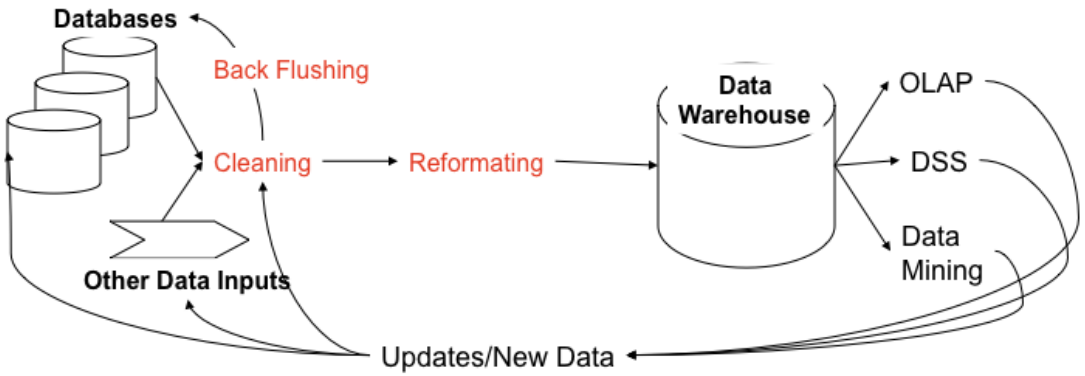
Data Warehouse is a subject oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions

Data Warehouse is a repository of data which is:

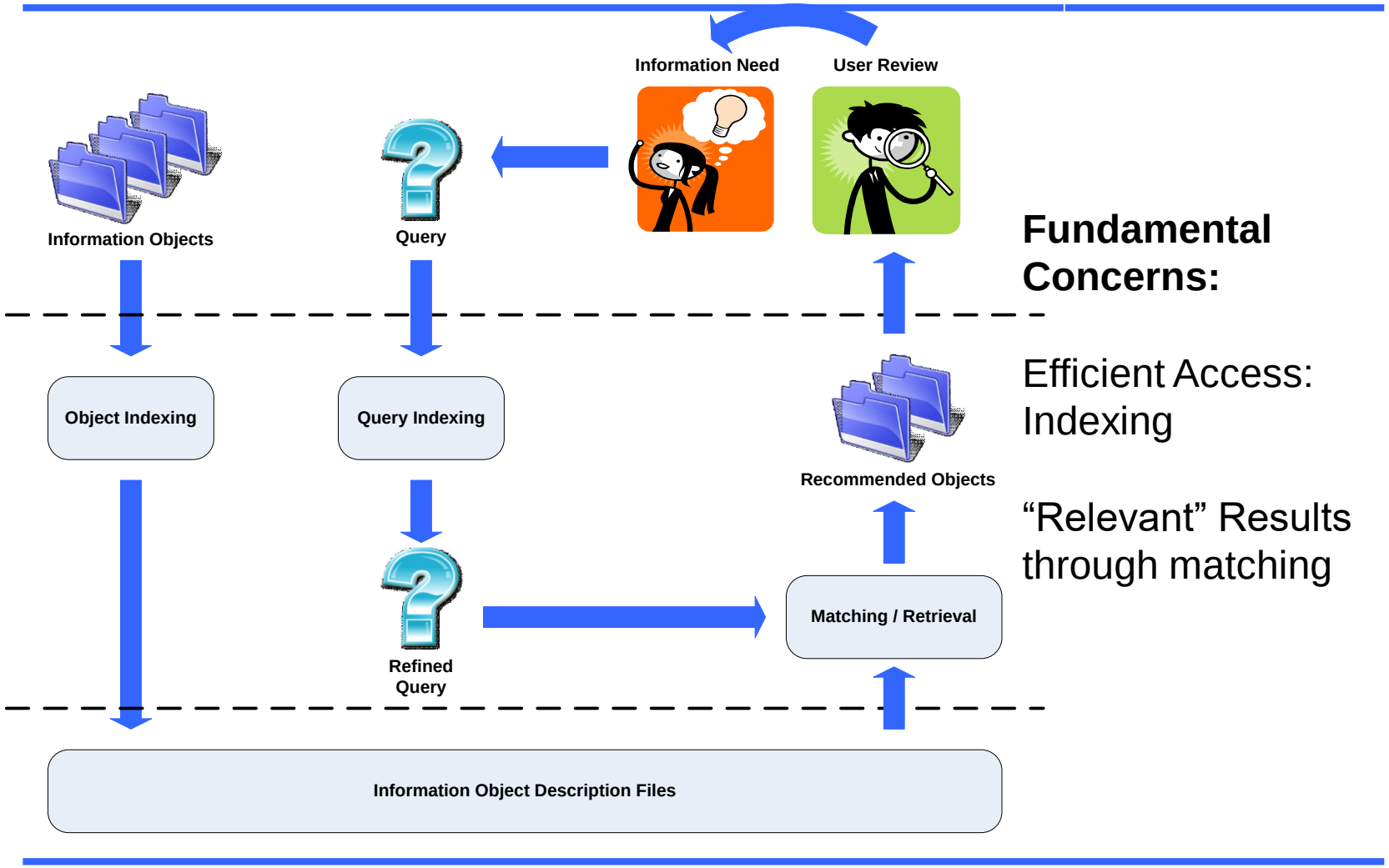
- Separate from operational systems and populated by data from these systems
- Provides a trend view of data
- Available entirely for the task of making data available to be interrogated by business users
- **Timestamped** and associated with defined periods of time, that is calendar periods or fiscal 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



Unstructured Approach: Information Retrieval



Fundamental Concerns:

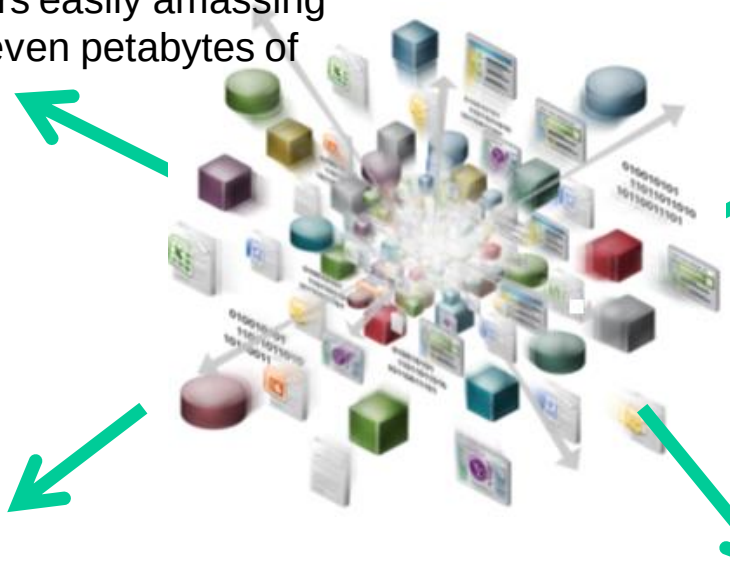
Efficient Access: Indexing

"Relevant" Results through matching

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.

Solution Trend for Velocity: "Big Data"

Desire to examine and derive new insights from information about:

enterprise (organisation, customers, suppliers and partners

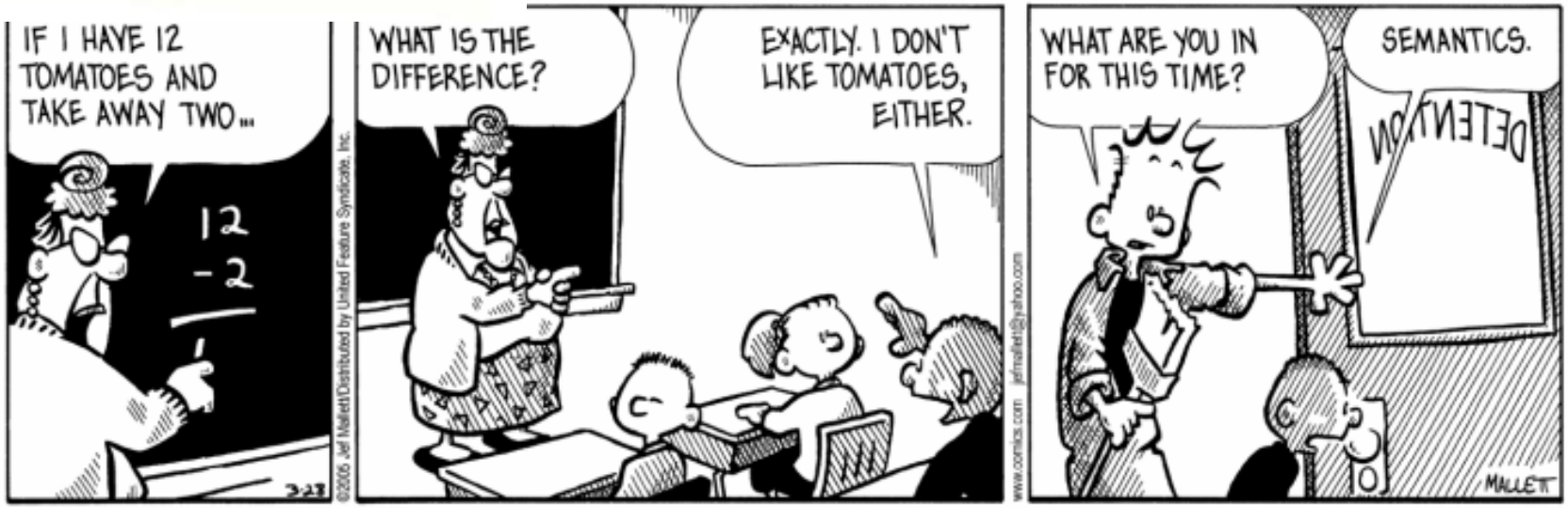
individuals (personsalisation, recommendations etc.)

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

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

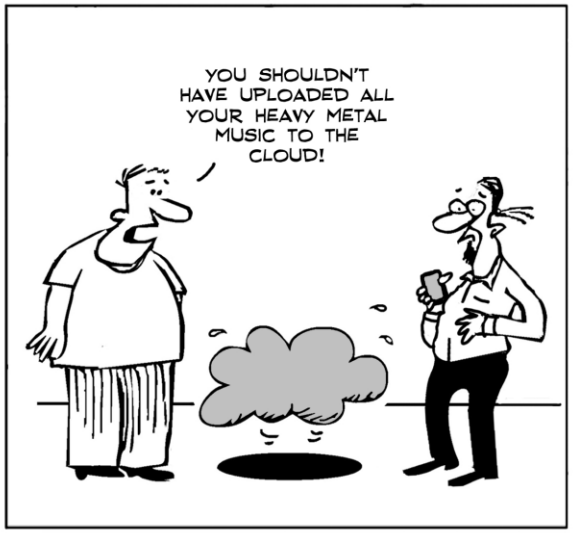
Solution Trend for coping with Variety:

Natural Language Processing (NLP)
and
Semantic Web Technologies



Solution Trend for coping with Volume: “The Cloud”

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



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

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