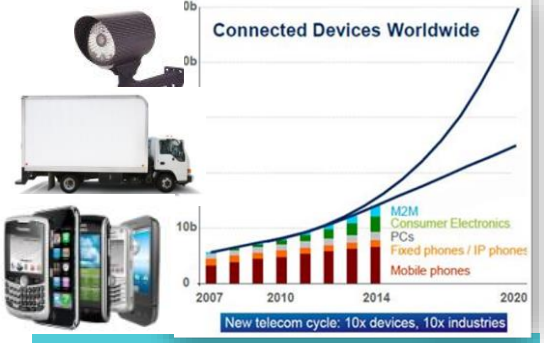
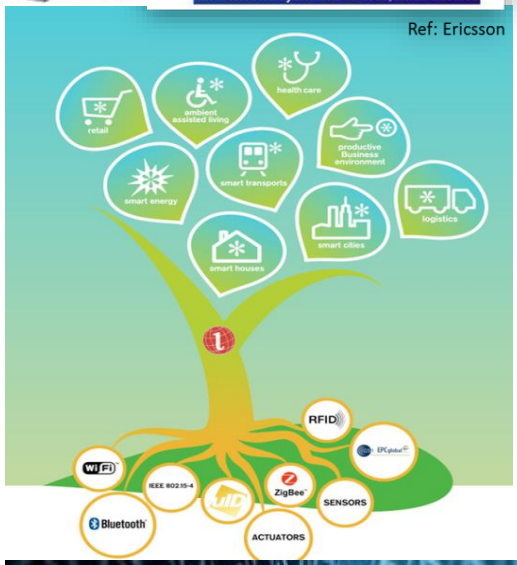
Big Data & Semantic Technology

John Davies
BT

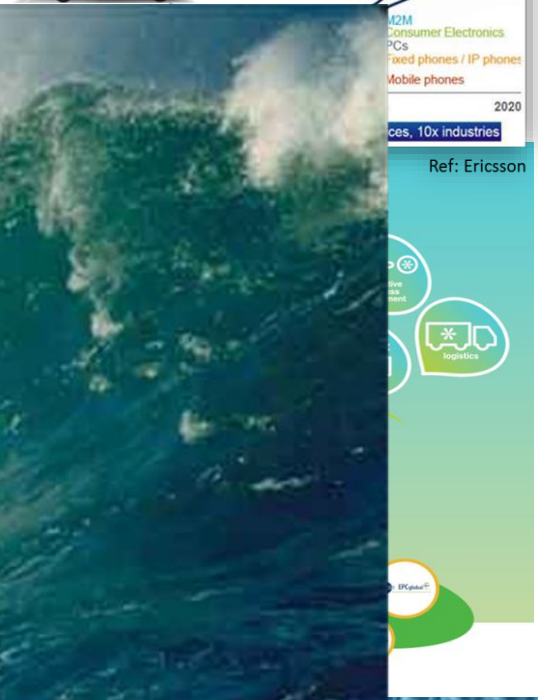
Big Data - motivation





- **Huge and rapidly growing data volumes** (*petabytes)*
  Internet companies – Google, Yahoo!, Facebook, Twitter...
  - large-scale parallel compute environments such as Hadoop, which are used by Google, Amazon etc to analyse big data sets, and create a high-performance computation facility
- **Increasing need to handle heterogeneous, unstructured data**
  sensor data, social media, call logs, textual info, …
- **Opportunity for large-scale, ad hoc analytics**
  but with separation from operational systems
- **BT – ~ few petabytes organisation**
  - probably a similar amount thrown away

Big Data - motivation



Big Data - motivation

Big Data: definition & implications

- Big Data can be deployed where traditional IT cannot deal with *volume or velocity* of data or required *processing* power

- Emergence of **technology** to address Big Data requirements: *scale-out*, commodity-based solutions
- Opportunity to change the **economics** of enterprise data – *reducing costs whilst increasing value*
- Enabling new approaches to analytics that facilitate the ***data-driven*** *(evidence-based)* enterprise

Implications

- Delivers new capabilities
  - Whole population product testing
    - Google: ‘Our 1 bn monthly unique visitors are our product testers’
  - Complete empirical studies on previously unanalysed data (e.g. pharma gene info)
- Does it change the economics of data?
- Who owns big data initiatives that cover the whole organisation?
- Skills are scarce

Big Data, Big Challenges

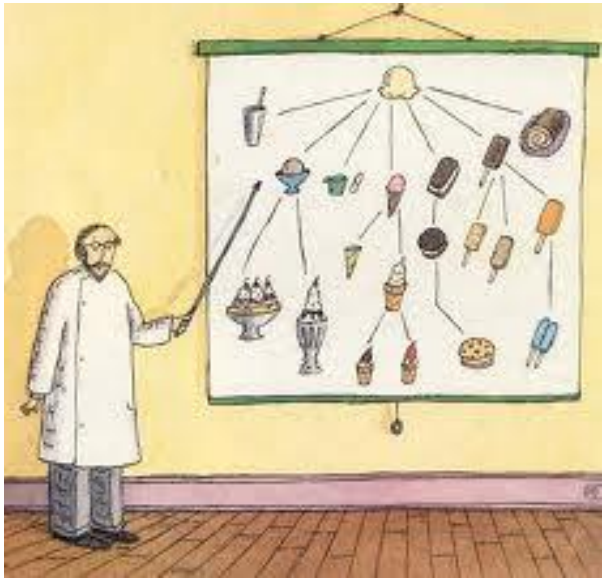
- Data discovery
- Data provenance, Trust
- Access Control

- Not new issues (encountered in RDBMS)
- Big data both *magnifies* them (new and immature field) provides a *new opportunity* to solve the problems however (*fresh impetus*)
- Tooling relatively weak

BT

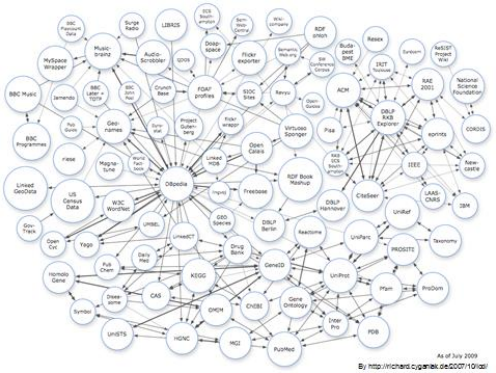
Semantic Technology & Big Data Challenges

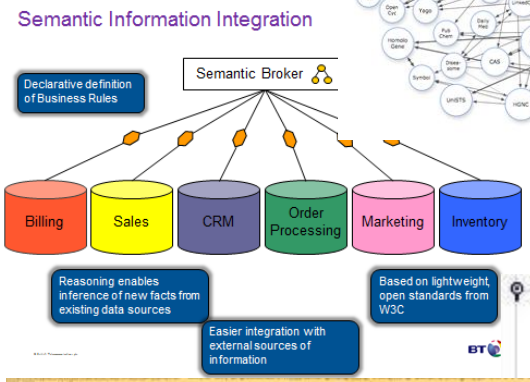
- Semantic tech community has been considering these issues for a while
- Data Discovery: semantic descriptions to aid discovery (voiD); semantic search
- Data provenance: PROV ontology
  - modelling agent-, object- and process-based provenance
- Trust:
  - Hoonoh – trust in info seeking
  - Deriving trust from provenance info

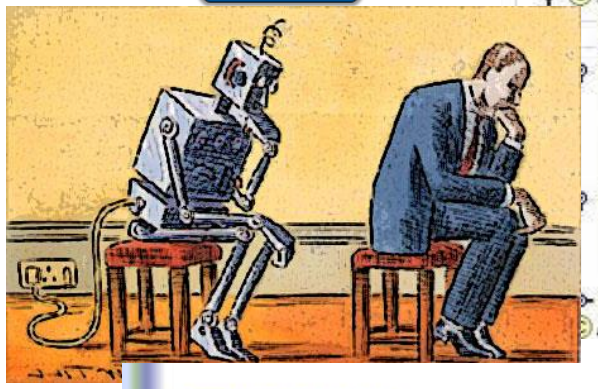


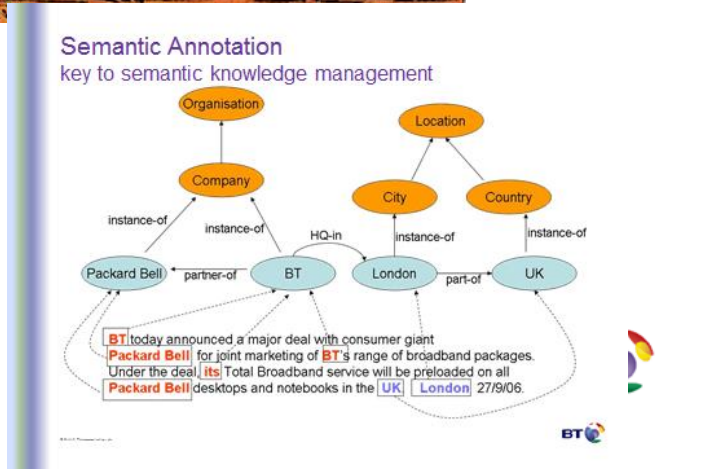
Semantic Technology

- Linked (Open) Data & the Data Web
- Information integration
- Reasoning & consistency
- Ontological modelling
- Semantic annotation of unstructured information









- Making (web) data machine processable
- *… Making (big) data machine processable*

Connecting Big Data

- Current focus on big data processing technology
- Longer term opportunity of integrating many data sources for bigger potential

  ⇨ Linked Data

  – lightweight, open standards, no re-engineering

- More connections, more network effects
- New relationships, new insights
- Higher value

Conclusions

- Emergence of **technology** to address Big Data requirements
- Opportunity to change the **economics** of enterprise data – *reducing costs whilst increasing value?*
- Big data amplifies problems, such as data discovery, provenance, data quality and data control problems, trust, etc
- Semantic Technology has been addressing some big data issues in another context – transferability?
- Linked Data is a way to inter-connect big data sources for greater value

BT