Karlsruhe Institute of Technology

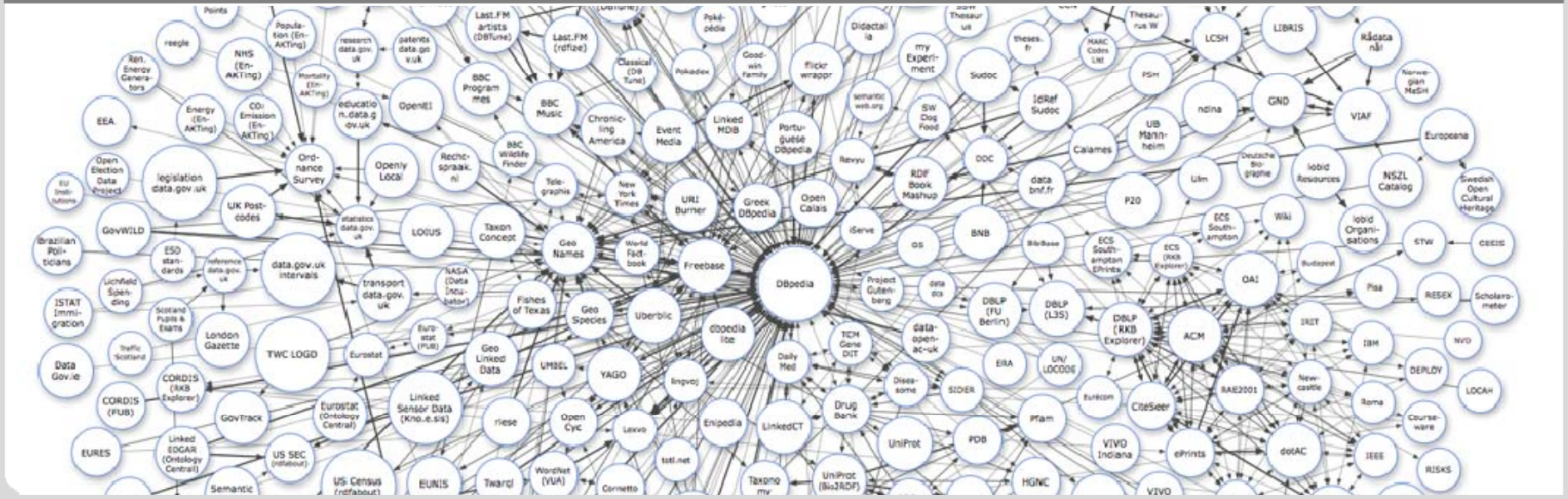
SINTEF

Institut "Jožef Stefan" Ljubljana, Slovenija

Introduction to Linked Data

Marko Grobelnik, Andreas Harth, Dumitru Roman

Big Linked Data Tutorial Semantic Days 2012



Tutorial Agenda

- Introduction to Linked Data (45 m – 60 m) Andreas
- Consuming Norwegian Linked Data (30 m) Titi
- Large Scale Linked Data Management (30 m) Andreas
- Big Data Intro and Analytics (60 m – 90 m) Marko
- Questions & Answers Session (30 m) all

Introduction to Linked Data (Andreas)

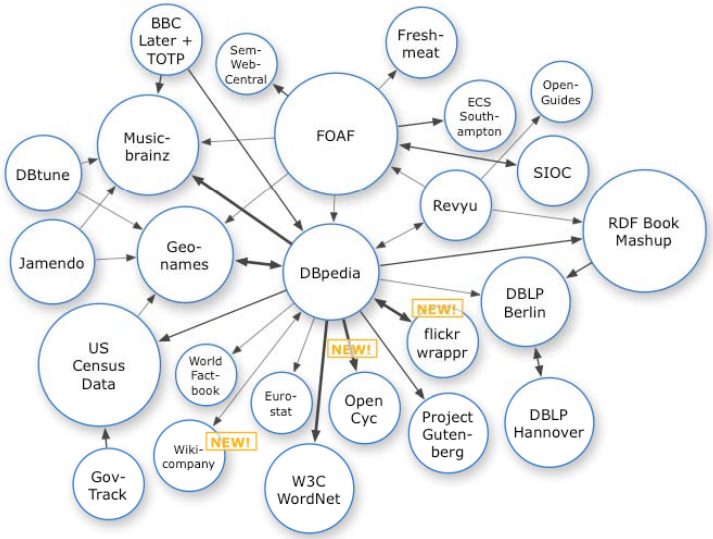
- Motivation
- Linked Data
  - Principles (Web Architecture and RDF, Resource Description Framework)
  - SPARQL RDF Query Language
- Ontology Languages
  - RDF Vocabulary Description Language (RDFS)
  - Web Ontology Language (OWL)
- Application Architectures
- Summary

MOTIVATION

Motivation

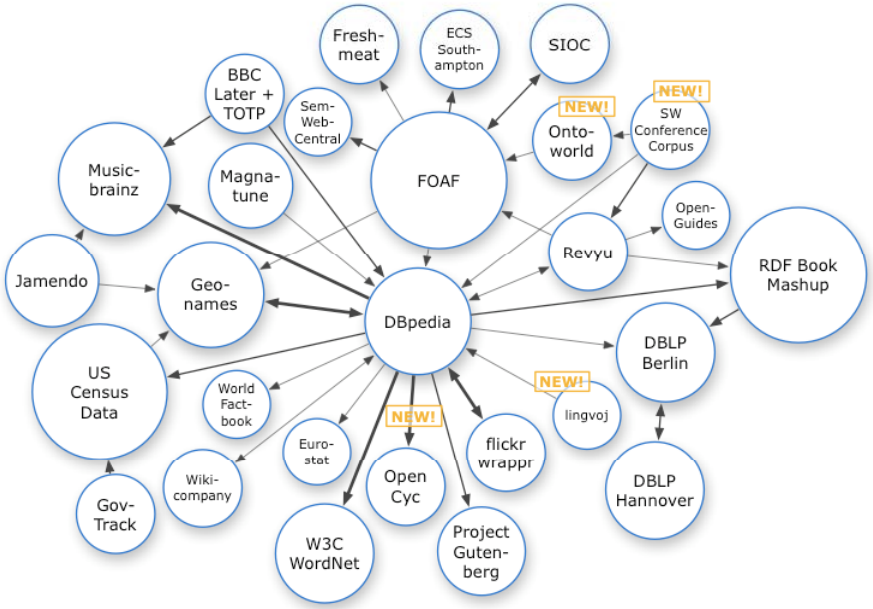
- With increased use of computers more and more data is being stored
  - Organisations rely on data for business decisions
  - Data drives policy decisions in government
  - Individuals rely on data from the Web for information and communication
- Data volumes explode
  - More and more data available on the Web is represented in Semantic Web standards
  - Linking Open Data (LOD) initiative
- Semantic Web technologies facilitate the integration of data from multiple sources
- Combining data from multiple sources enables insights

Linked Data on the Web



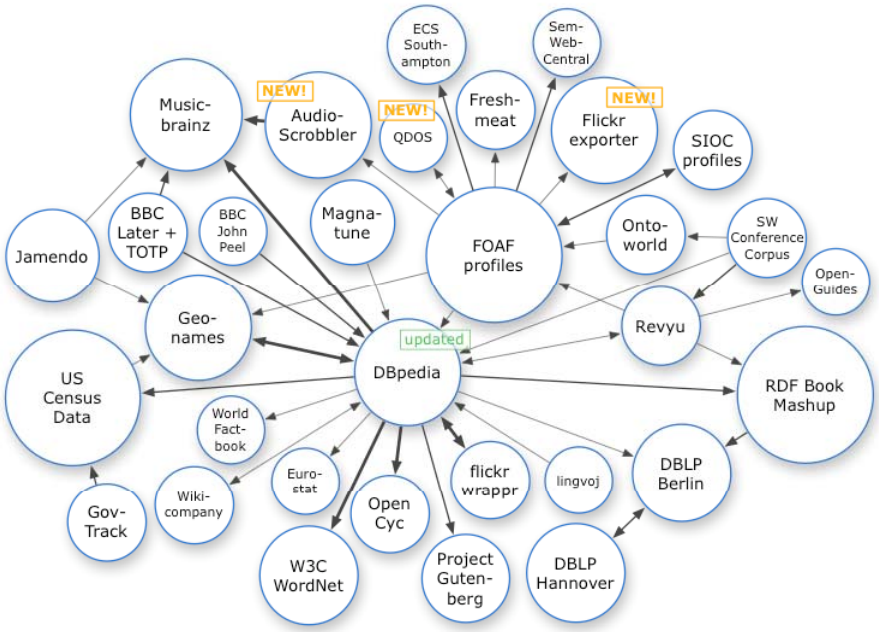
2007-10

Linked Data on the Web



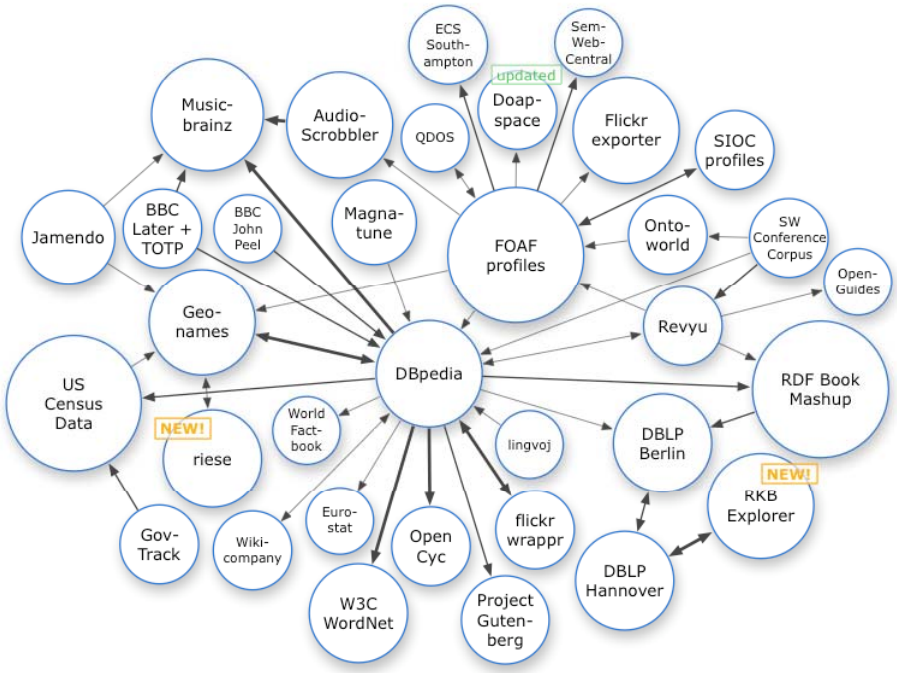
2007-11

Linked Data on the Web



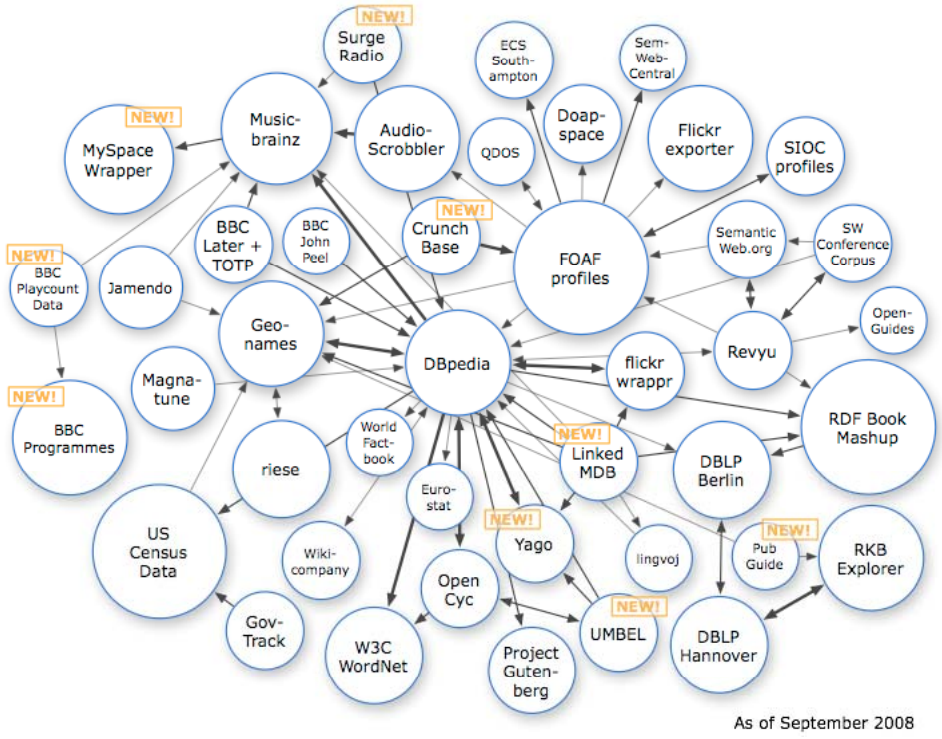
2008-02

Linked Data on the Web



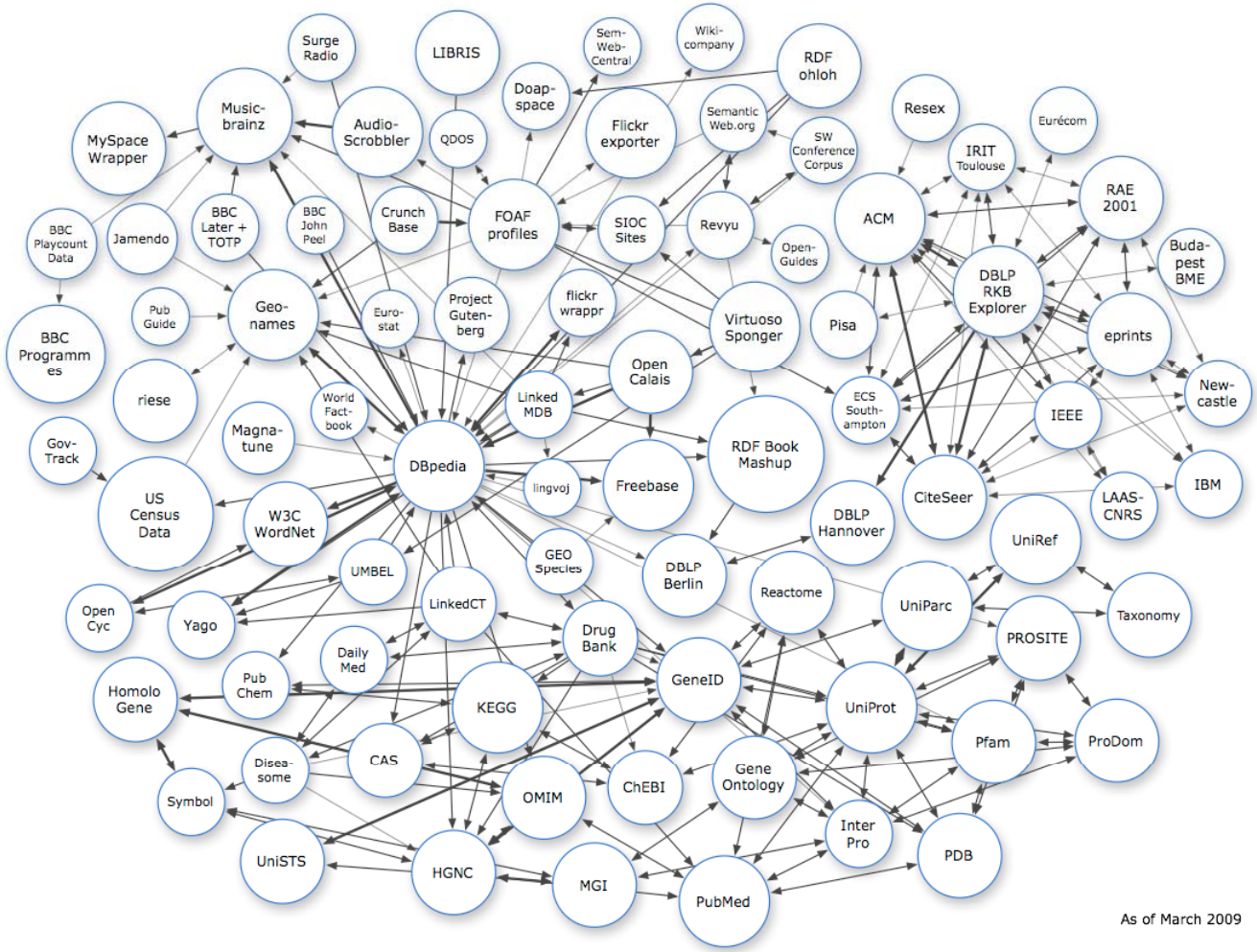
2008-03

Linked Data on the Web



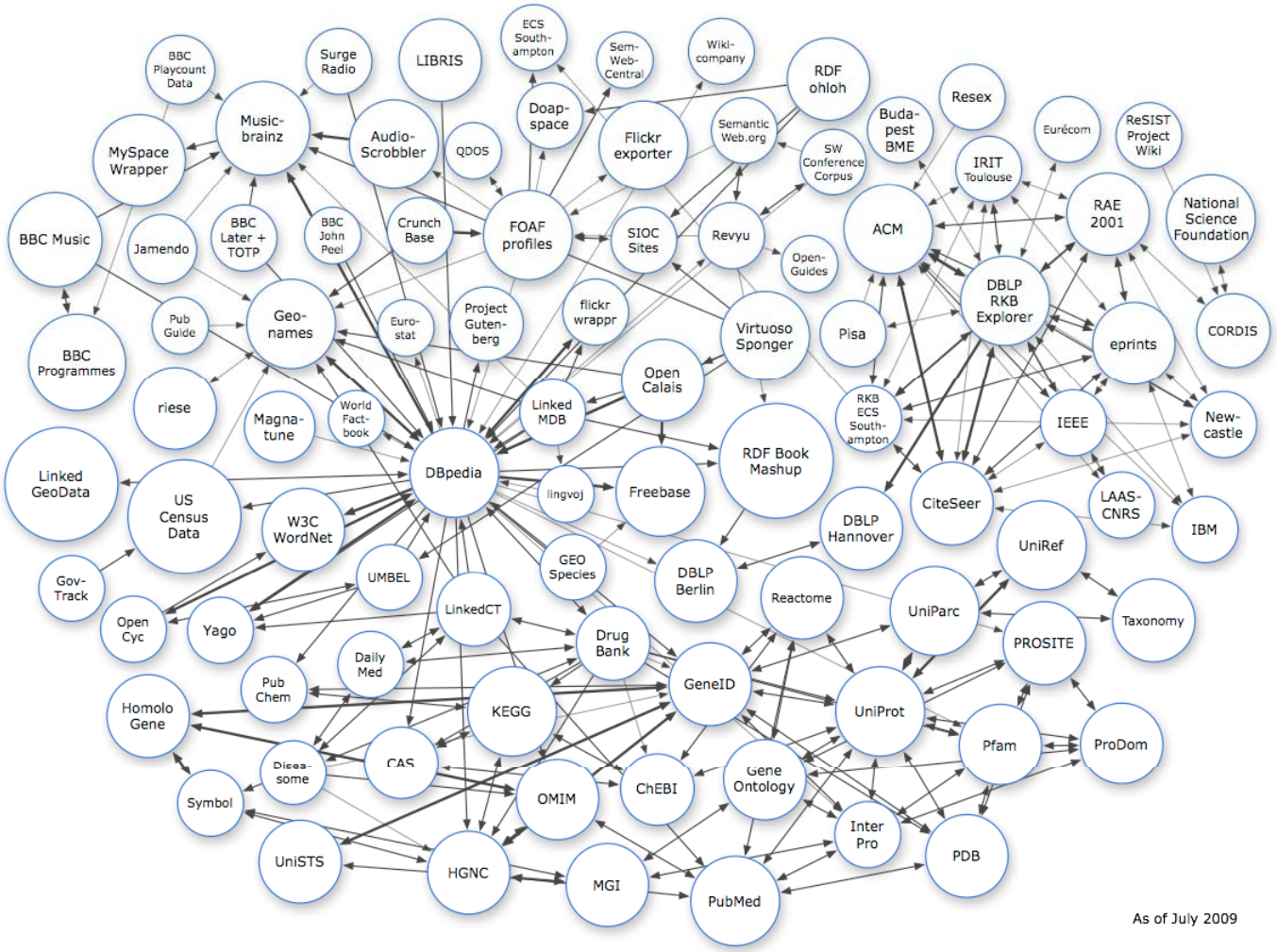
2008-09

Linked Data on the Web



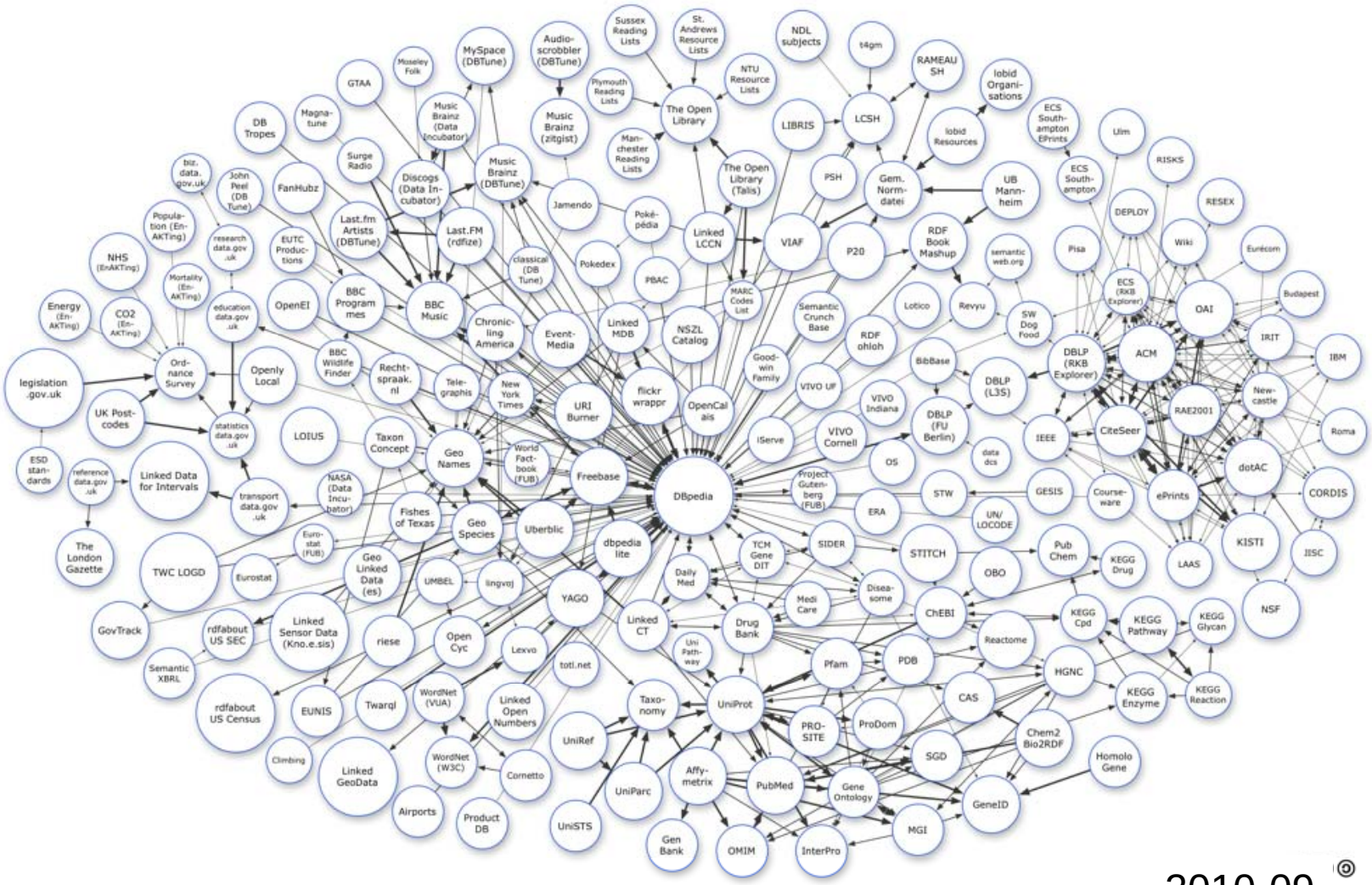
2009-03

Linked Data on the Web



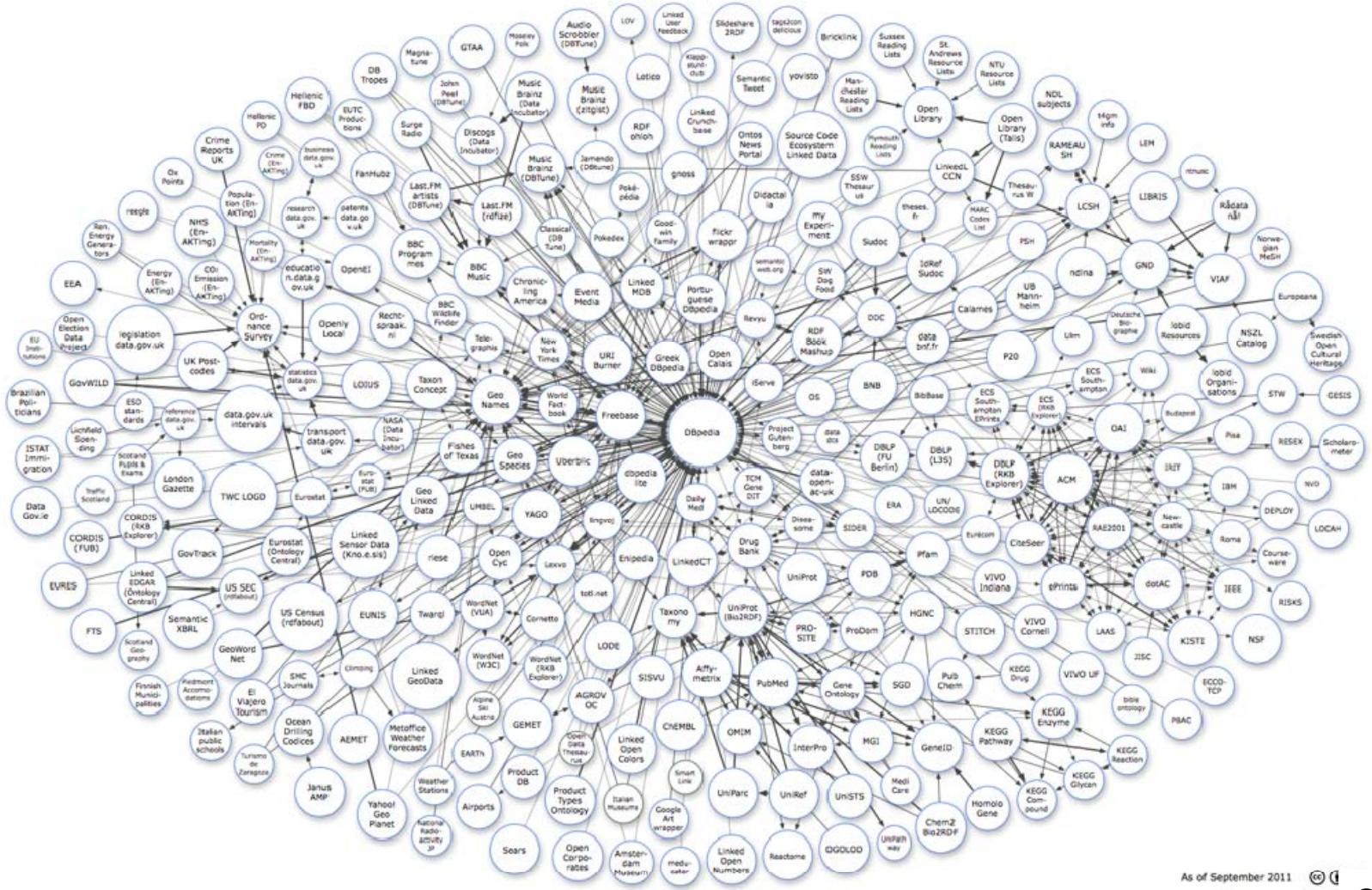
2009-07

Linked Data on the Web



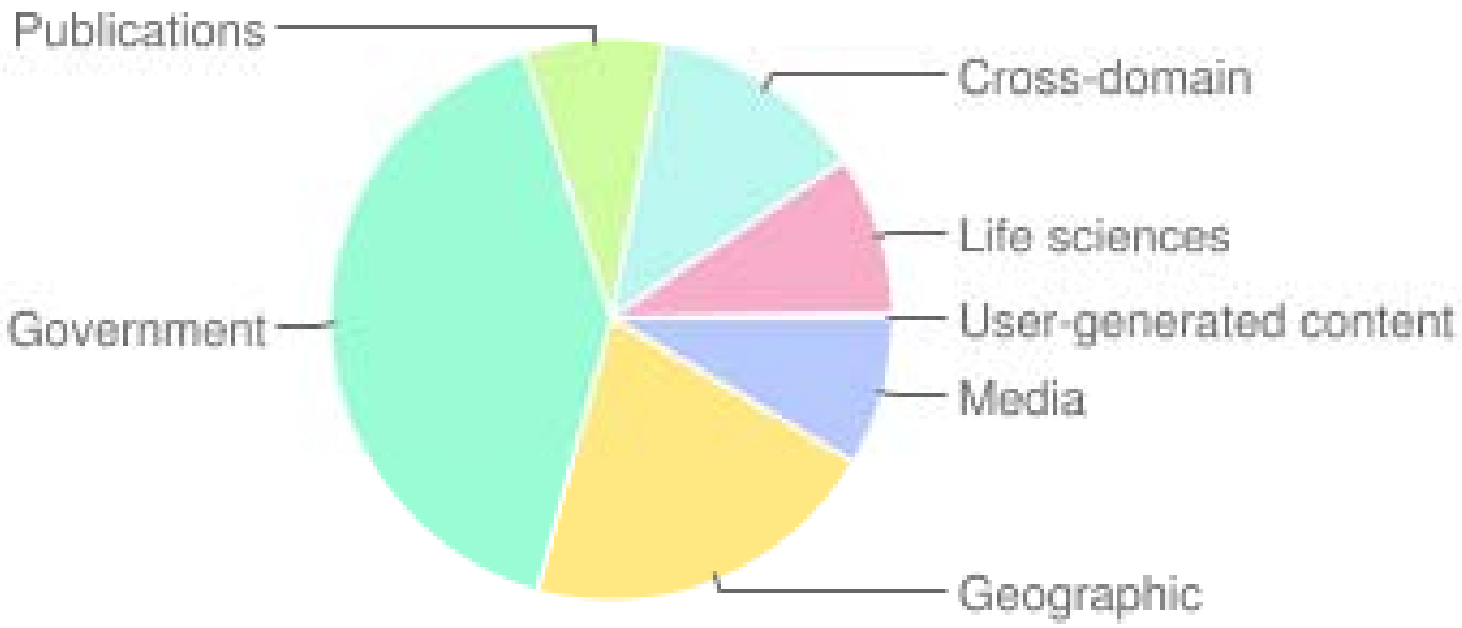
2010-09

Linked Data on the Web



2011-09

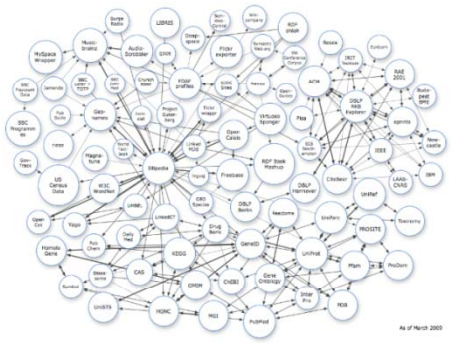
Types of Data in the Linking Open Data Cloud

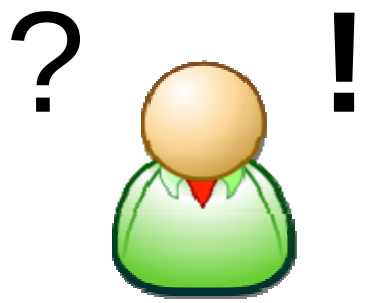


http://www4.wiwiss.fu-berlin.de/lodcloud/state/ (Sept 2010)

Scenario Overview

- Semantic Technologies facilitate access to data





- Q: data about Berlin?
- Q: famous people that died in Berlin?
- Q: data about Hegel?
- Q: Hegel's publications?
- Q: data about Marlene Dietrich?
- Q: Dietrich's songs?

DBpedia

- Linked Data version of Wikipedia
- Scripts that extract data (text, links, infoboxes) from Wikipedia
- Published as Linked Data
- Interlinking hub in the Linked Data web

- Berlin
  - http://dbpedia.org/resource/Berlin
- Hegel
  - http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_Hegel

- Marlene Dietrich
  - http://dbpedia.org/resource/Marlene_Dietrich

BBC Music

- Data about BBC (radio) programmes, artists, songs…
- Combination of BBC-internal data (playlists), MusicBrainz (artists, albums), Wikipedia (artists)
- Underpinning the BBC Music website
- Data published according to Linked Data principles

- Marlene Dietrich
  - http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5.rdf#artist

Virtual International Authority File (VIAF)

- Joint project of national libraries and related organisations
  - 21 institutions, among them the Library of Congress, Deutsche Nationalbibliothek, Bibliothèque nationale de France
- Provide access to “authority files”
- Matching and interlinking collections from participating institutions

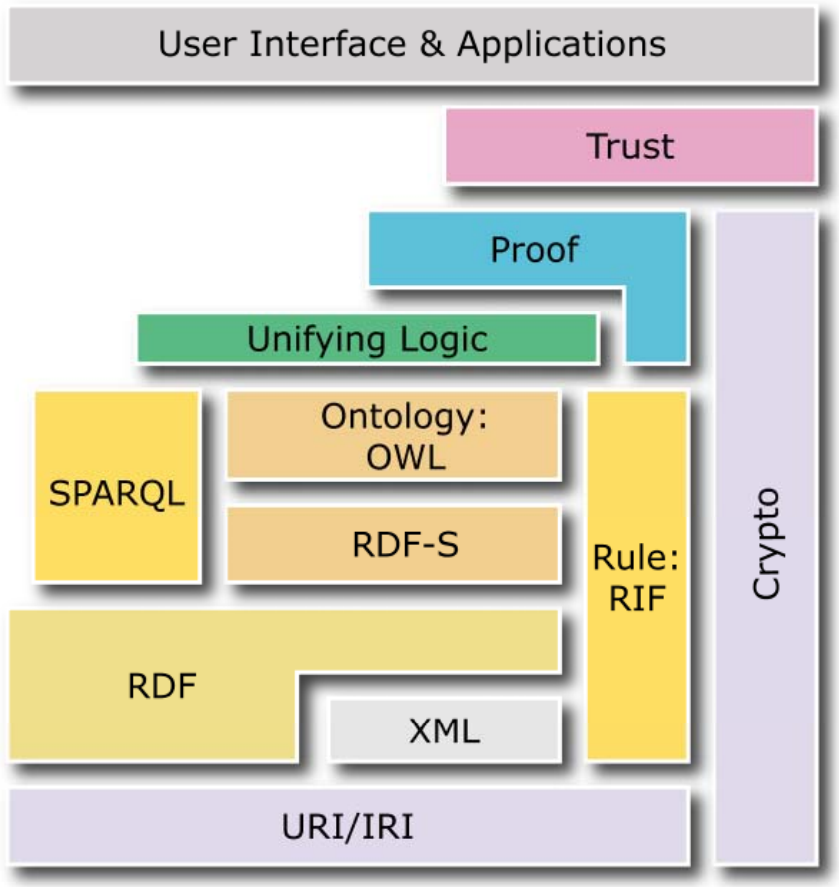
- Hegel
  - http://viaf.org/viaf/89774942/

- Marlene Dietrich
  - http://viaf.org/viaf/97773925/

LINKED DATA PRINCIPLES

Semantic Technologies

- Semantic Web technologies, standardised by the W3C, are mature:
  - RDF recommendation in 1999, update in 2004
  - RDFa (RDF in HTML) note in 2008
  - RDFS recommendation in 2004
  - SPARQL recommendation in 2008
  - OWL recommendation in 2004, update in 2009
- Linked Data is a subset of the Semantic Web stack, including web architecture:
  - IRI (IETF RFC 3987, 2005)
  - HTTP (IETF RFC 2616, 1999)



Linked Data Principles

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

http://www.w3.org/DesignIssues/LinkedData

1. Use URIs as Names for Things

- Use a unique identifier to denote things
- URIs are defined in RFC 2396

- Hegel, Georg Wilhelm Friedrich
  - http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_Hegel
  - http://viaf.org/viaf/89774942/
  - …
- Hegel, Georg Wilhelm Friedrich: Gesammelte Werke / Vorlesungen über die Logik
  - urn:isbn:978-3-7873-1964-0

Names for Things

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

2. Use HTTP URIs

- Enables “lookup” of URIs
- Via Hypertext Transfer Protocol (HTTP)
- Piggy-backs on hierarchical Domain Name System to guarantee uniqueness of identifiers
- Uses established HTTP infrastructure
- Connects logical level (thing) with physical level (source)
- Important: distinction between “thing URI” and “source URI” („other resource“ vs. „information resource“)

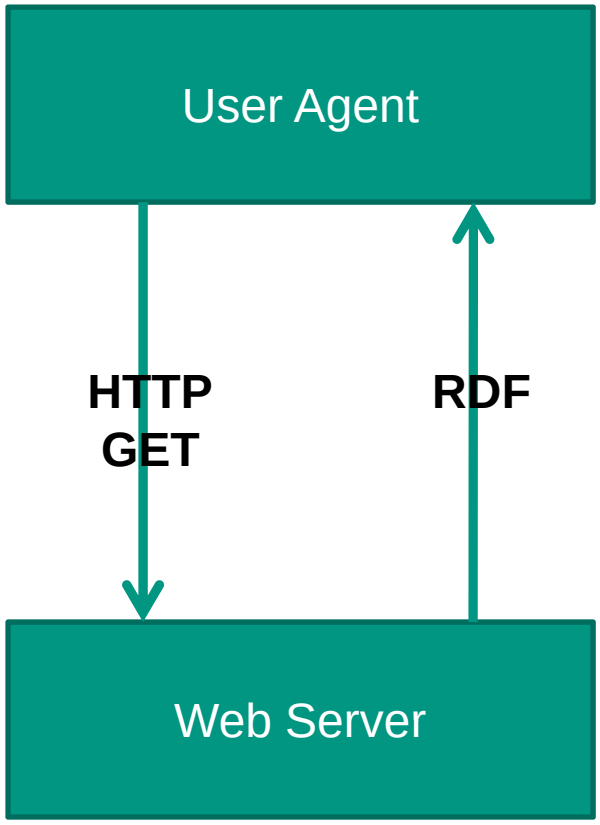
Information Resources vs. Other Resources

Marlene Dietrich, the person

File containing data about Marlene Dietrich

Name?
Creator?
Birth date?
Last change date?
License?
Copyright?
…

Correspondence between thing-URI and source-URI („hash URIs“)



http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5.rdf#artist

http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5.rdf

Hypertext Transfer Protocol (HTTP)

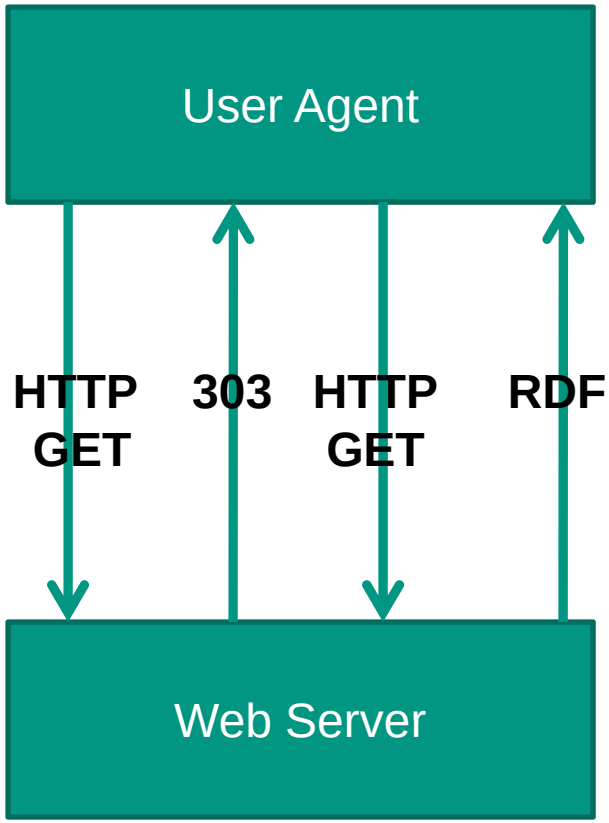
REQUEST

```
$ curl -H "Accept: application/rdf+xml" -v
  http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-
  02b20c7a9dd5.rdf#artist
> GET /music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5.rdf
  HTTP/1.1
> User-Agent: curl/7.25.0
> Host: bbc.co.uk
> Accept: application/rdf+xml
```

RESPONSE

```
< HTTP/1.1 200 OK
< Date: Tue, 08 May 2012 07:12:19 GMT
< Server: Apache/2.2.3 (Red Hat)
< Content-Type: application/rdf+xml
< Content-Length: 1956
<
{ [data not shown]
```

Correspondence between thing-URI and source-URI („slash URIs“)



http://dbpedia.org/resource/Marlene_Dietrich

http://dbpedia.org/data/Marlene_Dietrich

http://dbpedia.org/page/Marlene_Dietrich

3. Provide Useful Information

- When somebody looks up a URI, return data using the standards (RDF*, SPARQL)
- Resource Description Framework, a format for encoding graph-structured data (with URIs to identify nodes/vertices and links/edges)

Resource Description Framework

- Directed, labeled graph
- `triple(subject, predicate, object)`
  - subject: URI (or blank node)
  - predicate: URI
  - object: URI (or blank node) or RDF literal (string, integer, date…)

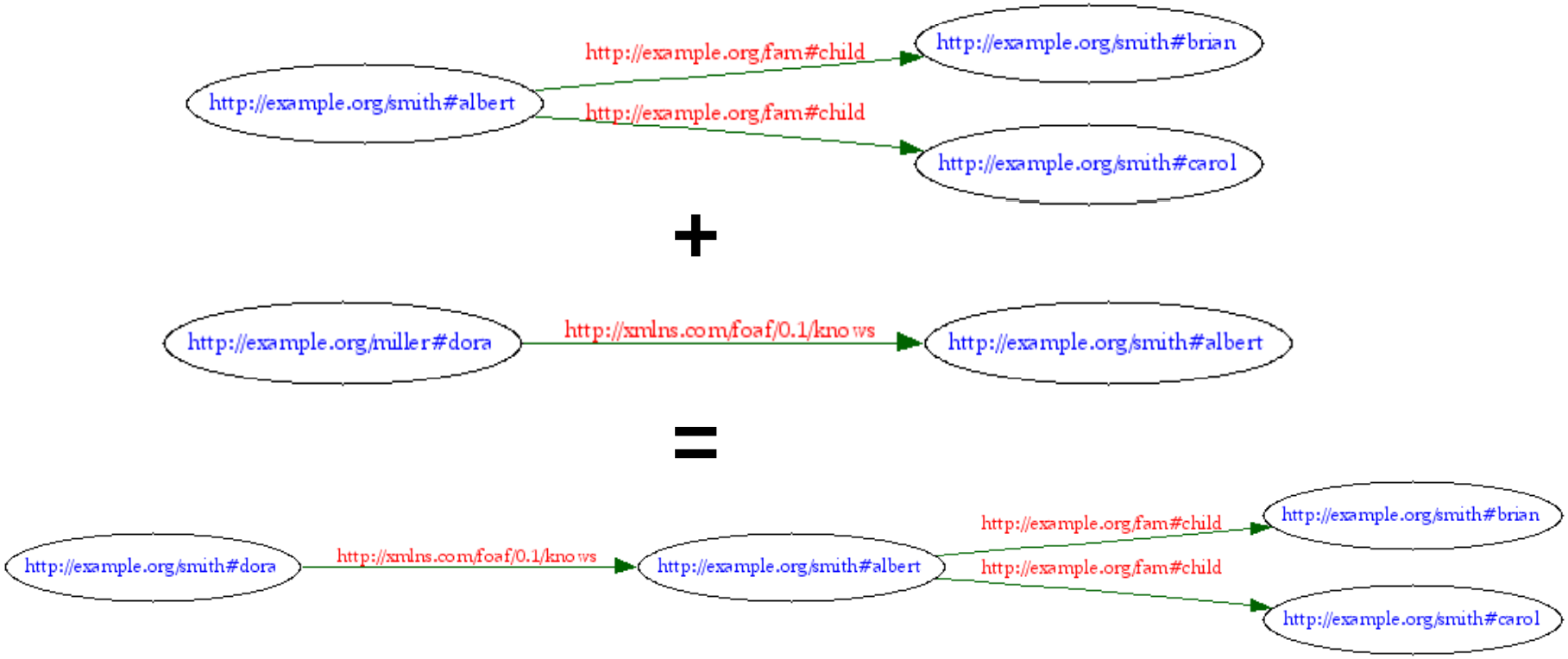
- RDF/XML is the most widely deployed serialisation
- Other serialisations possible (N-Triples, Turtle, Notation3…)

- Quadruples (or quads) used as internal representation when integrating data
- `quad(subject, predicate, object, context)`
  - context: URI (used to store origin of triple)

RDF Example

```
dbpedia:Georg_Wilhelm_Friedrich_Hegel rdf:type
foaf:Person .
dbpedia:Georg_Wilhelm_Friedrich_Hegel rdf:type
yago:PoliticalPhilosophers .
dbpedia:Georg_Wilhelm_Friedrich_Hegel
rdfs:comment "Georg Wilhelm Friedrich Hegel var
en tysk filosof."@no .
dbpedia:Georg_Wilhelm_Friedrich_Hegel dbpedia-
owl:influenced dbpedia:Francis_Fukuyama .
dbpedia:Georg_Wilhelm_Friedrich_Hegel dbpedia-
owl:influenced dbpedia:Friedrich_Nietzsche .
```

Merging Data with RDF



4. Link to Other URIs

- Enable people (and machines) to jump from server to server
- External links vs. internal links (for any predicate)
- Special owl:sameAs links to denote equivalence of identifiers (useful for data merging)

Equivalences via owl:sameAs

http://viaf.org/viaf/89774942/

- http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_He
- http://www.idref.fr/026917467/id
- http://libris.kb.se/resource/auth/190350
- http://d-nb.info/gnd/118547739

http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5#artist

- http://dbpedia.org/resource/Marlene_Dietrich

http://viaf.org/viaf/97773925/

- http://dbpedia.org/resource/Marlene_Dietrich .
- http://d-nb.info/gnd/118525565
- http://libris.kb.se/resource/auth/238817
- http://www.idref.fr/027561844/id

http://dbpedia.org/resource/Berlin

- http://mpii.de/yago/resource/Berlin
- http://data.nytimes.com/N50987186835223032381 - Berlin (Germany)
- http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Berlin
- http://data.nytimes.com/16057429728088573361 - Gaspe Peninsula (Quebec) (?)

SPARQL RDF PROTOCOL AND QUERY LANGUAGE

SPARQL

- SPARQL Protocol and RDF Query Language
- Query language for RDF graphs
- “SQL for RDF”
- SPARQL specification consists of
  - Query language
  - Result formats (representation of results in RDF and XML)
  - Query protocol (mechanisms to pose queries and retrieve results)

Simple Query Example

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT *
WHERE {
  ?s dct:subject
    <http://dbpedia.org/resource/Category:People_from_Stavanger> .
  ?s rdfs:label ?name.
}
```

- Main part is query pattern (`WHERE` clause)
  - Using Turtle syntax for RDF
  - Query patterns may contain variables (`?s, ?name`)
- Shortcuts for URIs (`PREFIX`)
- Query results via selection of variables (`SELECT`)

Query Results

- Table with one row per result

?s	?name
http://dbpedia.org/resource/Erik_Nevland	"Erik Nevland"@no
http://dbpedia.org/resource/Jan_Simonsen	"Jan Simonsen"@no
http://dbpedia.org/resource/Laila_Goody	"Laila Goody"@no
http://dbpedia.org/resource/Henriette_Henriksen	"Henriette Henriksen"@no
http://dbpedia.org/resource/Guri_Hjeltnes	"Guri Hjeltnes"@no
http://dbpedia.org/resource/Johan_E._Holand	"Johan E. Holand"@no
http://dbpedia.org/resource/Kristian_Valen	"Kristian Valen"@no
…	…

Further Functionality

- Optional triple patterns (e.g., return name and optionally birthdate if available)
- Unions (e.g., return material scientists and also physicists)
- Filter (e.g., only return scientists born before 1970)
- Result formats (e.g., return RDF triples instead of results table)
- Modificators (e.g., sort results, only return unique results)

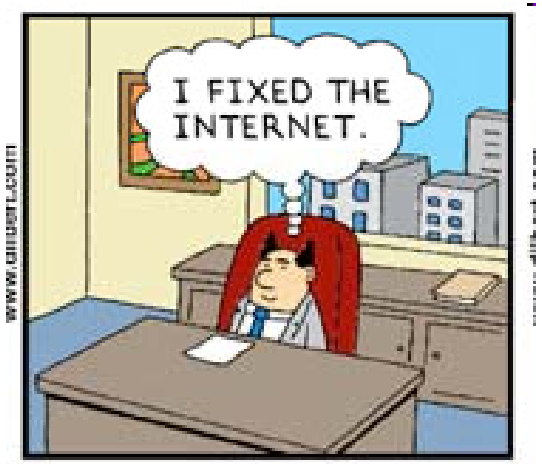
Benefits of Linked Data

- Explicit, simple data representation
  - Common data representation (Resource Description Framework, RDF) hides underlying technologies and systems
- Distributed System
  - Decentralised distributed ownership and control facilitates adoption and scalability
- Cross-referencing
  - Allows for linking and referencing of existing data, via reuse of URIs
- Loose coupling with common language layer
  - Large scale systems require loose coupling, via HTTP as common access protocol
- Ease of publishing and consumption
  - Simple and easy-to-use systems and technologies to facilitate uptake
- Incremental data integration
  - Start with merged RDF graphs and provide mappings as you go

Challenges (I)

- Ramp-up cost for data conversion
  - May be alleviated by semi-automatic mappings and adequate tool support for manual conversion
- Integrated data may be messy at first
  - But can be refined as need arises
- Distributed creation and loose coordination may result in inconsistencies
  - Can be detected, diagnosed, and fixed with appropriate tools

The Pedantic Web Group



- Get the community to contact publishers about errors/issues as they arise
- Get involved: http://pedantic-web.org/
- 137 members!

- Acknowledgements to: Aidan Hogan, Alex Passant, Me, Antoine Zimmermann, Axel Polleres, Michael Hausenblas, Richard Cyganiak, Stéphane Corlosquet

Challenges (II)

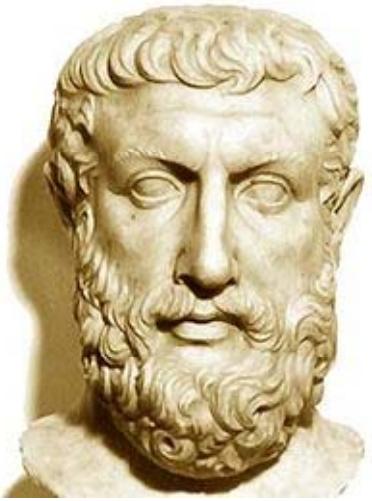
- Often very much oriented towards individuals
- Little possibilities for expressing schema knowledge
- Different data sources have different ways of representing the same facts

- Ontology languages (RDFS, OWL) solve that drawback
- RDFS and OWL are layered on top of RDF

ONTOLOGY LANGUAGES

Ontology in Philosophy

- Term exists only in singular (there are no “ontologies”)
- Ontology is concerned with the study of the nature of being, existence or reality as such
- Discussed by Aristoteles (Sokrates), Thomas von Aquin, Descartes, Kant, Hegel, Wittgenstein, Heidegger, Quine, ...



Ontology in Informatics

“An Ontology is a
formal specification > interpretable by machines
of a shared > based on consensus
conceptualisation > describes terminology
of a domain of interest” > covers a specific topic

Studer, Benjamins and Fensel (1998) based on Gruber (1993) and Borst (1997)

Schema Knowledge

- RDF provides universal mechanism for the representation of facts using triples
- Possible to describe individuals and their relations
- Required: describe generic sets of individuals (classes), e.g., people, chemical compounds, organisations…
- Required: specification of logical connections between individuals, classes and properties to describe their meaning, e.g., “researchers write papers”, “materials are chemical compounds”
- In database-speak: schema knowledge

Schema Knowledge with RDFS

- RDF Vocabulary Description Language (RDFS)
- Allows for specification of schema (also: terminological) knowledge
- RDFS is a special RDF vocabulary (every RDFS document is an RDF document)
- RDFS vocabulary is generic: allows to specify the semantics of other vocabularies (and as such is a kind of “metavocabulary”)
- Thus, RDFS is an ontology language (but a lightweight ontology language)
- “A little semantics goes a long way” (Hendler, 1997)

Classes and Instances

- Property `rdf:type` defines the subject of a triple as of type of the object
- Object of the triple is interpreted as identifier for the class, which contains the resources denoted via subject of the triple
  Example:
  "The individual Hegel is of type Person"
  `dbpedia:Georg_Wilhelm_Friedrich_Hegel rdf:type foaf:Person .`
- Class membership is not exclusive:
  Example:
  `dbpedia:Georg_Wilhelm_Friedrich_Hegel rdf:type yago:PoliticalPhilosophers .`
- Instances and classes both use same syntax for URIs, so no syntactical distinction

Subclasses - Motivation

- Given triple
  - `dbpedia:Georg_Wilhelm_Friedrich_Hegel rdf:type yago:PoliticalPhilosophers .`
  - and a query for all `foaf:Person` instances
    - we do not get any results
- We could add the triple
  - `dbpedia:Georg_Wilhelm_Friedrich_Hegel rdf:type foaf:Person .`
    - but would solve the problem only for one instance

Subclasses

- Solution:
  - Make one statement which says that every scientist is a person
  - Which means every instance of class `yago:PoliticalPhilosophers` is also an instance of class `foaf:Person`

- Realised via `rdfs:subClass` property

  Example:
  "The class of political philosophers is a subclass of the class of persons"

```
yago:PoliticalPhilosophers rdfs:subClassOf
foaf:Person .
```

Subclasses

- `rdfs:subClassOf` is reflexive, that is, every class is a subclass of itself

  Example:

  ```
  yago:PoliticalPhilosophers rdfs:subClassOf
  yago:PoliticalPhilosophers .
  ```

- Possible to equate two classes via reciprocal subclass relations:

  Example:

  ```
  dbpedia:Person rdfs:subClassOf foaf:Person .
  foaf:Person rdfs:subClassOf dbpedia:Person .
  ```

Class Hierarchies

- Typically, ontologies contain not only single subclass relations, but class hierarchies
  Example:

```
yago:PoliticalPhilosophers rdfs:subClassOf
                                    yago:Philosophers .
yago:Philopsophers rdfs:subClassOf dbpedia:Person .
dbpedia:Person rdfs:subClassOf dbpedia:Mammal .
```

- Transitivity of `rdfs:subClassOf` is part of the RDFS semantics, which means e.g., the following holds:

```
Example:
dbpedia:Philopsophers rdfs:subClassOf dbpedia:Mammal .
```

Further RDFS Primitives

- Property hierarchies via `rdfs:subPropertyOf`
- Restrictions on properties via `rdfs:domain` and `rdfs:range`
- Lists and collections
- Reification (statements about statements)
- Annotations via `rdfs:label` or `rdfs:comment`

RDFS Summary

- RDFS can be used to describe semantic aspects of specific domains
- On the basis of RDFS it is possible to infer implicit knowledge
- However, the primitives of RDFS have limited expressivity

Web Ontology Language OWL

- Fragment of first-order logics
- Five variants: OWL EL, OWL RL, OWL QL, OWL DL, OWL Full
- OWL DL is decidable and has a corresponding description logics SROIQ (D)
- OWL documents are RDF documents
- Three building blocks are
  - Classes (comparable to classes in RDFS)
  - Individuals (comparable to instances in RDFS)
  - Roles (comparable to properties in RDFS)
- OWL contains primitives to specify elaborate expressions, e.g. two classes are disjoint
- OWL allows for complex reasoning tasks such as consistency check, but may be computationally expensive

Equivalence

- OWL allows for specification of equivalence; needed in data integration scenarios
- Between individuals: owl:sameAs
  Example:
  `<http://viaf.org/viaf/97773925/> owl:sameAs <http://dbpedia.org/resource/Marlene_Dietrich> .`

- Between properties: owl:equivalentProperty
- Between classes: owl:equivalentClass
  Example:
  `dbpedia:Person owl:equivalentClass foaf:Person .`

- However, equivalences are often implicitly stated in the data

Inverse Functional Properties

- Possible to define “uniquely identifying properties” useful for object consolidation

- E.g. (hypothetical) from

```
ex:passportNo rdf:type owl:inverseFunctionalProperty .
```

- and

```
dbpedia:Marlene_Dietrich ex:passportNo “12033-89-5” .
freebase:en.marlene_dietrich ex:passportNo ”12033-89-5” .
```

- follows:

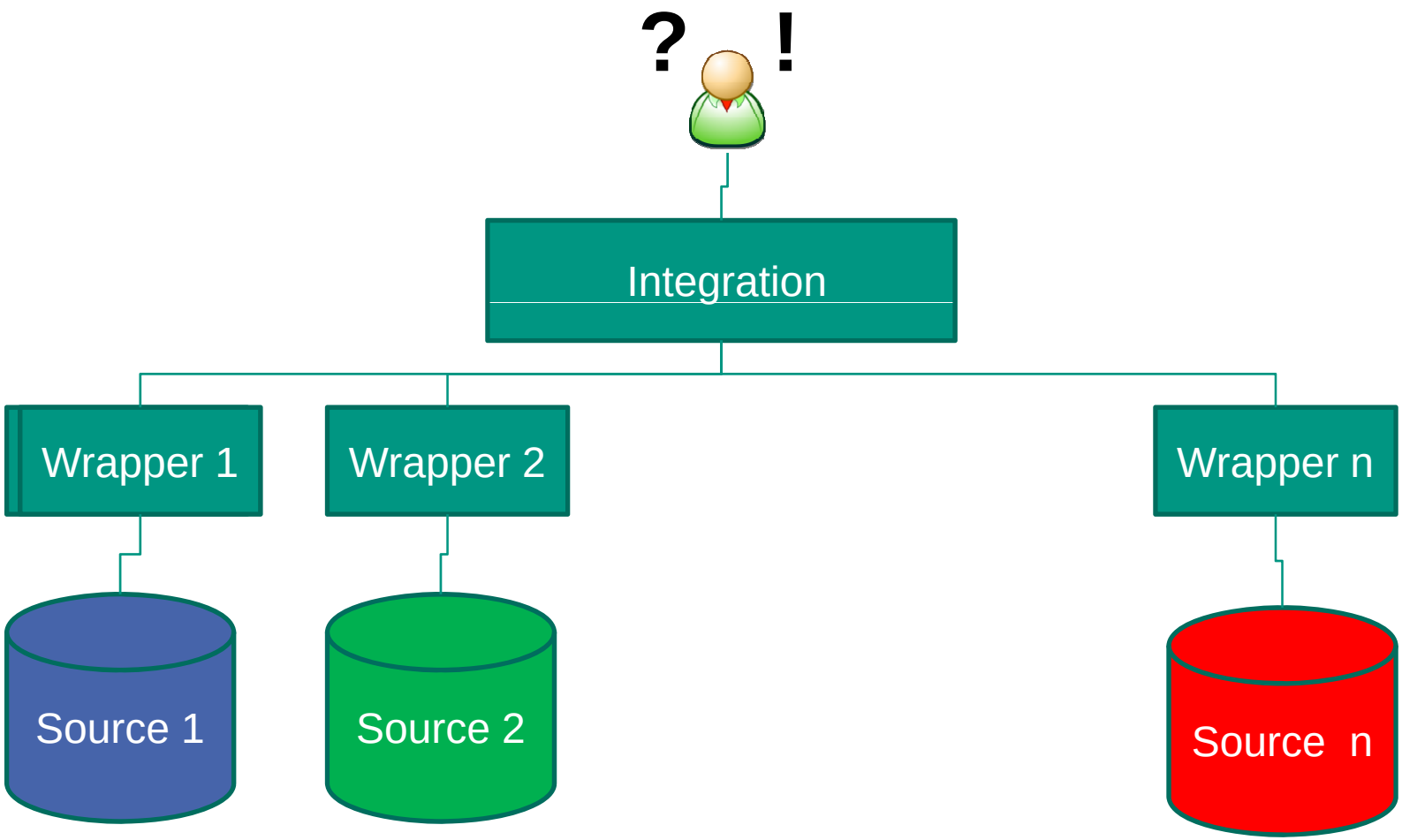
```
dbpedia:Marlene_Dietrich owl:sameAs
freebase:en.marlene_dietrich .
```

Further OWL Primitives

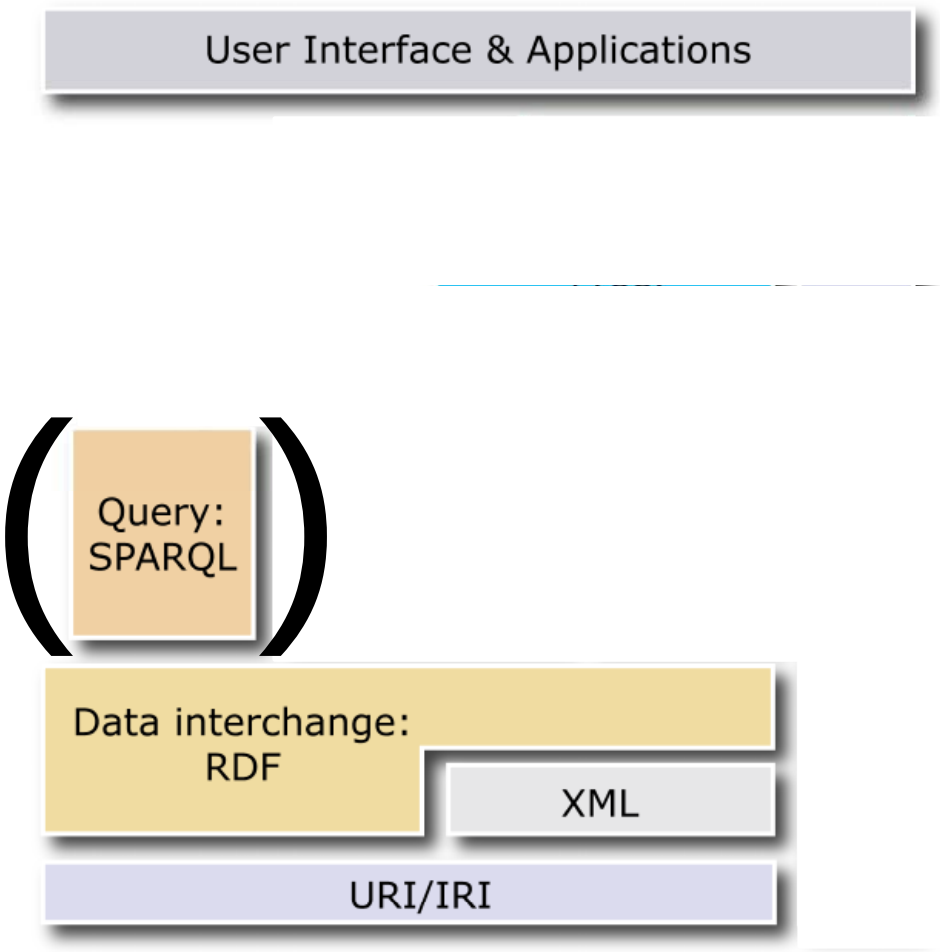
- Property characteristics: inverse properties, symmetric properties
- Property cardinality: minimum cardinality, maximum cardinality
- Class restrictions
- Property chains
- …

LINKED DATA APPLICATION ARCHITECTURES

Data Integration System Architecture



Semantic Web Components



Linked Data: Minimal Components

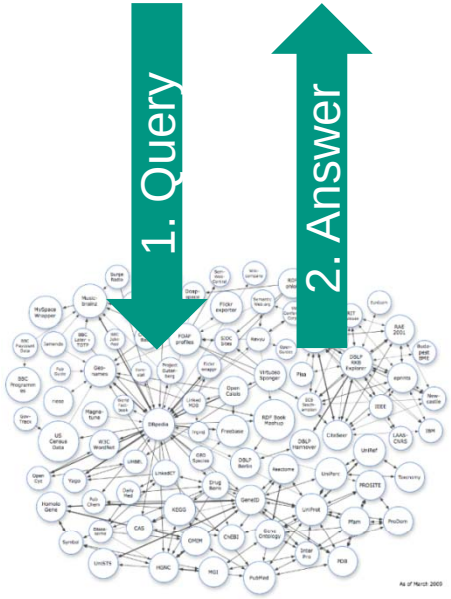
User Interface & Applications

Query:
SPARQL

Data interchange:
RDF

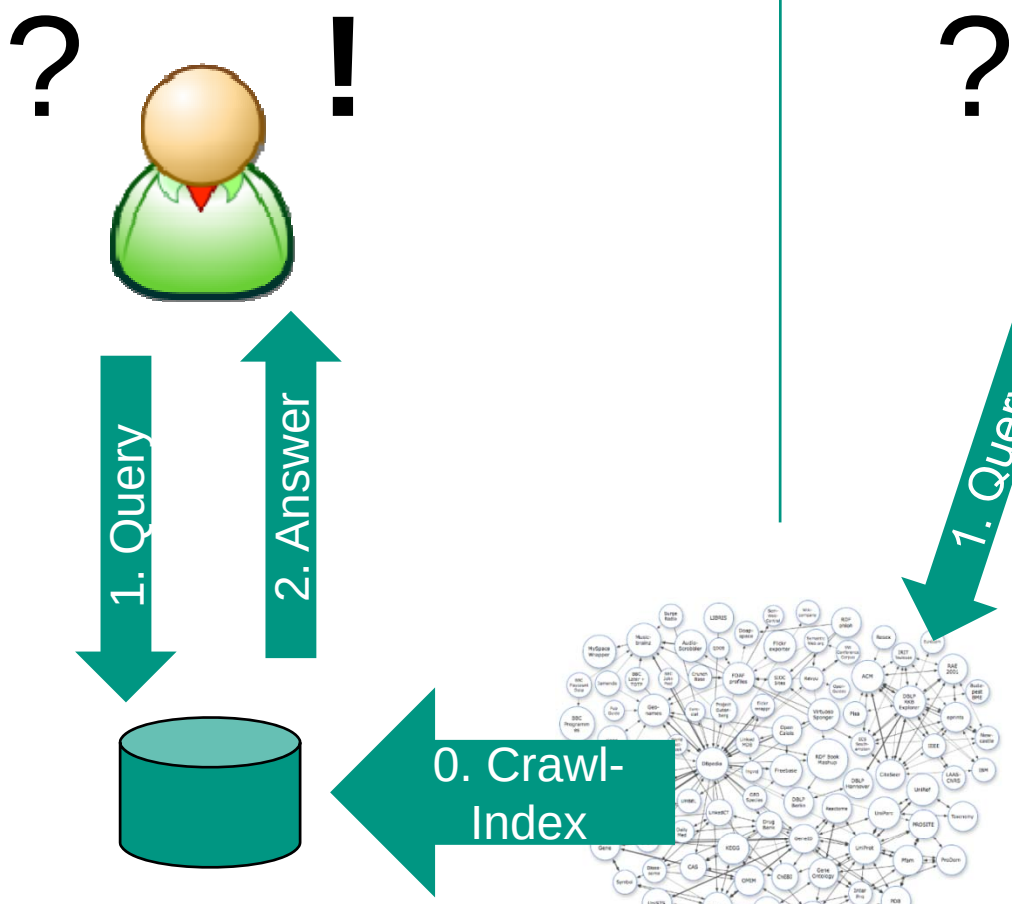
XML

URI/IRI

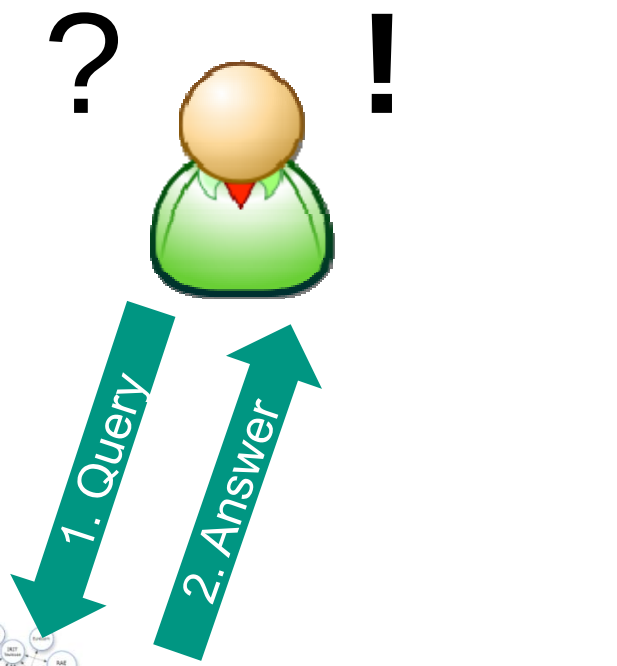


Architecture Styles

Warehousing/ Crawl-Index-Serve

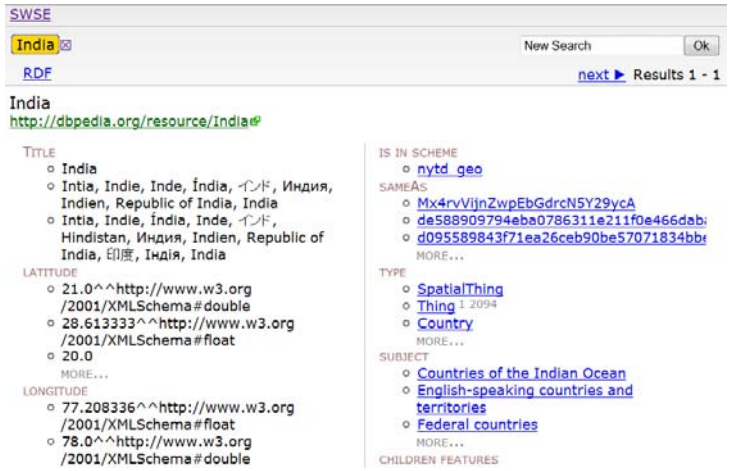


Virtual Integration/ Distributed Querying



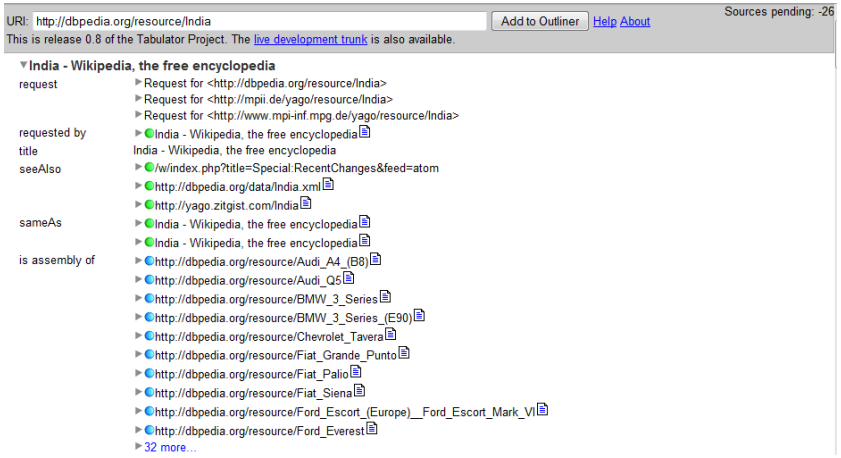
Basic Application: Entity Browsing

Warehousing/ Crawl-Index-Serve



SWSE, Falcons, Sindice, Watson, FactForge…

Virtual Integration/ Distributed Querying



Tabulator, Disco, Zitgist…

SUMMARY

Summary

- The Linked Data Web is a large, decentralised, complex system built on simple principles
  - identify resource via HTTP URIs
  - provide RDF that links to other URIs upon lookup
- Current trend around Linked Data allows for a re-think of components in Semantic Web Layer Cake
- Data publishers and consumers coordinate little
- Web of Data grows rapidly and covers a large variety of domains
- Algorithms operating over a common access protocol and data model
- Ontology languages provide integration and mapping between disparate sources
- First commercial applications emerging

Attribution

- Slides from my SWT-2 lectures and WWW 2010 SILD tutorial
- Slides about RDFS and OWL adapted from SWT-1 lecture (Rudolph, Kroetzsch, Harth)
- Linking Open Data cloud diagrams, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/
- Images of Berlin, Hegel and Dietrich via Wikipedia
- Hendler 97: http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html
- Borst 97: “Construction of Engineering Ontologies”, Ph.D. Thesis, University of Twente 1997.
- Studer, Benjamins, Fensel 98: “Knowledge Engineering: Principles and Methods”, DKE 25(1-2):161-198.
- Gruber 93: “Towards principles for the design of ontologies used for knowledge sharing”, Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer.

Karlsruhe Institute of Technology

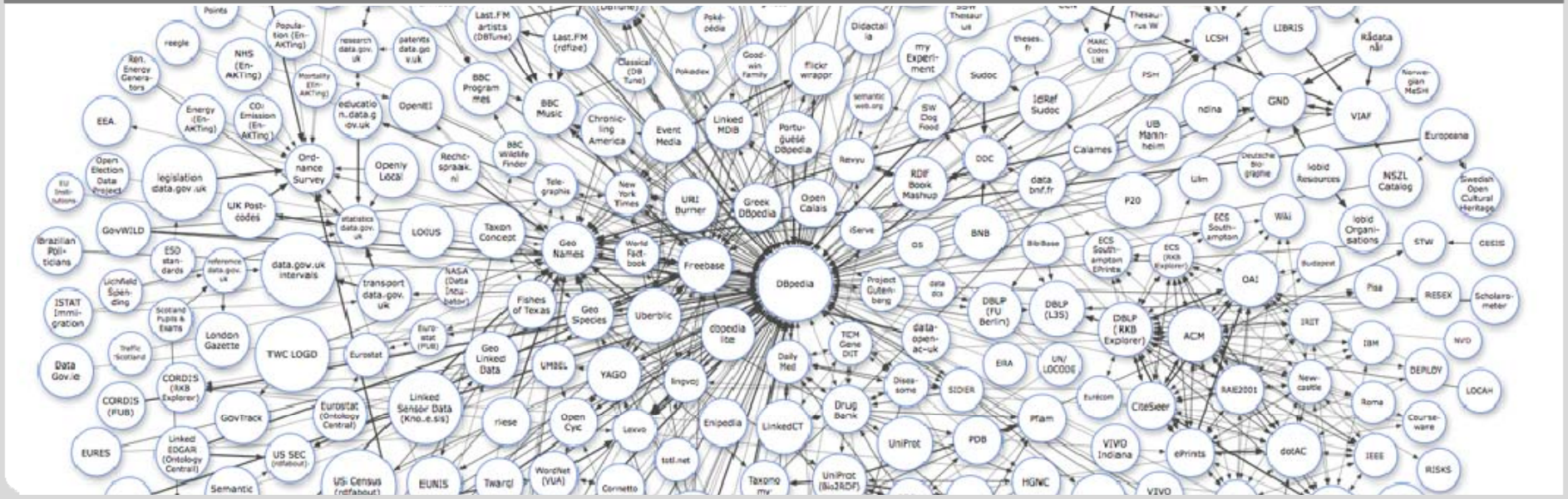
SINTEF

Institut "Jožef Stefan" Ljubljana, Slovenija

Consuming Norwegian Linked Data

Marko Grobelnik, Andreas Harth, Dumitru Roman

Big Linked Data Tutorial Semantic Days 2012



Open Data Movement

- Open Government - establish a modern cooperation among politicians, public administration, industry and private citizens by enabling more transparency, democracy, participation and collaboration
- The Open Government Partnership (launched on September 20, 2011)

- 8 founding governments
- 43 national governments commitments to OG

http://www.opengovpartnership.org/

- Key enablers: free access to information and the possibility to freely use and re-use this information => Open Government Data (OGD)

Open Data Movement (cont')

- "Open Government Data" - worldwide movement to open up government/public administration data
  - Targeted to both human and machine-readable non-proprietary formats - for re-use
- "Open Data"
  - Data beyond just governmental institutions
  - Includes data from relevant stakeholder groups (e.g. citizens, industry, NGOs, science or education, etc.)
- Examples:
  - DIFI (http://data.norge.no/); UN (http://data.worldbank.org); Open Knowledge Foundation (http://okfn.org); New York Times (http://data.nytimes.com)

Open Government Data Principles
(http://sunlightfoundation.com/policy/documents/ten-open-data-principles/)

1. Data must be complete
2. Data must be primary
3. Data must be timely
4. Data must be accessible
5. Data must be machine-processable
6. Access must be non-discriminatory
7. Data formats must be non-proprietary
8. Data must be license-free
9. Data must be permanently available
10. Usage Costs

From Open Data to Linked Open Data

- Crucial for data to be put into a context - new knowledge and more powerful services and applications
  - Interoperability and standards are key

- 5 Stars Model

*	Information is available on the Web (any format) under an open license
**	Information is available as structured data (e.g. Excel instead of an image scan of a table)
***	Non-proprietary formats are used (e.g. CSV instead of Excel)
****	URI identification is used so that people can point at individual data
*****	Data is linked to other data to provide context

What are the costs and benefits of ★ web data?

As a consumer ...	As a publisher ...
✓ You can see it. ✓ You can print it. ✓ You can store it locally (on your hard drive or on an USB stick). ✓ You can enter the data manually into another system.	✓ It is easy to publish.

What are the costs and benefits of ★★ web data?

As a consumer, you can do everything that you could do with ★ web data, plus:	As a publisher ...
✓ You can directly process it with proprietary software to aggregate it, perform calculations, visualise it, etc. ✓ You can export it into another (structured) format.	✓ It is easy to publish.

What are the costs and benefits of ★★★ web data?

As a consumer, you can do everything that you could do with ★★ web data, plus:	As a publisher ...
✓ You do not have to pay for a format over which a single entity has exclusive control	✓ It is easy to publish.

What are the costs and benefits of ★★★★ web data?

As a consumer, you can do everything that you could do with ★★★ web data, plus:	As a publisher ...
✓ You can link to it from any other place, either on the web or locally. ✓ You can bookmark it. ✓ You can re-use parts of the data.	✓ You will need to invest some time slicing and dicing your data. ✓ You will need to assign URIs to data items and think about how to represent the data. ✓ You have fine-granular control over the data items and can optimise their access (e.g. load balancing, caching, etc.)

What are the costs and benefits of ★★★★★ web data?

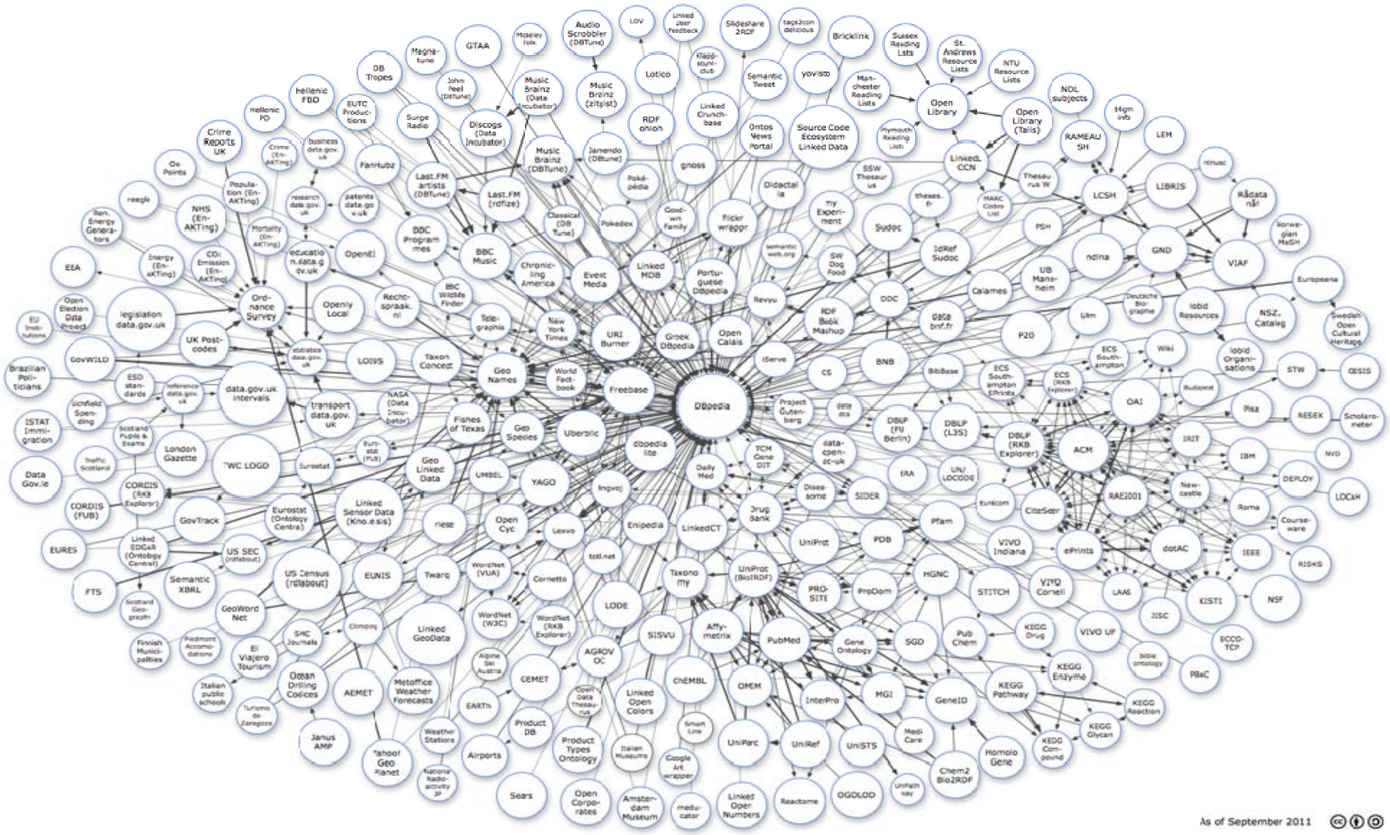
As a consumer, you can do everything that you could do with ★★★★ web data, plus:	As a publisher ...
✓ You can discover new data of interest while consuming other information. ✓ You have access to the data schema.	✓ You will need to invest resources to link your data to other data on the web. ✓ You make your data discoverable. ✓ You increase the value of your data.

Taken from *Linked Open Data: The Essentials* (by E. Bauer, M. Kaltenböck), in turn based on http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

Linked Data

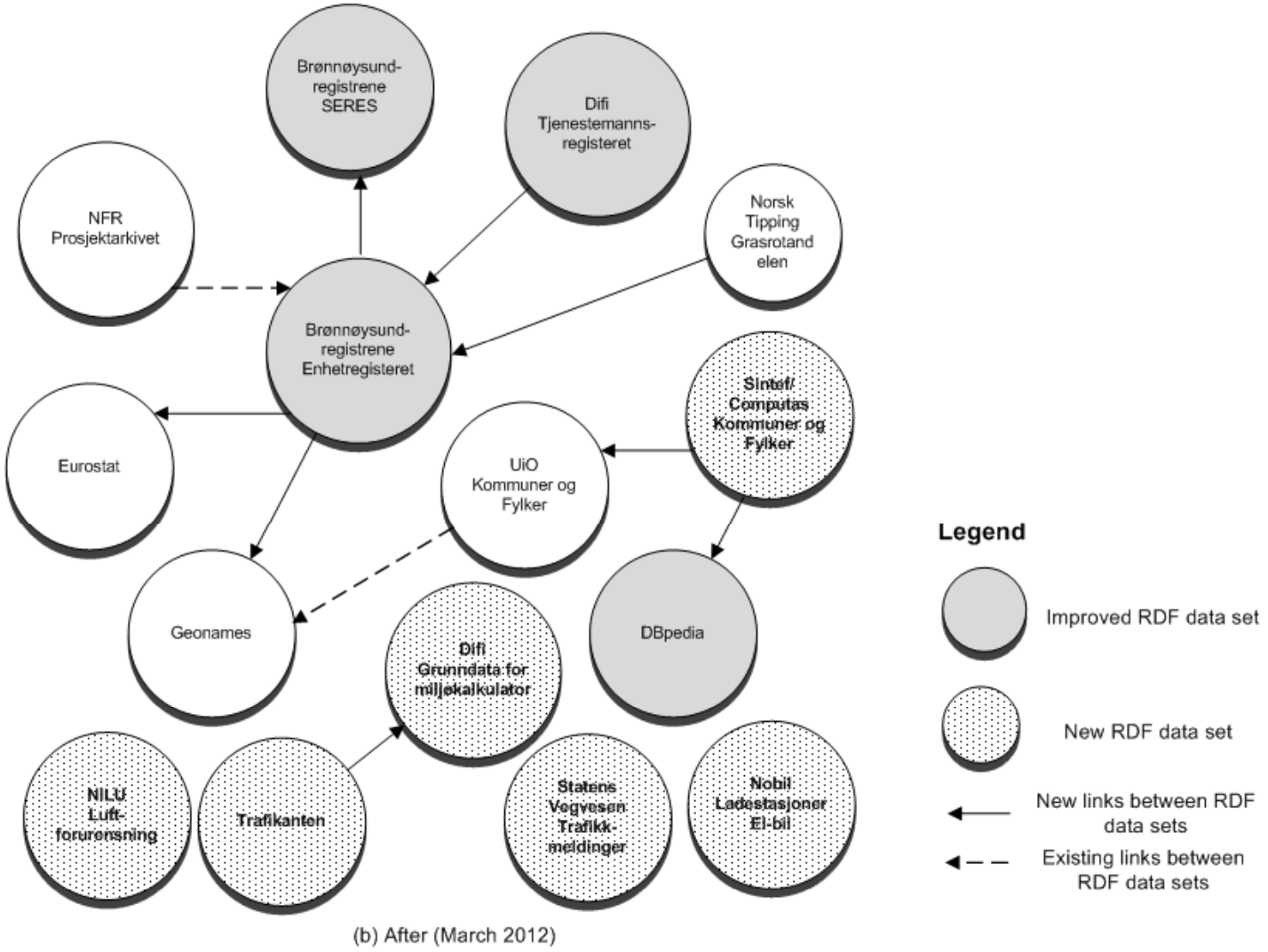
http://linkeddata.org/

- A term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF



Linking Open Data cloud diagram, by Richard Cygania and Anja Jentzsch. http://lod-cloud.net/

Examples of data sets from the Norwegian LOD



(b) After (March 2012)

LOD example – companies data

http://opendata.computas.no/lod/id/enhet/986429360

KARDE AS, en Enhet

Organisasjonsnr: 986429360

Organisasjonsform: Aksjeselskap (AS)

Nacekode:
70.220 - Bedriftsrådgivning og annen administrativ rådgivning

Antall ansatte: 8 (per 15.01.2011)

Adresse:
Skafjellveien 20
3070 SANDE I VESTFOLD
Norge

Kontaktperson:
Terje Johan Grimstad

- andre roller:
  - Daglig leder/ adm.direktør i KARDE AS
  - Styremedlem i KARDE AS
  - Kontaktperson i TGR INVEST AS
  - Styrets leder i TGR INVEST AS
  - Styremedlem i TELLU AS
  - Styremedlem i PALPACOM AS

Styremedlemmer:

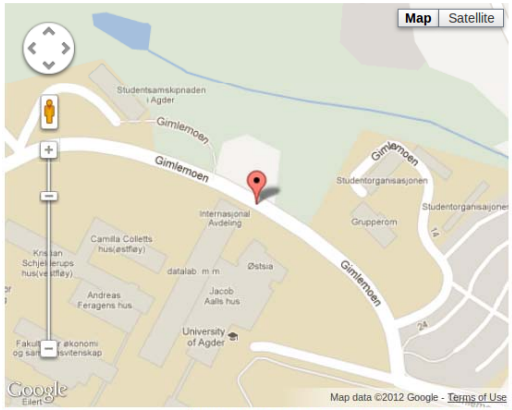
- Styrets leder
  - Arthur Berg Reinertsen
    - andre roller:
      - Daglig leder/ adm.direktør i KARDE AS
      - Styrets leder i KARDE AS
      - Kontaktperson i ABR INVEST AS
      - Styrets leder i ABR INVEST AS
      - Styrets leder i TELLU AS
      - Styrets leder i PALPACOM AS

```
<http://opendata.computas.no/lod/id/enhet/986429360>
      a       <http://opendata.computas.no/dict/guid/Brønnøysundregistrene/Begrep/Enhet/4974> ;
      vocab:adresse
              [ a
<http://opendata.computas.no/dict/guid/Brønnøysundregistrene/Begrep/Adresse/4967> , vcard:Address
;
                vocab:gateadresse "Skafjellveien 20" ;
                vocab:kommunenavn "SANDE" ;
                vocab:kommunenr "0713" ;
                vocab:land
                        [ a       geoinfo:self_governing ,
<http://opendata.computas.no/dict/guid/Brønnøysundregistrene/Begrep/Land/4908> ;
                          rdfs:seeAlso <http://www.geonames.org/countries/NOR/> ;
                          vocab:landNavn "Norge" ;
                          vocab:landkode "NOR" ;
                          geoinfo:codeISO3 "NOR"
                        ] ;
                vocab:postnr "3070" ;
                vocab:poststed "SANDE I VESTFOLD" ;
                hvor:kommunenr "0713" ;
                hvor:postnr "3070" ;
                hvor:poststed "SANDE I VESTFOLD" ;
                vcard:street-address
                        "Skafjellveien 20"
              ] ;
      vocab:ansvarskapital
              [ a       vocab:Ansvarskapital ;
                vocab:kapitalbeløp "100000.0"^^<http://www.w3.org/2001/XMLSchema#double> ;
                vocab:kapitaltype "Aksjekapital"@no ;
                vocab:kapitalvaluta "NOK"
              ] ;
      vocab:antAnsattePåDato
              [ a       vocab:AntAnsattePåDato ;
                vocab:antAnsatte "8"^^<http://www.w3.org/2001/XMLSchema#long> ;
                vocab:gyldigFraDato "2011-01-
14T23:00:00Z"^^<http://www.w3.org/2001/XMLSchema#dateTime>
              ] ;
      vocab:enhetsnavn "KARDE AS" ;
      vocab:erEnhetIRolle _:b22 , _:b1 , _:b13 , _:b14 , _:b24 , _:b32 , _:b23 , _:b28 , _:b6 ,
_:b33 , _:b8 , _:b30 ;
      vocab:nacekode
              [ a       nace:Activity ,
<http://opendata.computas.no/dict/guid/Brønnøysundregistrene/Begrep/Næringskode/4960> ;
                rdfs:seeAlso "http://www4.ssb.no/stabas/ItemsFrames.asp?ID=5552001&Language=nb" ;
                nace:broader <http://ec.europa.eu/eurostat/ramon/rdfdata/nace_r2/70.22> ;
                nace:code "70.220" ;
                nace:name "Bedriftsrådgivning og annen administrativ rådgivning"@no ;
                vocab:nacekodekode "70.220" ;
                vocab:nacekodetekst "Bedriftsrådgivning og annen administrativ rådgivning"@no
              ] ;
```

LOD example – electric car charging stations

http://opendata.computas.no/nobil/id/chargingStation/1121

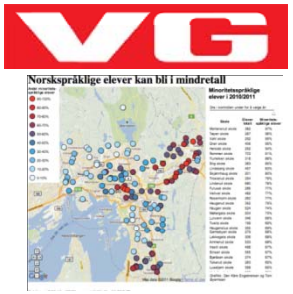
Universitetet i Agder, Kristiansand

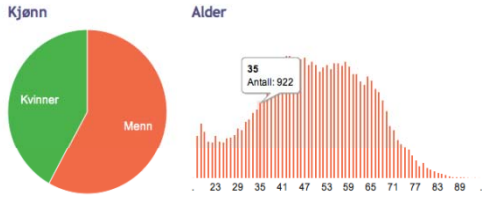


Picture:	http://www.nobil.no/img/ladestasjonbilder/1121.jpg
Thumbnail:	http://www.nobil.no/img/ladestasjonbilder/tn_1121.jpg
Owner:	Universitetet i Agder
Available slots:	2
Place type:	Gateplan
Charging station ID:	1121
Availability:	Besøkende
Contact info:	drift-krs@uia.no
Place description:	Til høyre for hovedbygning. Ingen ordentlige parkeringsplasser, men et skilt og to stikkontakter i en vegg.
Time limit (hours):	0
Street address:	Gimlemoen 25
Postal code:	4604
Municipal number:	1001
Postal area:	KRISTIANSAND S
Latitude:	58.1644
Longitude:	8.0038
Geometry:	POINT(58.1644 8.0038)
Point:	8.0038 58.1644
Charging speed:	16A
Access key:	Åpen
Charging outlets:	2

Example application: Monitoring regional development in Norwegian municipalities

- Data journalism for monitoring regional developments in municipalities in Norway
  - Journalists spend significant time collecting and aggregating data for analyzing trends and monitoring regional developments
  - E.g.
    - How many job openings are there in the municipalities?
    - What's the distribution of gender and age in the political landscape?
    - Where are accidents happening?
- **Problem**: How can we speed up and improve the process of collecting and aggregating data for monitoring regional developments?
- **Added value proposition**: Enable smarter/faster monitoring of regional developments





Example application: Monitoring regional development in Norwegian municipalities (cont')

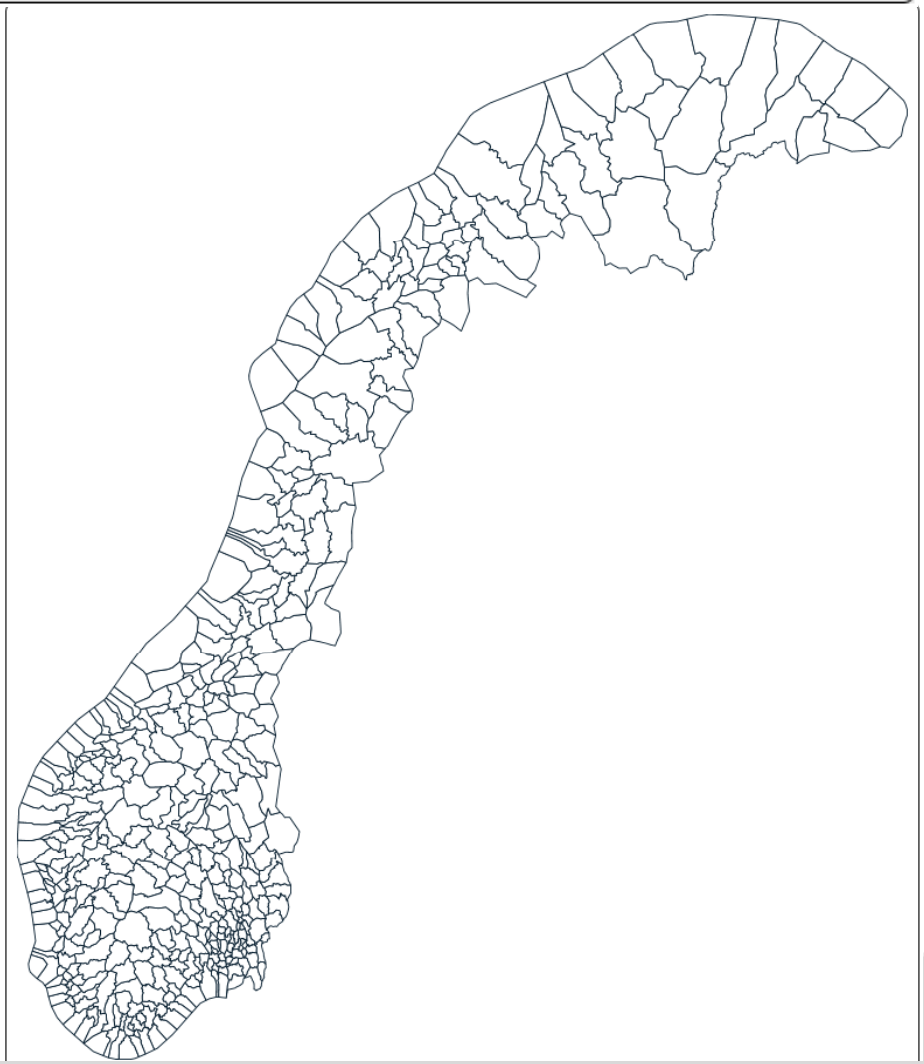
computas

Linked Data Visualizer

SINTEF

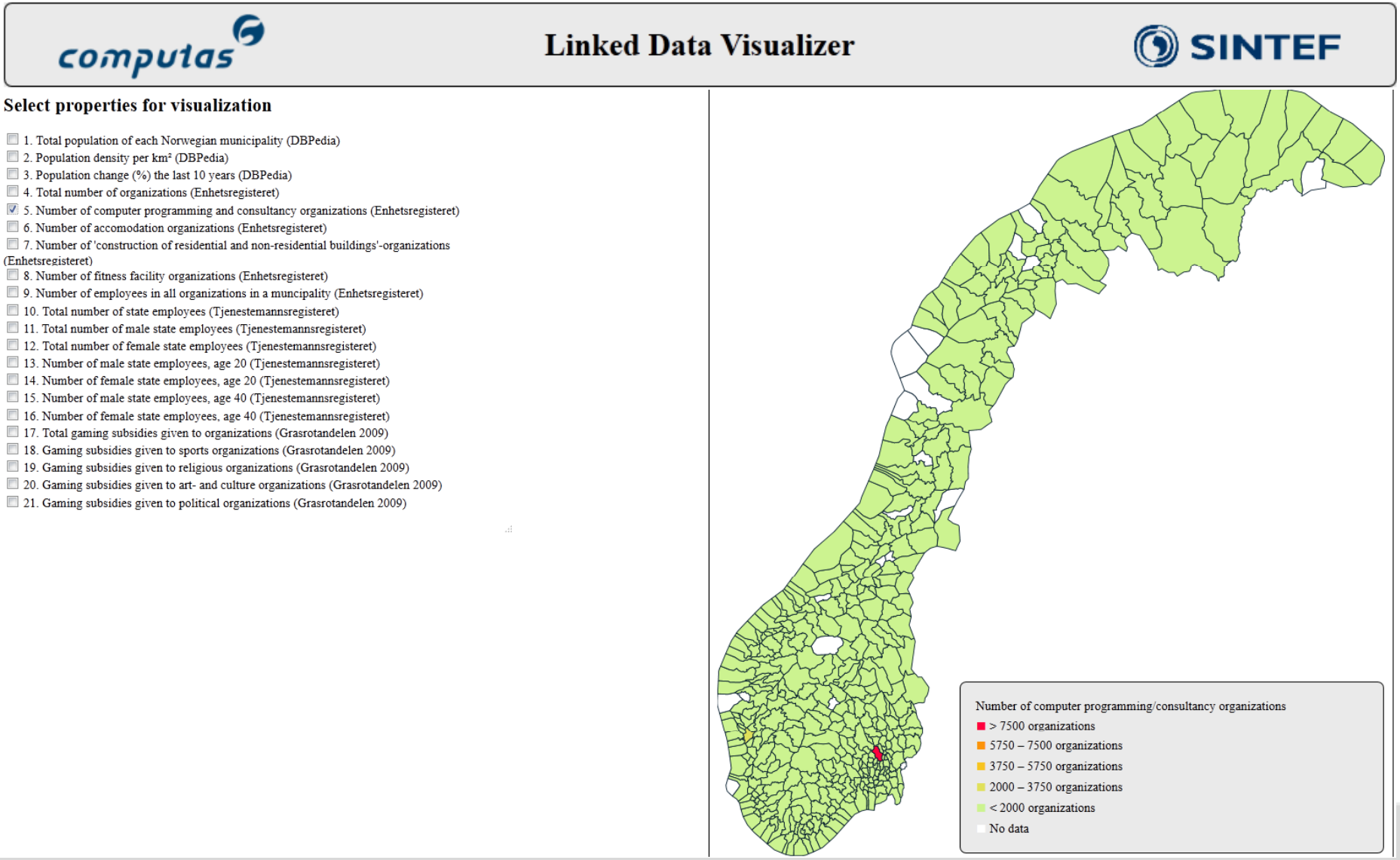
Select properties for visualization

- 1. Total population of each Norwegian municipality (DBPedia)
- 2. Population density per km² (DBPedia)
- 3. Population change (%) the last 10 years (DBPedia)
- 4. Total number of organizations (Enhetsregisteret)
- 5. Number of computer programming and consultancy organizations (Enhetsregisteret)
- 6. Number of accomodation organizations (Enhetsregisteret)
- 7. Number of 'construction of residential and non-residential buildings'-organizations (Enhetsregisteret)
- 8. Number of fitness facility organizations (Enhetsregisteret)
- 9. Number of employees in all organizations in a muncipality (Enhetsregisteret)
- 10. Total number of state employees (Tjenestemannsregisteret)
- 11. Total number of male state employees (Tjenestemannsregisteret)
- 12. Total number of female state employees (Tjenestemannsregisteret)
- 13. Number of male state employees, age 20 (Tjenestemannsregisteret)
- 14. Number of female state employees, age 20 (Tjenestemannsregisteret)
- 15. Number of male state employees, age 40 (Tjenestemannsregisteret)
- 16. Number of female state employees, age 40 (Tjenestemannsregisteret)
- 17. Total gaming subsidies given to organizations (Grasrotandelen 2009)
- 18. Gaming subsidies given to sports organizations (Grasrotandelen 2009)
- 19. Gaming subsidies given to religious organizations (Grasrotandelen 2009)
- 20. Gaming subsidies given to art- and culture organizations (Grasrotandelen 2009)
- 21. Gaming subsidies given to political organizations (Grasrotandelen 2009)

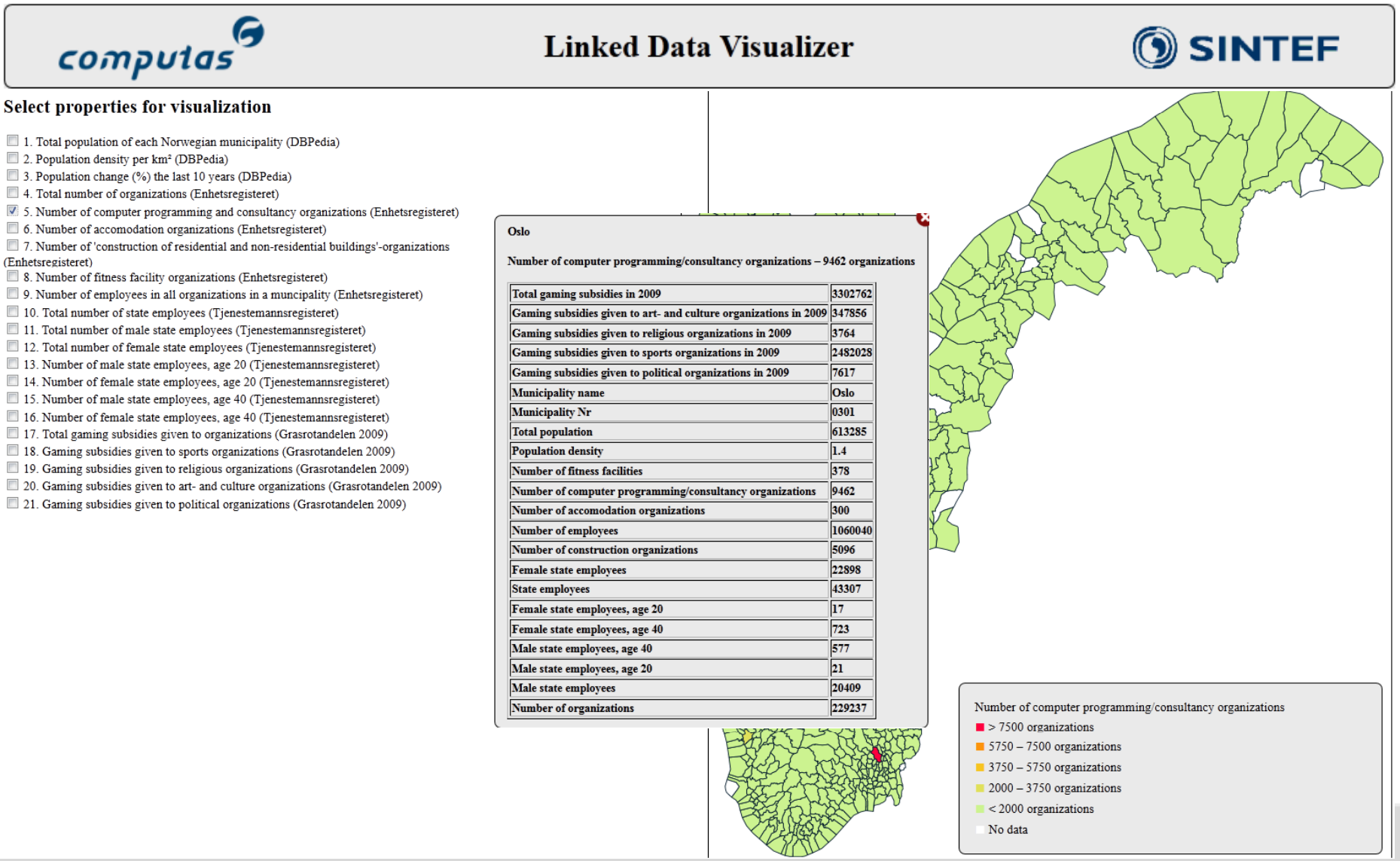


http://opendata.computas.no/RegionalDevelopm

Example application: Monitoring regional development in Norwegian municipalities (cont')



Example application: Monitoring regional development in Norwegian municipalities (cont')



Example application: Monitoring regional development in Norwegian municipalities (cont')

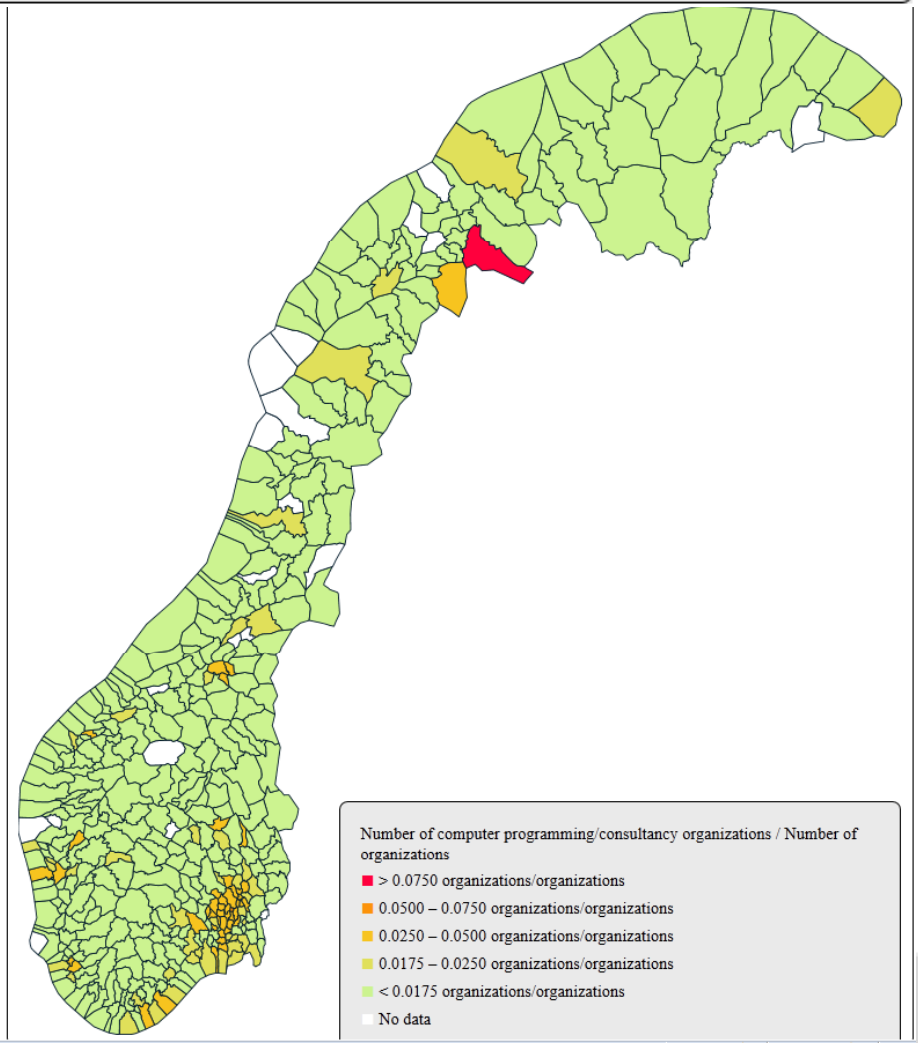
computas

Linked Data Visualizer

SINTEF

Select properties for visualization

- [ ] 1. Total population of each Norwegian municipality (DBPedia)
- [ ] 2. Population density per km² (DBPedia)
- [ ] 3. Population change (%) the last 10 years (DBPedia)
- [x] 4. Total number of organizations (Enhetsregisteret)
- [x] 5. Number of computer programming and consultancy organizations (Enhetsregisteret)
- [ ] 6. Number of accomodation organizations (Enhetsregisteret)
- [ ] 7. Number of 'construction of residential and non-residential buildings'-organizations (Enhetsregisteret)
- [ ] 8. Number of fitness facility organizations (Enhetsregisteret)
- [ ] 9. Number of employees in all organizations in a muncipality (Enhetsregisteret)
- [ ] 10. Total number of state employees (Tjenestemannsregisteret)
- [ ] 11. Total number of male state employees (Tjenestemannsregisteret)
- [ ] 12. Total number of female state employees (Tjenestemannsregisteret)
- [ ] 13. Number of male state employees, age 20 (Tjenestemannsregisteret)
- [ ] 14. Number of female state employees, age 20 (Tjenestemannsregisteret)
- [ ] 15. Number of male state employees, age 40 (Tjenestemannsregisteret)
- [ ] 16. Number of female state employees, age 40 (Tjenestemannsregisteret)
- [ ] 17. Total gaming subsidies given to organizations (Grasrotandelen 2009)
- [ ] 18. Gaming subsidies given to sports organizations (Grasrotandelen 2009)
- [ ] 19. Gaming subsidies given to religious organizations (Grasrotandelen 2009)
- [ ] 20. Gaming subsidies given to art- and culture organizations (Grasrotandelen 2009)
- [ ] 21. Gaming subsidies given to political organizations (Grasrotandelen 2009)

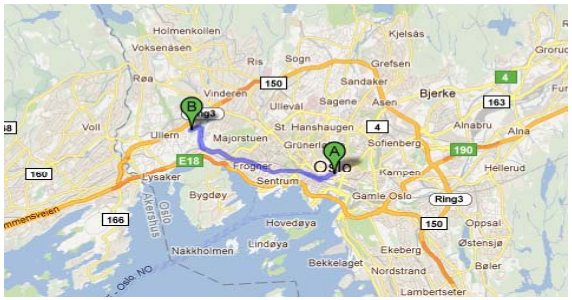


opendata.computas.no/RegionalDevelopment/#

Example application: Decision support for environmentally friendly behaviour

- **Problem statement**: Faced with different transportation options for a short trip, which are the most environmental-friendly options given constraints like time, weather, traffic and private preferences.

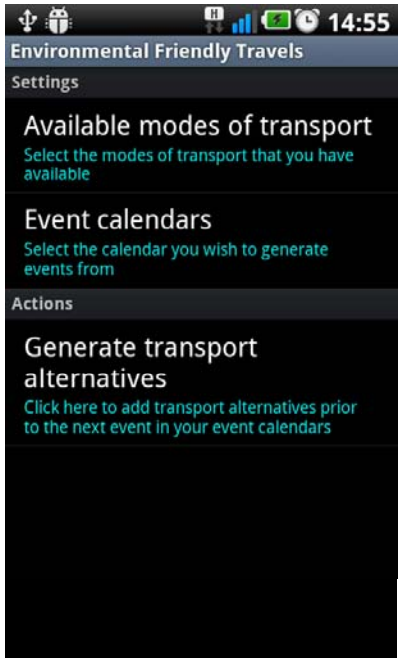
- Typically different options
  - Public transportation (bus/tram/metro/train)
  - Private car (electric/gas/diesel) car, taxi
  - Cycling, walking

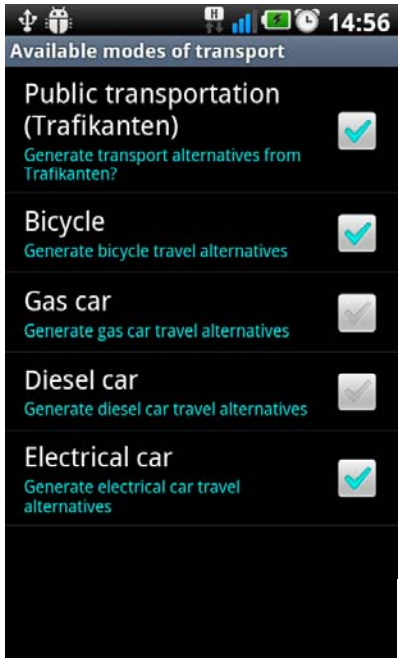


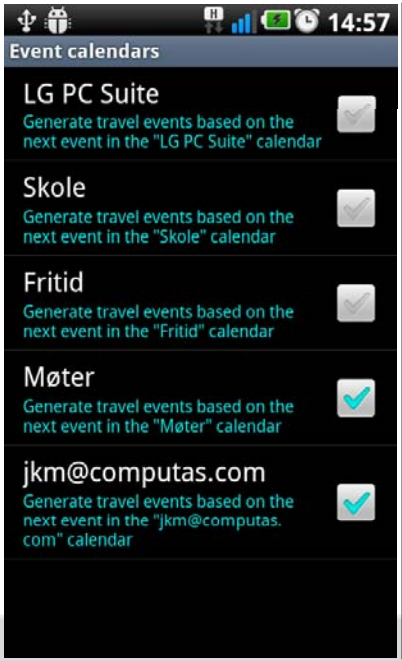
- Constraints: Time, avoid bad weather, polluted zones, traffic, private preferences

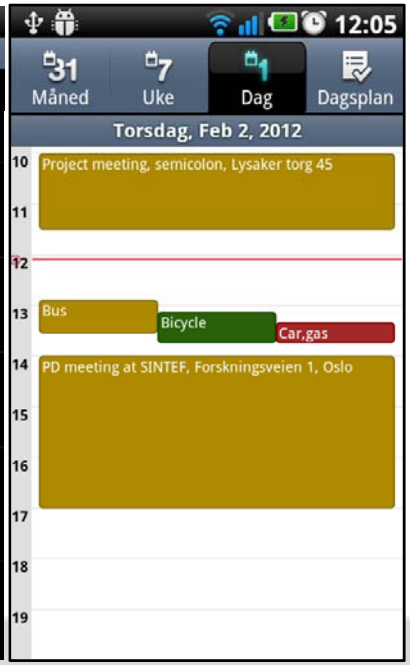
- Environmental parameters: CO2 emissions, energy efficiency

- **Added value proposition**: Enable smarter/faster environmental friendly decision making for local trips when options are available
  - Assist the user's decision making wrt travelling from his current position to the position of the next event in the user's calendar









http://opendata.computas.no/EnvironmentalFriendly/

Karlsruhe Institute of Technology

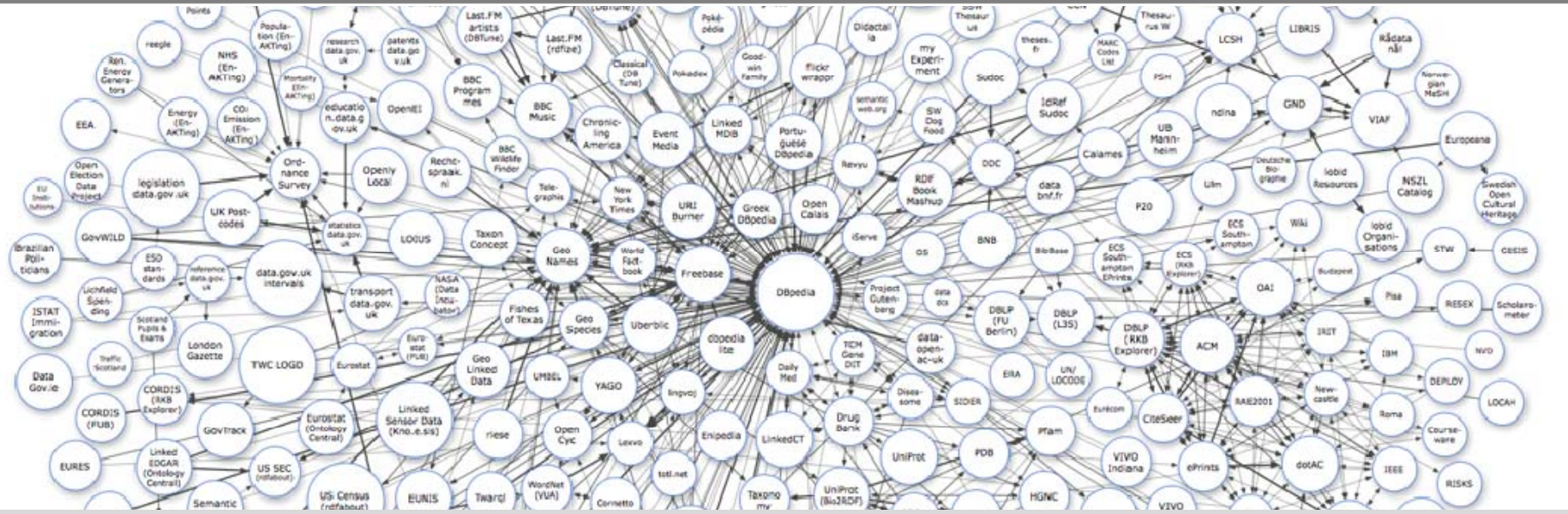
SINTEF

Institut "Jožef Stefan" Ljubljana, Slovenija

Large-scale Linked Data Management

Marko Grobelnik, Andreas Harth (Günter Ladwig), Dumitru Roman

Big Linked Data Tutorial Semantic Days 2012

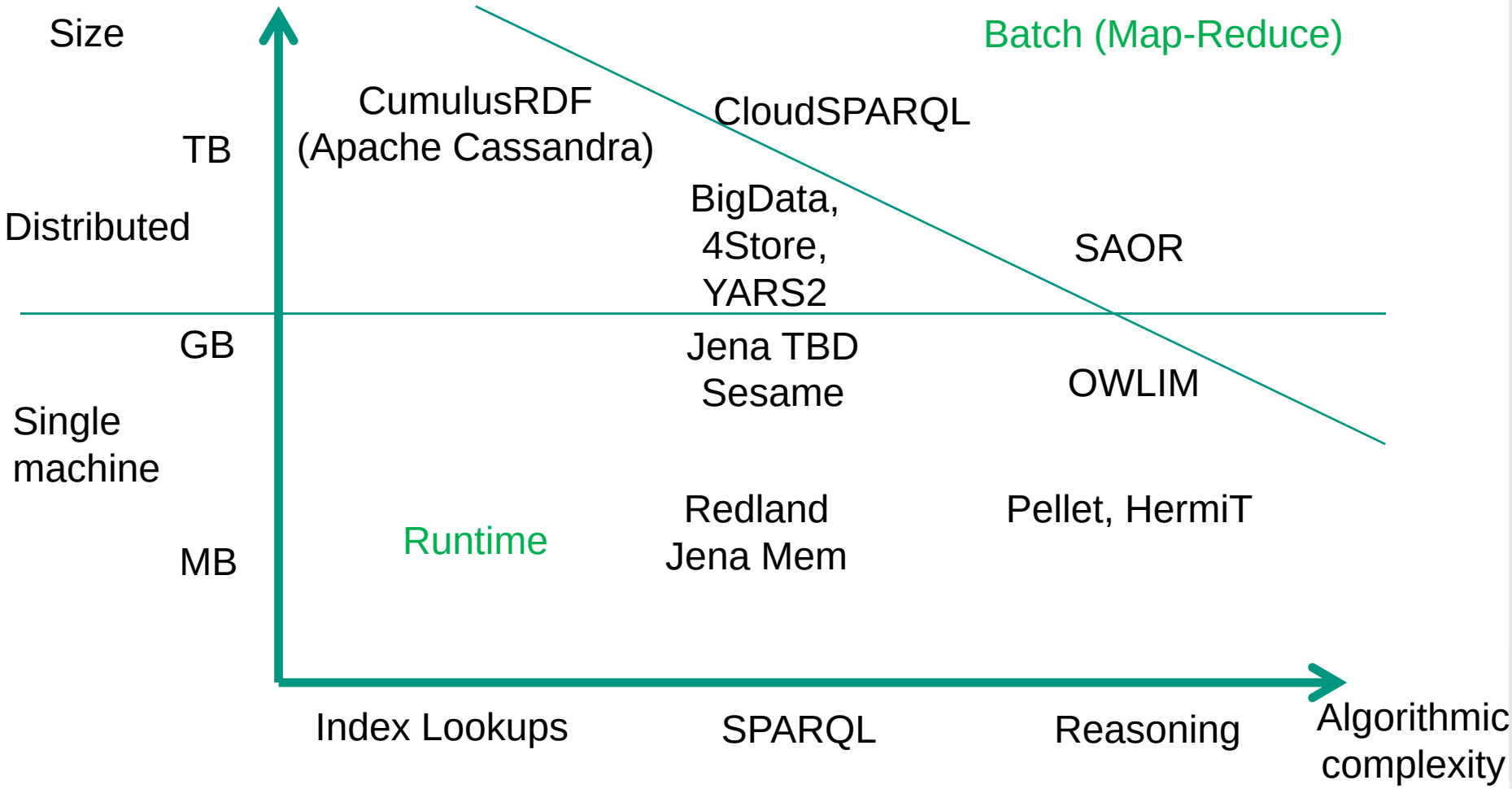


Large-Scale Linked Data Management (Andreas)

- Motivation
- Preliminaries
  - Apache Cassandra
  - CumulusRDF
- Storage Layouts
  - Storage Model
  - Hierarchical Layout
  - Flat Layout
- Evaluation
- Conclusion

MOTIVATION

Linked Data Storage and Retrieval at Scale



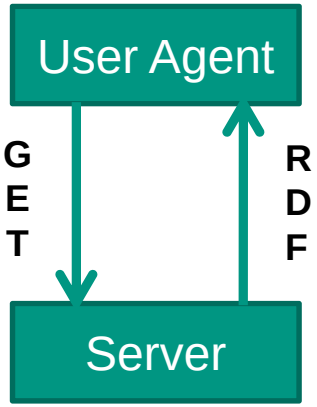
Linked Data Management

- RDF data accessible via HTTP lookups
- Many datasets cover descriptions of millions of entities
- Publishers often use full-fledged triple stores
  - Complex query processing capabilities not necessary for Linked Data lookups
- Trend towards specialized data management systems **tailored for specific use cases**
- Distributed key-value stores
  - Simple (often nested) data model
  - No (expensive) joins
  - High availability and scalability
- We investigate applicability of key-value stores for managing and publishing Linked Data

Linked Data Lookups

- Dereferencing URI *t* should return RDF graph describing *t*
  - Exact content is only lightly specified
- Common practice (e.g. DBpedia) is to return
  - all triples with the given URI as subject and
  - some triples with the given URI as object
- Other options
  - Only triples with the given URI as subject
  - Concise Bounded Descriptions

http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5#artist



http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5.rdf

Triple Patterns

- A triple pattern is an RDF triple that may contain variables instead of RDF terms in any position

```
?s dbpprop:birthPlace dbpedia:Karlsruhe .
```

or

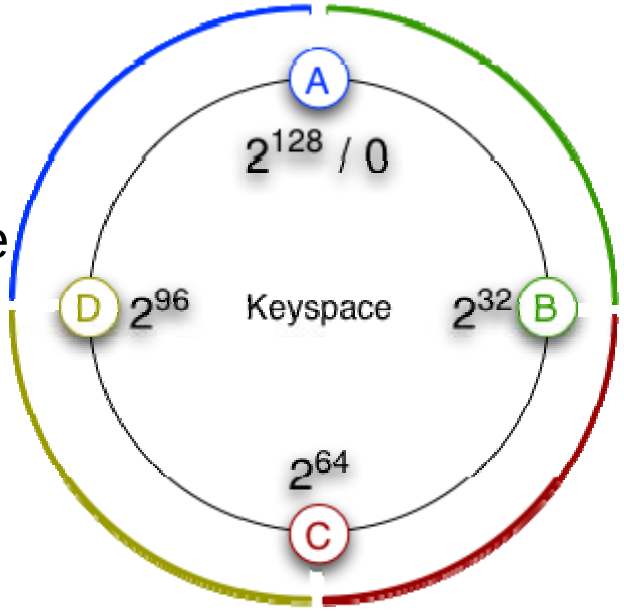
```
?s foaf:name ?o .
```

- Linked Data Lookup on *t* translates into two triple patterns lookups
  - (t ? ?)
  - (? ? t)
- At least three indexes to cover all possible triple patterns (with prefix lookups)

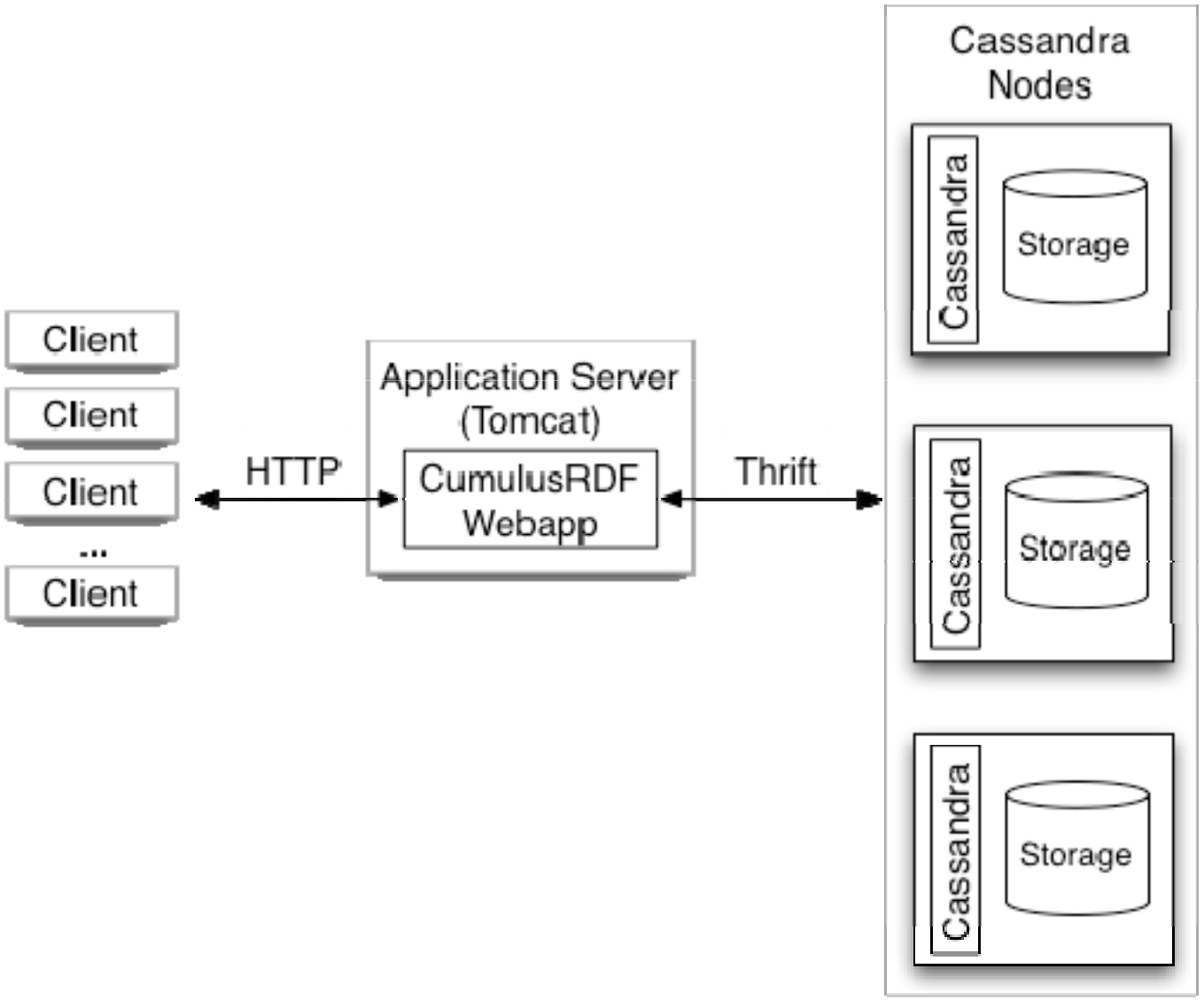
Patterns	Index
? ? ?	Any
s ? ?	SPO
? p ?	POS
? ? o	OSP
s p ?	SPO
? p o	POS
s ? o	OSP
s p o	Any

Apache Cassandra

- Open source data management system
- Distributed key-value store (DHT-based)
  - Nested key-value data model
  - Schema-less
- Decentralized
  - Every node in the cluster has the same role
  - No single point of failure
- Elastic
  - Throughput increases linearly as machines are added with no downtime
- Fault-tolerant
  - Data can be replicated



CumulusRDF


Cassandra Nodes
Cassandra
Storage
Cassandra
Storage
Cassandra
Storage
Client
Client
Client
...
Client
HTTP
Application Server (Tomcat)
CumulusRDF Webapp
Thrift

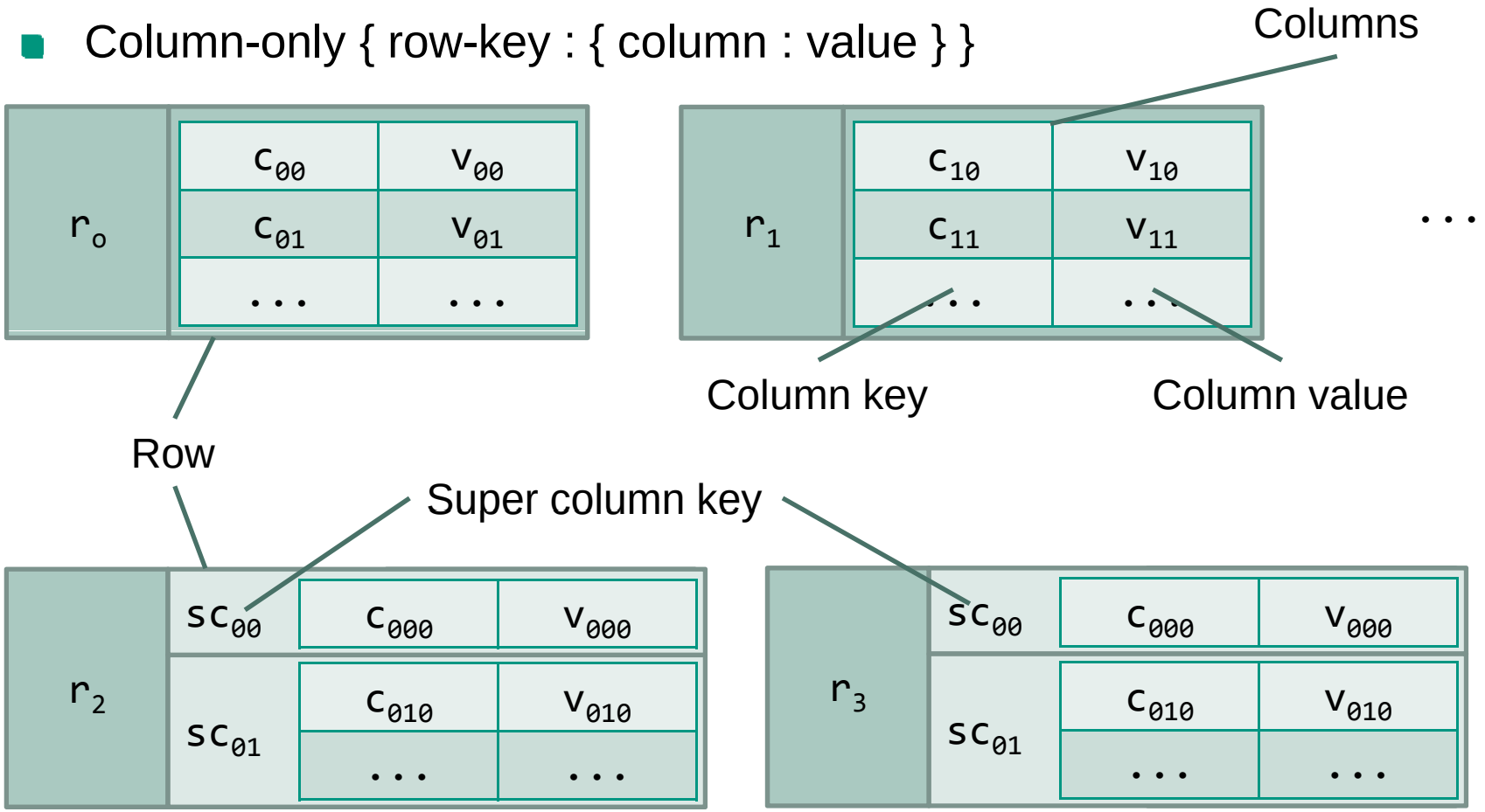
CumulusRDF Functionality

- Distributed deployment to enable scale (more data and also more clients) by adding more machines (via Cassandra)
- Geographical replication (via Cassandra)
- Write-optimised indices with eventual consistency (via Cassandra)
- Triple pattern lookups (via CumulusRDF index structures)
- Linked Data Lookups (via CumulusRDF index structures)

STORAGE LAYOUTS

Nested Key-Value Storage Model

- Column-only { row-key : { column : value } }



- Super columns { row-key : { supercolumn : { column : value } } }

Nested Key-Value Storage Model

- Secondary indexes map column values to rows
  - { value : row-key }

- Cassandra limitations
  - Entire rows always stored on a single node
  - No range queries on row keys
  - Columns are stored in specified order and allow for range queries

Hierarchical Layout

- Uses super columns
- RDF terms occupy row, supercolumn and column positions
  - Value is empty
- Three indexes SPO, POS, OSP cover all possible triple pattern
- Example: SPO index
  - SPO: `{ s : { p : { o : - } } }`

Row key	Super column key	Column key	Value
dbp:Jaws	foaf:name	"Jaws"	-
	rdf:type	dbp:Film	-
		dbp:Work	-

Flat Layout

- Uses columns only
  - Range queries on column keys allow **prefix lookups**
- Concatenate second & third position to form column key
  - SPO { s : { po : - } }
  - *po* is the concatenation of predicate and object
  - For (*sp?*) we perform a prefix lookup on *p* in row with key *s*

Row key	Column key	Value
dbp:Jaws	foaf:name “Jaws”	-
	rdf:type dbp:Film	-
	rdf:type dbp:Work	-

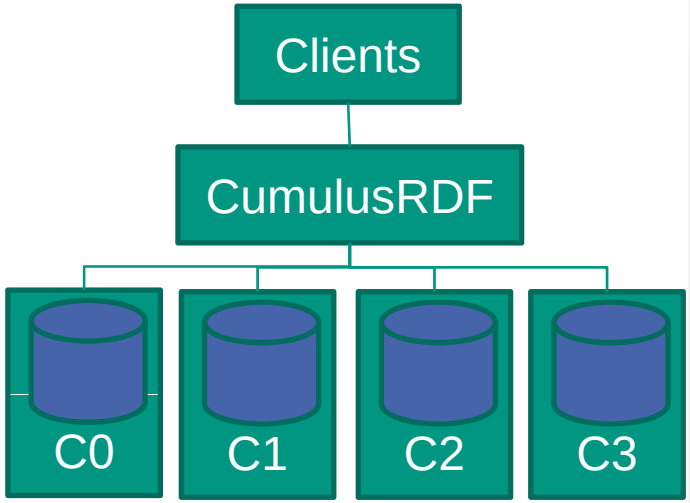
POS Index

- RDF data is skewed: many triples may share the same predicate (*rdf:type* is a prime example)
  - *p* as row key will result in a very uneven distribution
  - Cassandra cannot split rows among several nodes
- We take advantage of Cassandra's secondary indexes
- Use *po* as row key
  - { *po* : { *s* : - } }
  - Smaller rows, better distribution
  - No range queries on rows key: no prefix lookup!
- In each row we add a special column 'p' which has *p* as its value
  - { *po* : { 'p' : *p* } }
- Secondary index on column 'p' allows retrieval of all *po* row keys for a given *p*

EVALUATION

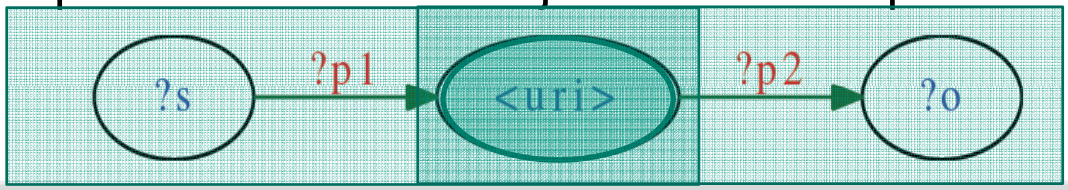
Evaluation

- System: 4 node cluster on virtualized infrastructure
  - 2 CPUs, 4GB RAM, 80GB disk per node
- Dataset: DBpedia 3.6 subset
  - 120M triples (all w/o multilingual labels)
- Triple pattern queries
  - 1M sampled S, SP, SPO, SO, and O patterns from dataset
  - Output: all matching triples
- Linked Data lookup queries
  - 2M resource lookups from DBpedia logs (1.2M unique)
  - Output: all triples with URI as subject and 10k triples with URI as object



C0-C3: Cassandra nodes

10k



all

Results – Storage Layout

Index	Node 1	Node 2	Node 3	Node 4	Std. Dev.	Max. Row
SPO Hier	4.41	4.40	4.41	4.41	0.01	0.0002
SPO Flat	4.36	4.36	4.36	4.36	0.00	0.0004
OSP Hier	5.86	6.00	5.75	6.96	0.56	1.16
OSP Flat	5.66	5.77	5.54	6.61	0.49	0.96
POS Hier	4.43	3.68	4.69	1.08	1.65	2.40
POS Sec	7.35	7.43	7.38	8.05	0.33	0.56

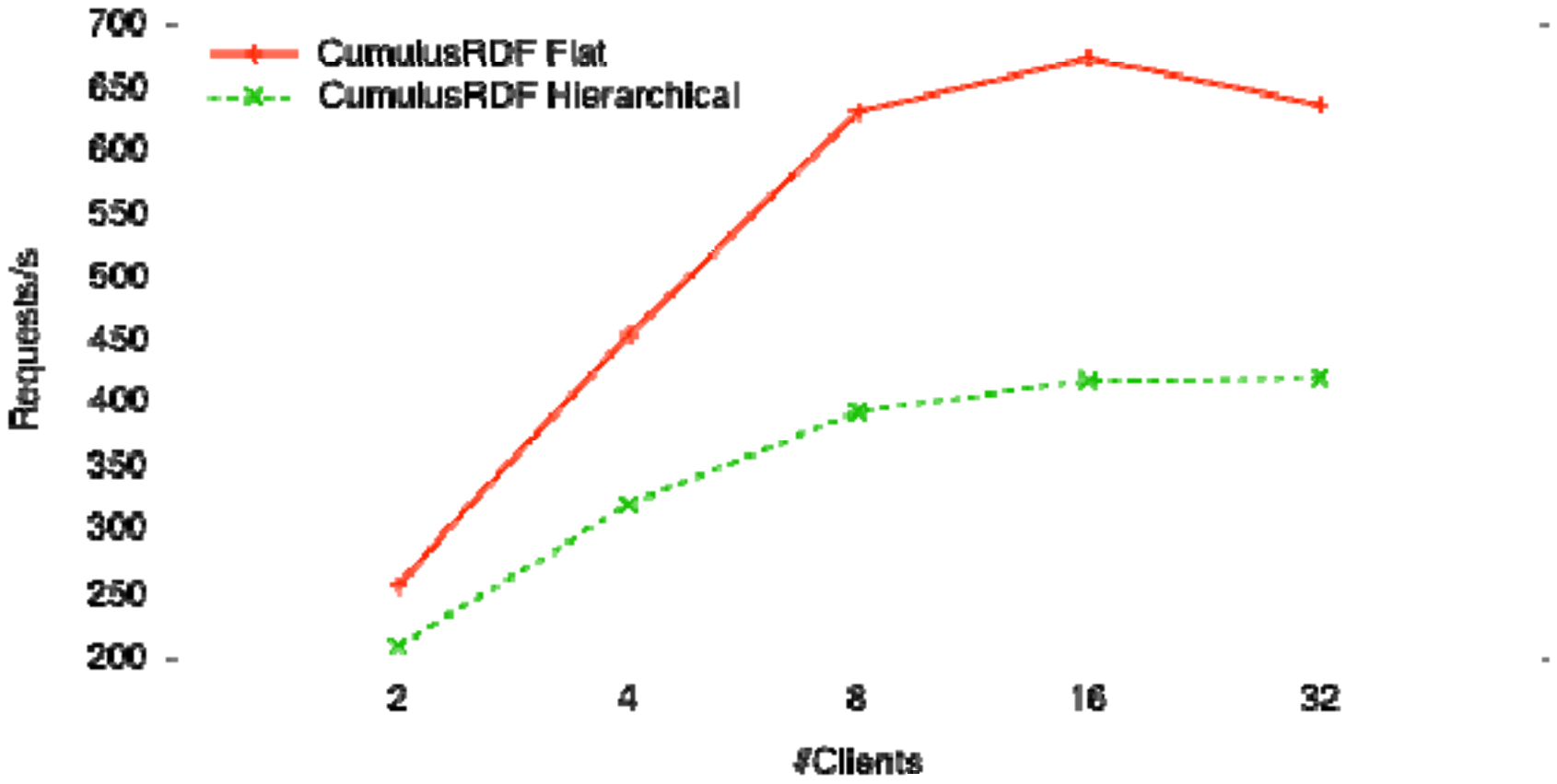
Values in GB

SPO Flat: { s : { po : - } }, OSP

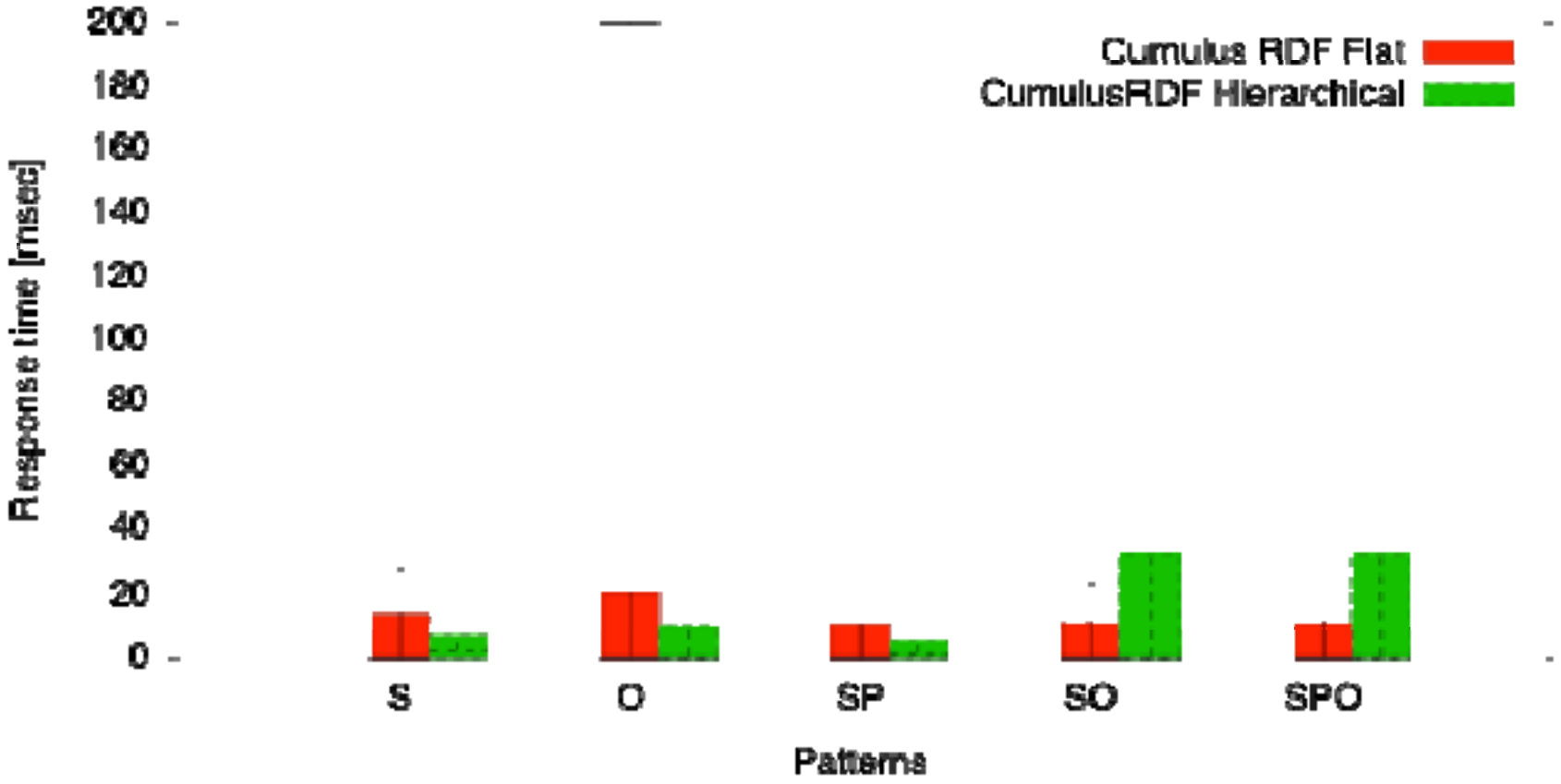
POS Sec: { *po* : { ‘p’ : *p* } }

SPO Hier: { s : { p : { o : - } } }, OSP, POS

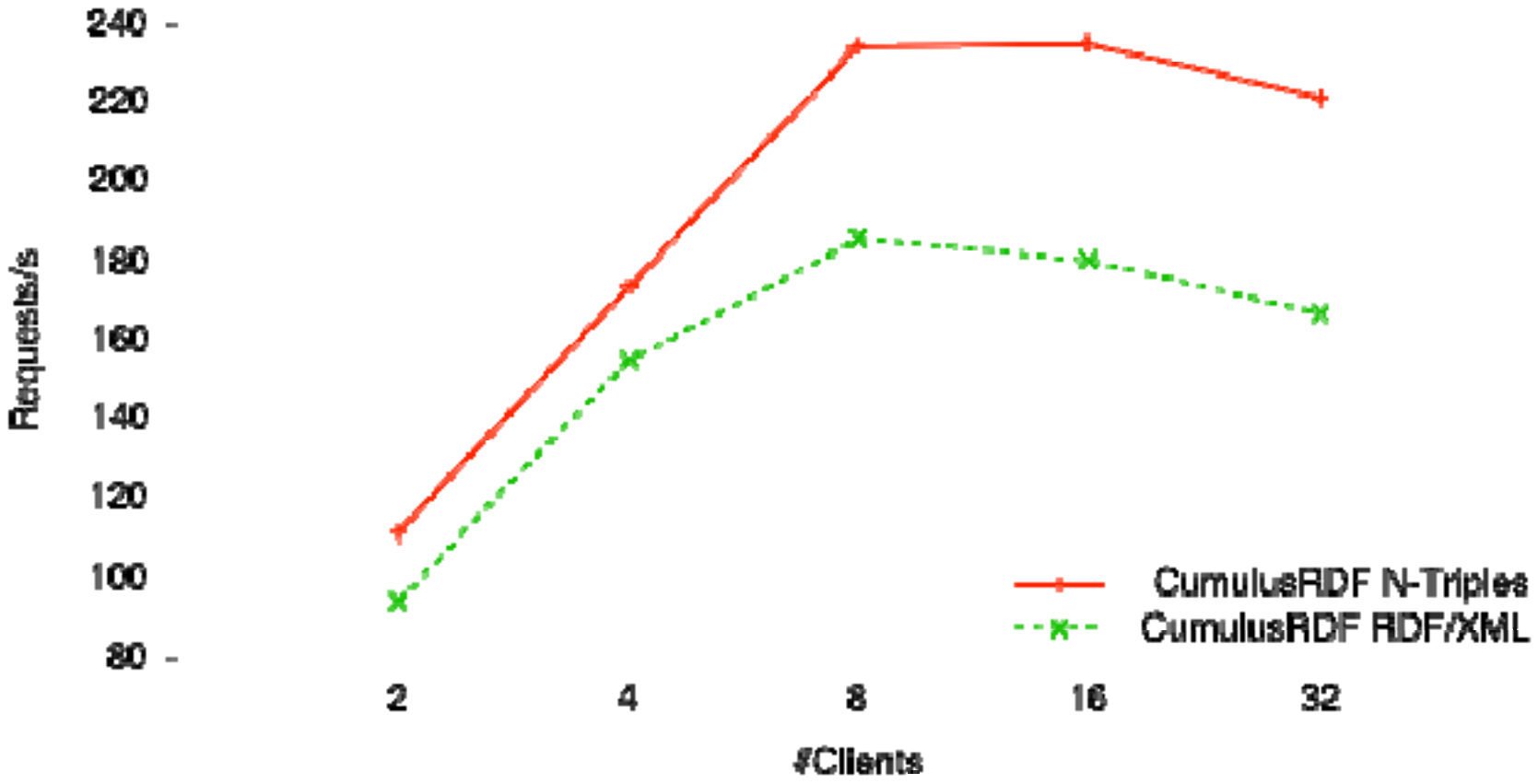
Results – Pattern Lookups



Results – Pattern Lookups



Results – Linked Data Lookups



Conclusion

- We evaluated two index schemes for RDF on nested key-value stores to support Linked Data lookups
  - Flat indexing gives best overall results
  - Output format impacts performance (N-Triples v RDF/XML)
- Apache Cassandra is a viable alternative to full-fledged triple stores for Linked Data lookups
- Future work
  - Automatic generation and maintenance of dataset statistics
  - Evaluate insert and update performance

- Get CumulusRDF at http://code.google.com/p/cumulusrdf

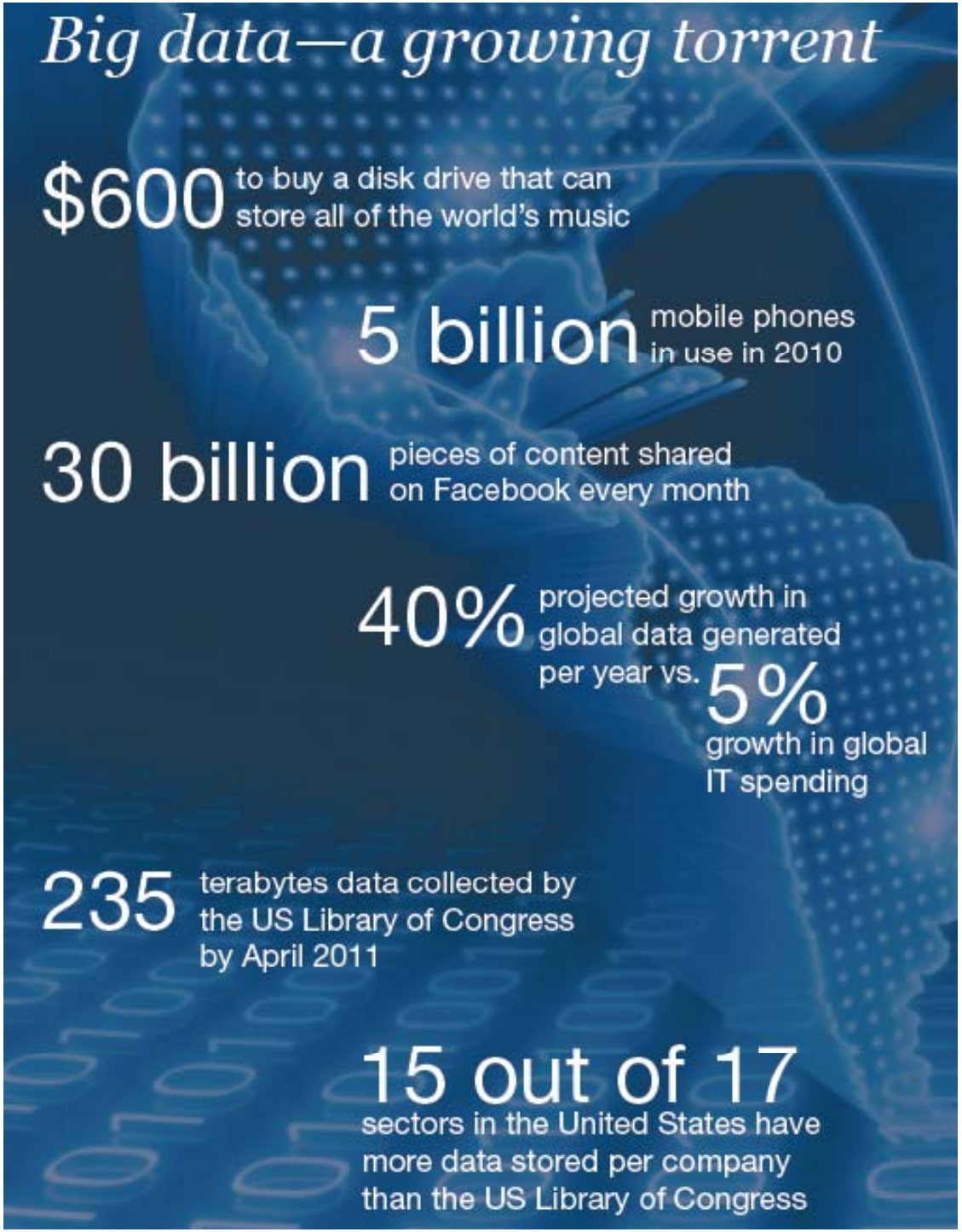
Big-Data Tutorial

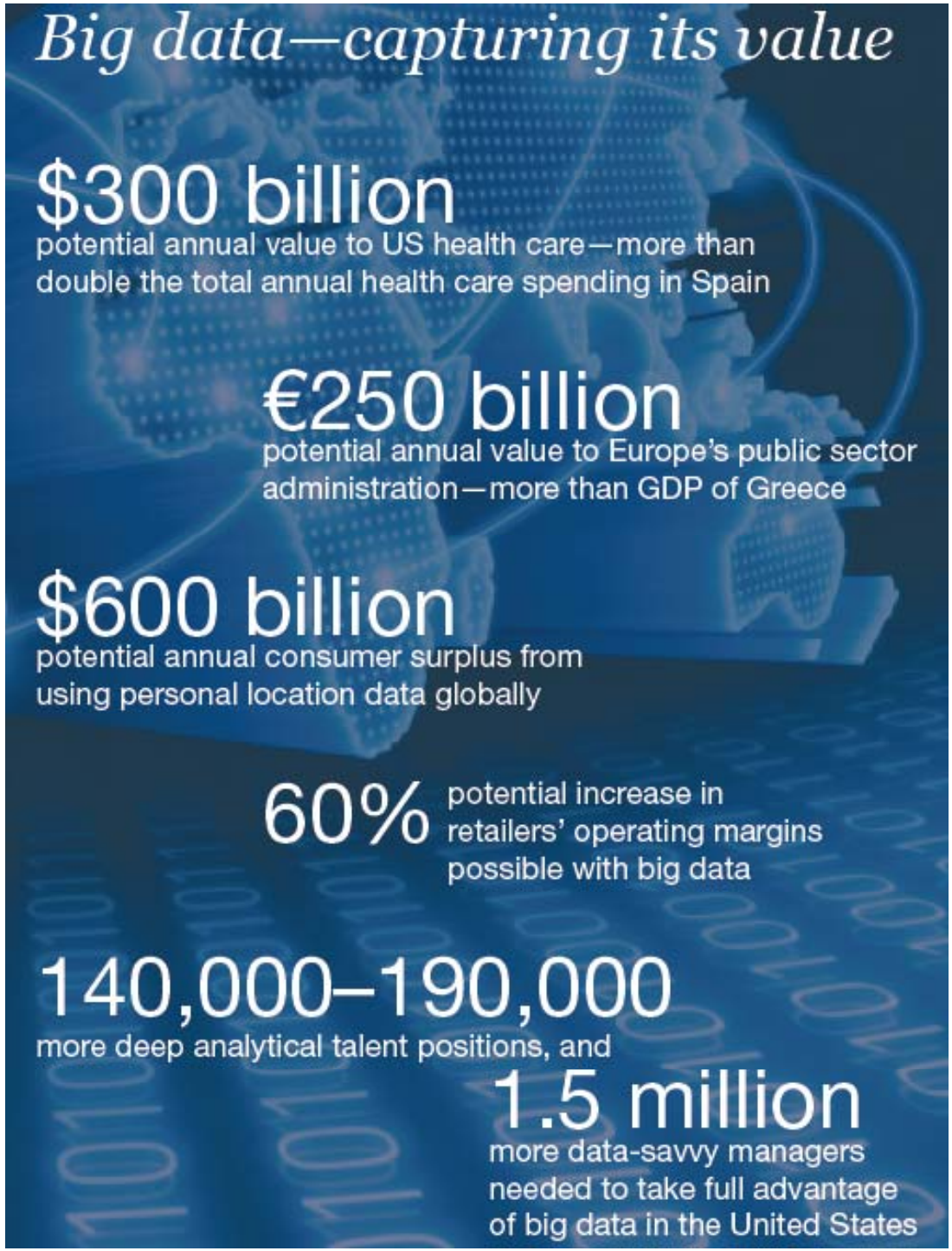
Marko Grobelnik
marko.grobelnik@ijs.si
Jozef Stefan Institute
Ljubljana, Slovenia

Stavanger, May 8th 2012

Outline

- Introduction
  - What is Big data?
  - Why Big-Data?
  - When Big-Data is really a problem?
- Techniques
- Tools
- Applications
- Literature

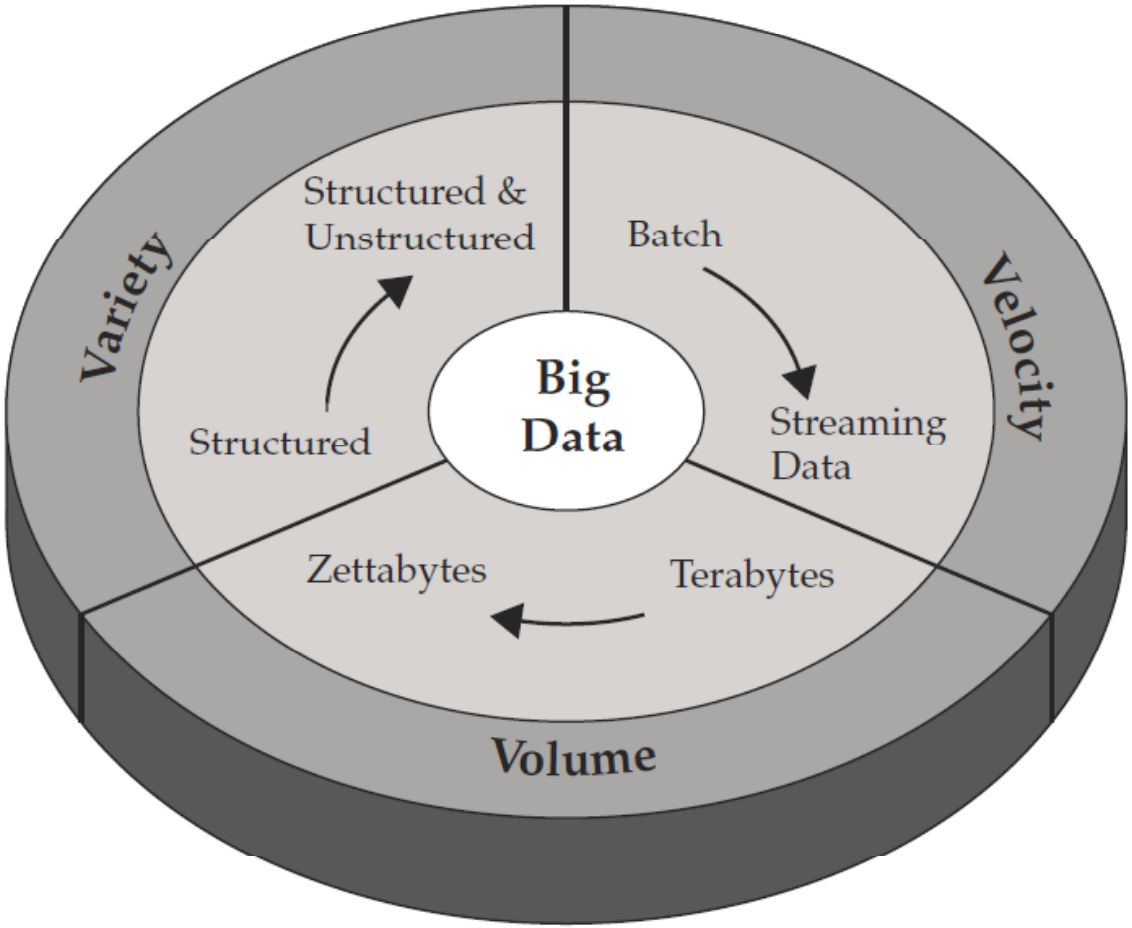

Big data—a growing torrent
$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated per year vs. 5% growth in global IT spending
235 terabytes data collected by the US Library of Congress by April 2011
15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress


Big data—capturing its value
$300 billion
potential annual value to US health care—more than double the total annual health care spending in Spain
€250 billion
potential annual value to Europe's public sector administration—more than GDP of Greece
$600 billion
potential annual consumer surplus from using personal location data globally
60%
potential increase in retailers' operating margins possible with big data
140,000–190,000
more deep analytical talent positions, and
1.5 million
more data-savvy managers needed to take full advantage of big data in the United States

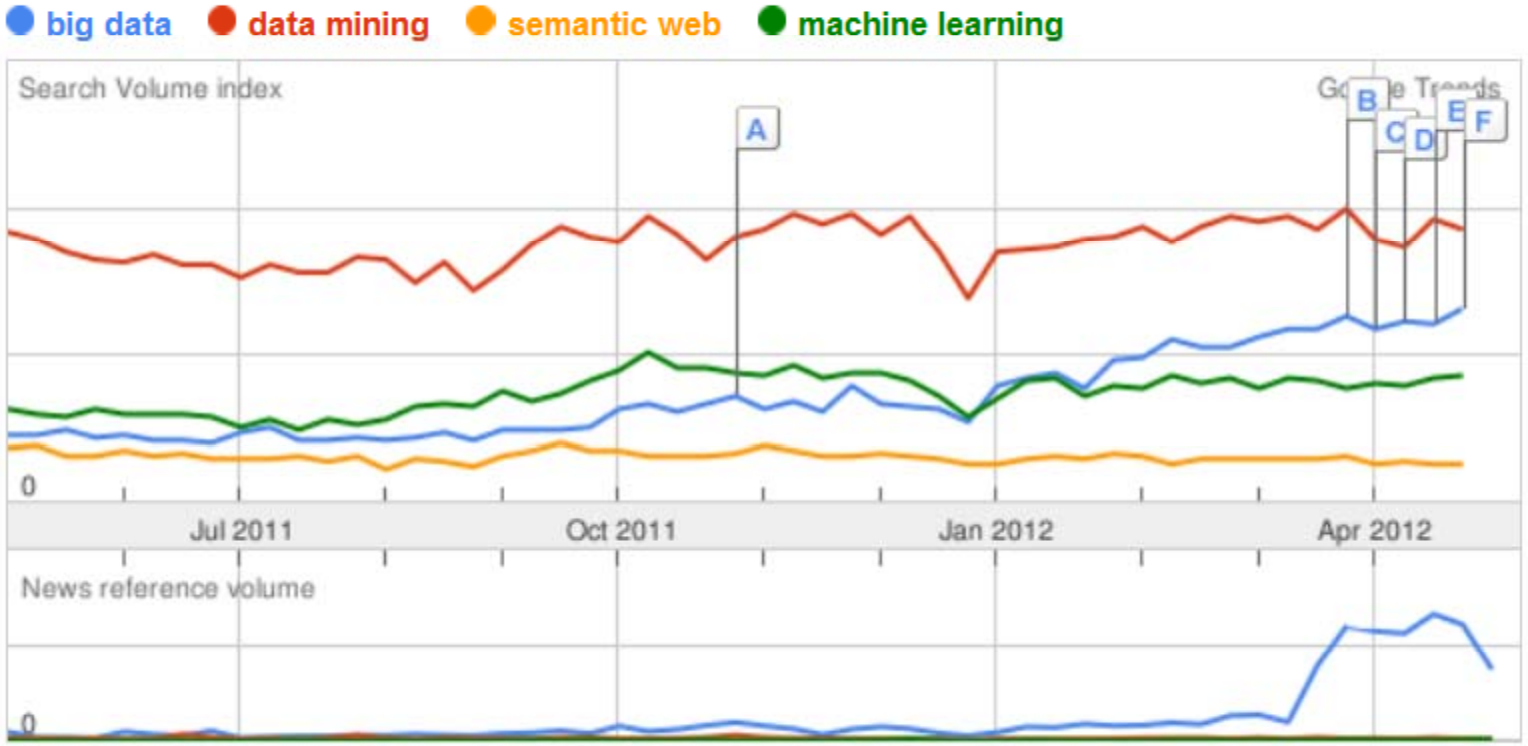
What is Big-Data?

- 'Big-data' is similar to 'Small-data', but bigger
- …but having data bigger consequently requires different approaches:
  - techniques, tools & architectures
- …to solve:
  - New problems…
  - …and old problems in a better way.

Characterization of Big-Data: volume, velocity, variety (V3)



Big-Data popularity on the Web



A [Spectra Logic Delivers ExaScale Storage for 'Big Data'; Announces Series of Products and Advancements and Unveils World's Highest Capacity Storage System]
MarketWatch - Nov 1 2011

B [Webcast: Obama Goes Big on Big Data]
Wired News - Mar 27 2012

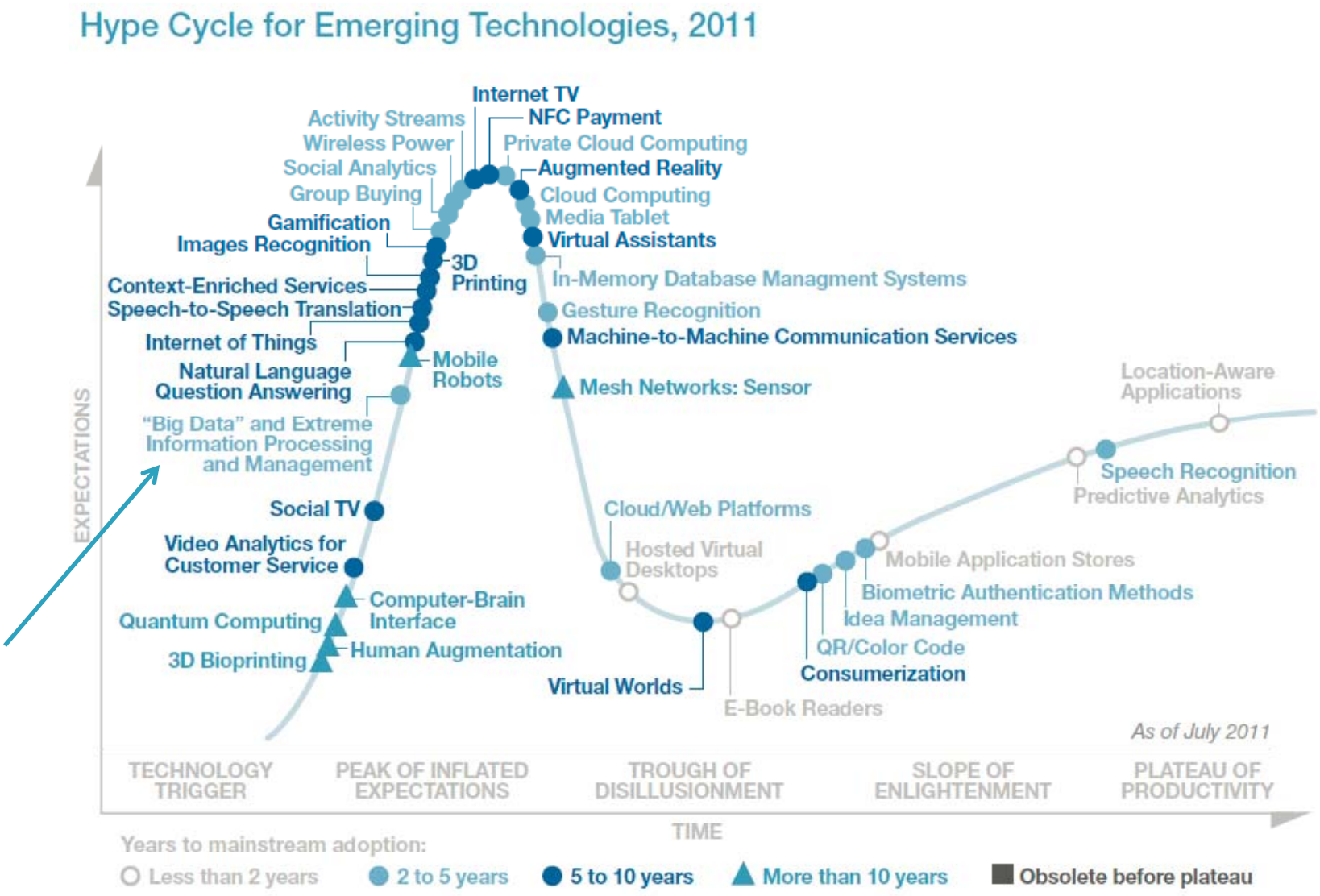
C [Cisco Joins Forces with EMC to Advance IT Skills in Cloud, Big Data and Data Center Technologies]
Justmeans - Apr 3 2012

D [Ferranti Unveils its MECOMS™ "Big Data" Strategy for Utility Meter Data Management and Real Time Billing]
Victoria Times Colonist - Apr 10 2012

E [Deconstructing Big Data - BuildZoom Launches an Article Series that Reveals the Hype and Substance Behind Big Data]
Houston Chronicle - Apr 17 2012

F [Harvard Releases Big Data for Books]
New York Times - Apr 24 2012

Big-Data in Gartner Hype-Cycle 2011



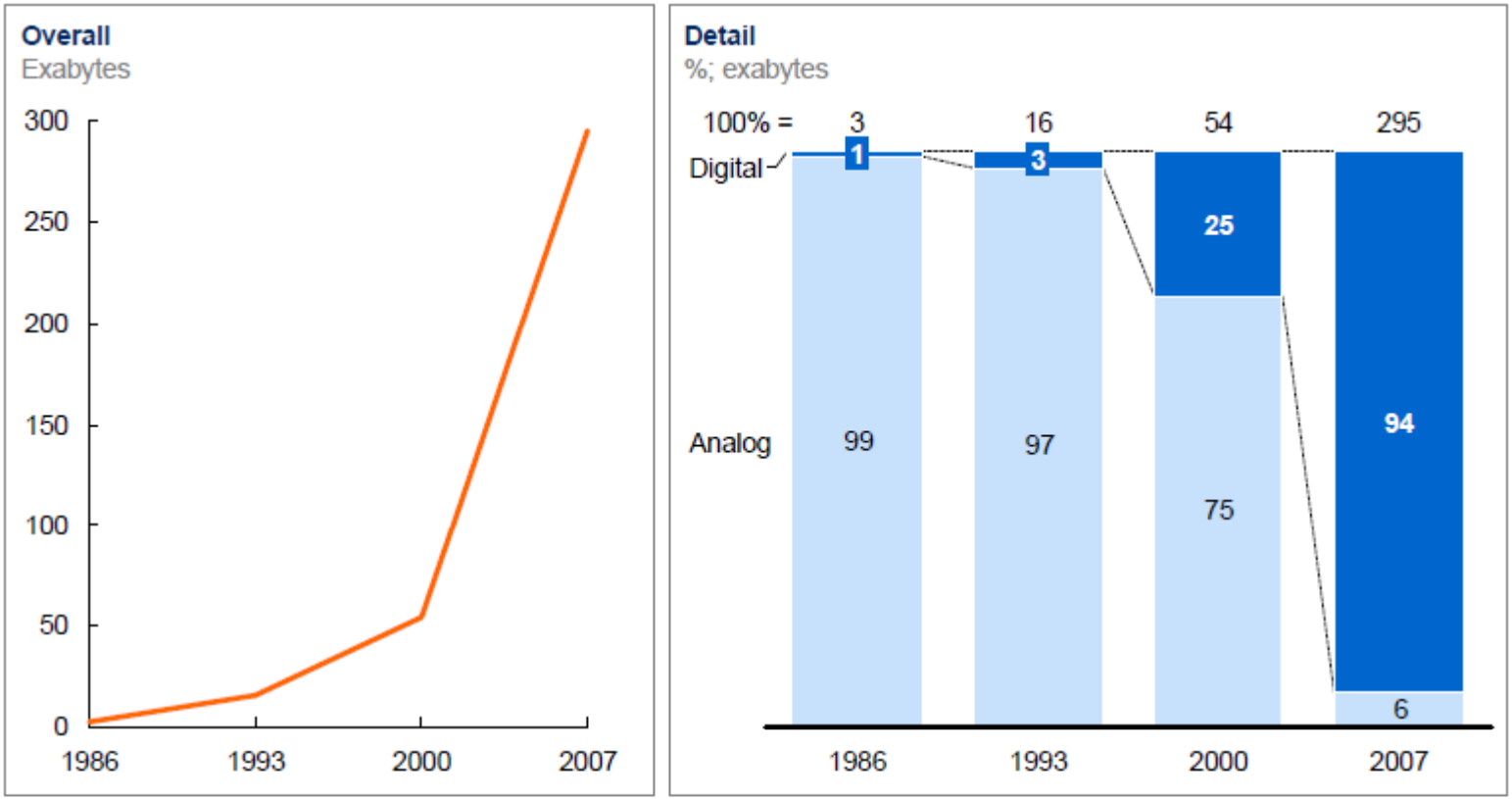
Why Big-Data?

- Key enablers for the growth of “Big Data” are:
  - Increase of storage capacities
  - Increase of processing power
  - Availability of data

Enabler: Data storage

Data storage has grown significantly, shifting markedly from analog to digital after 2000

Global installed, optimally compressed, storage



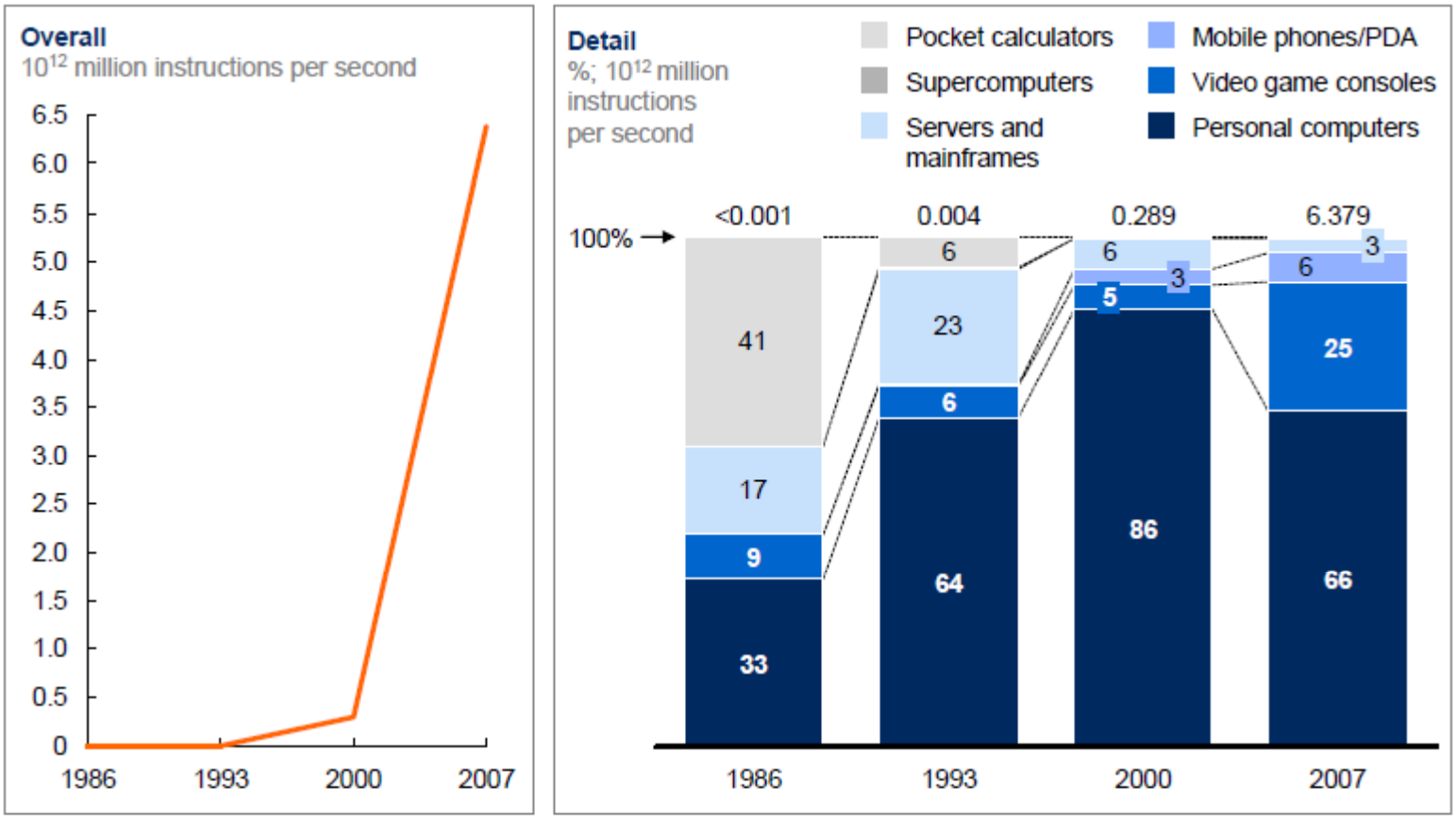
NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Enabler: Computation capacity

Computation capacity has also risen sharply

Global installed computation to handle information

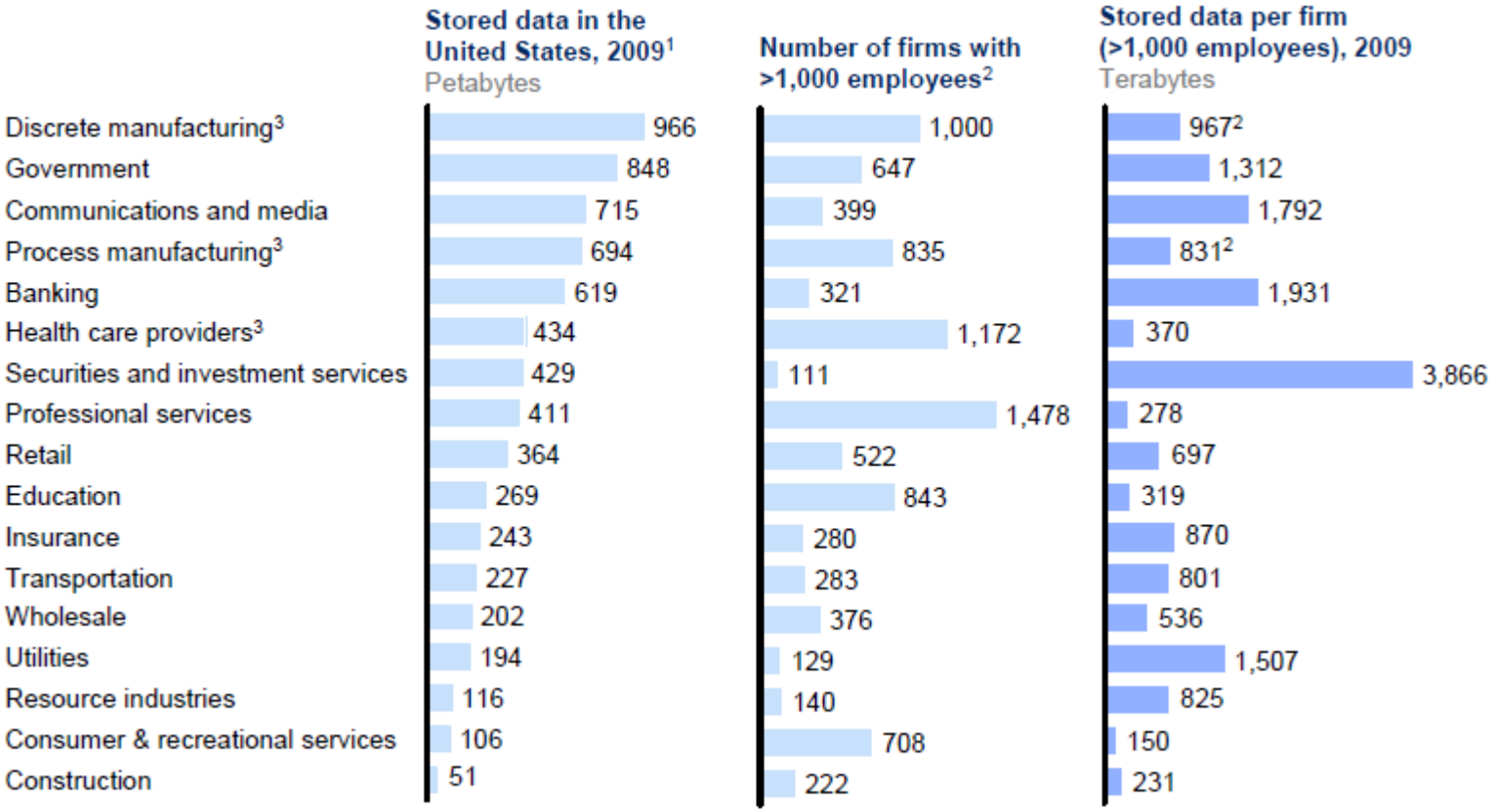


NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Enabler: Data availability

Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte

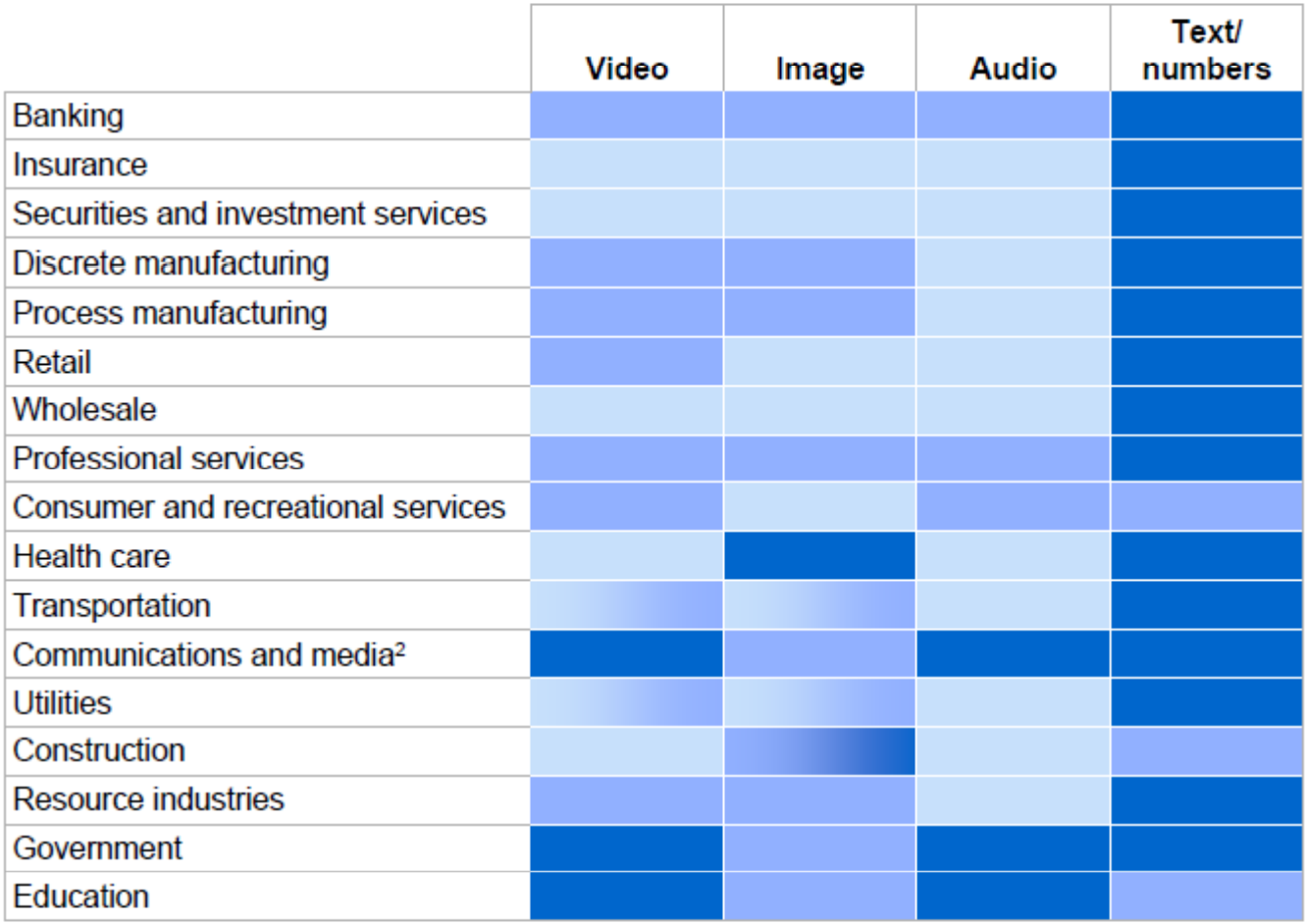


1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

Type of available data

The type of data generated and stored varies by sector[1]



1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.
SOURCE: McKinsey Global Institute analysis

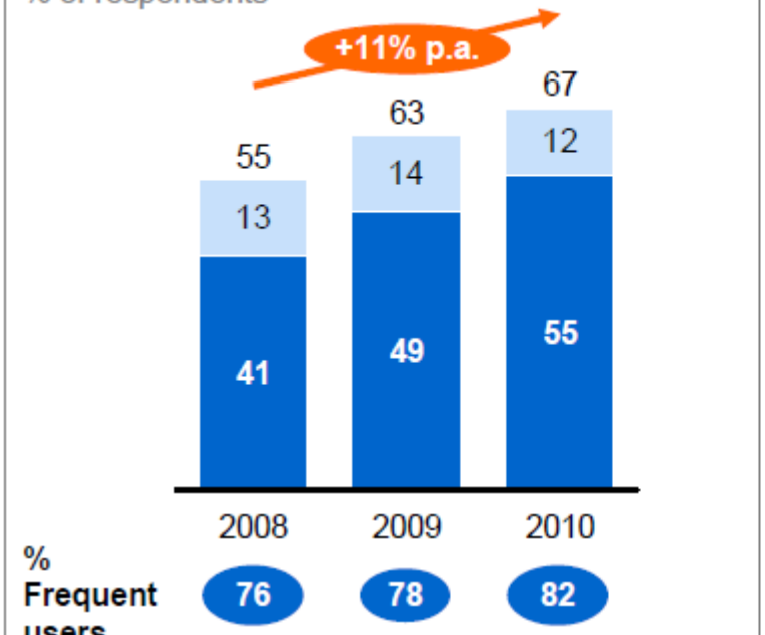
Data available from social networks and mobile devices

The penetration of social networks is increasing online and on smartphones; frequent users are increasing as a share of total users[1]

Frequent user[2]

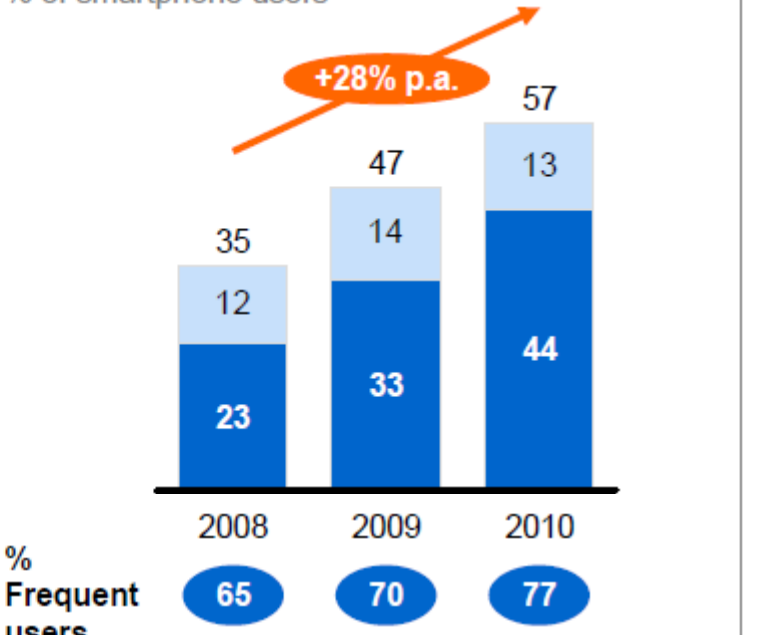
Social networking penetration on the PC is slowing, but frequent users are still increasing

% of respondents



Social networking penetration of smartphones has nearly doubled since 2008

% of smartphone users



1 Based on penetration of users who browse social network sites. For consistency, we exclude Twitter-specific questions (added to survey in 2009) and location-based mobile social networks (e.g., Foursquare, added to survey in 2010).

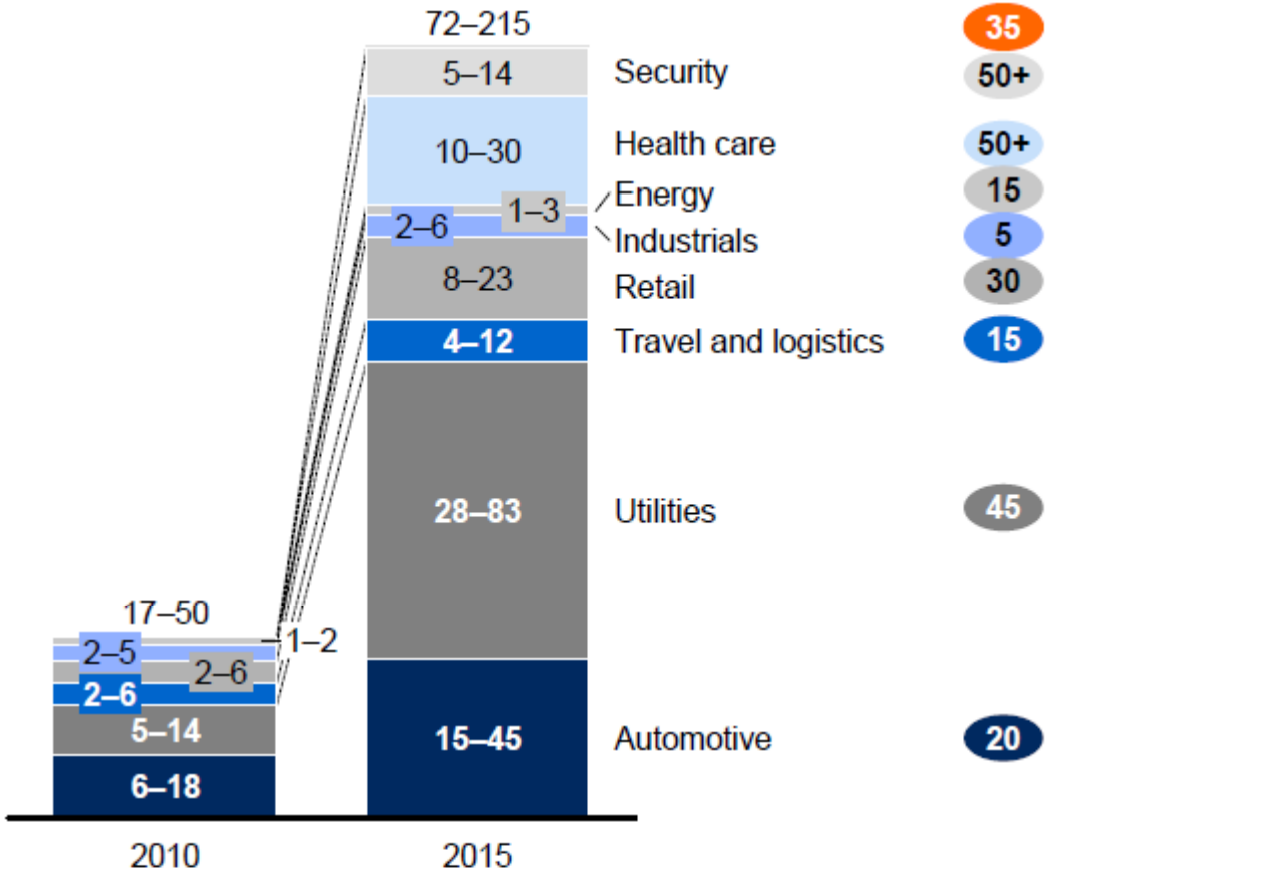
2 Frequent users defined as those that use social networking at least once a week.

SOURCE: McKinsey iConsumer Survey

Data available from "Internet of Things"

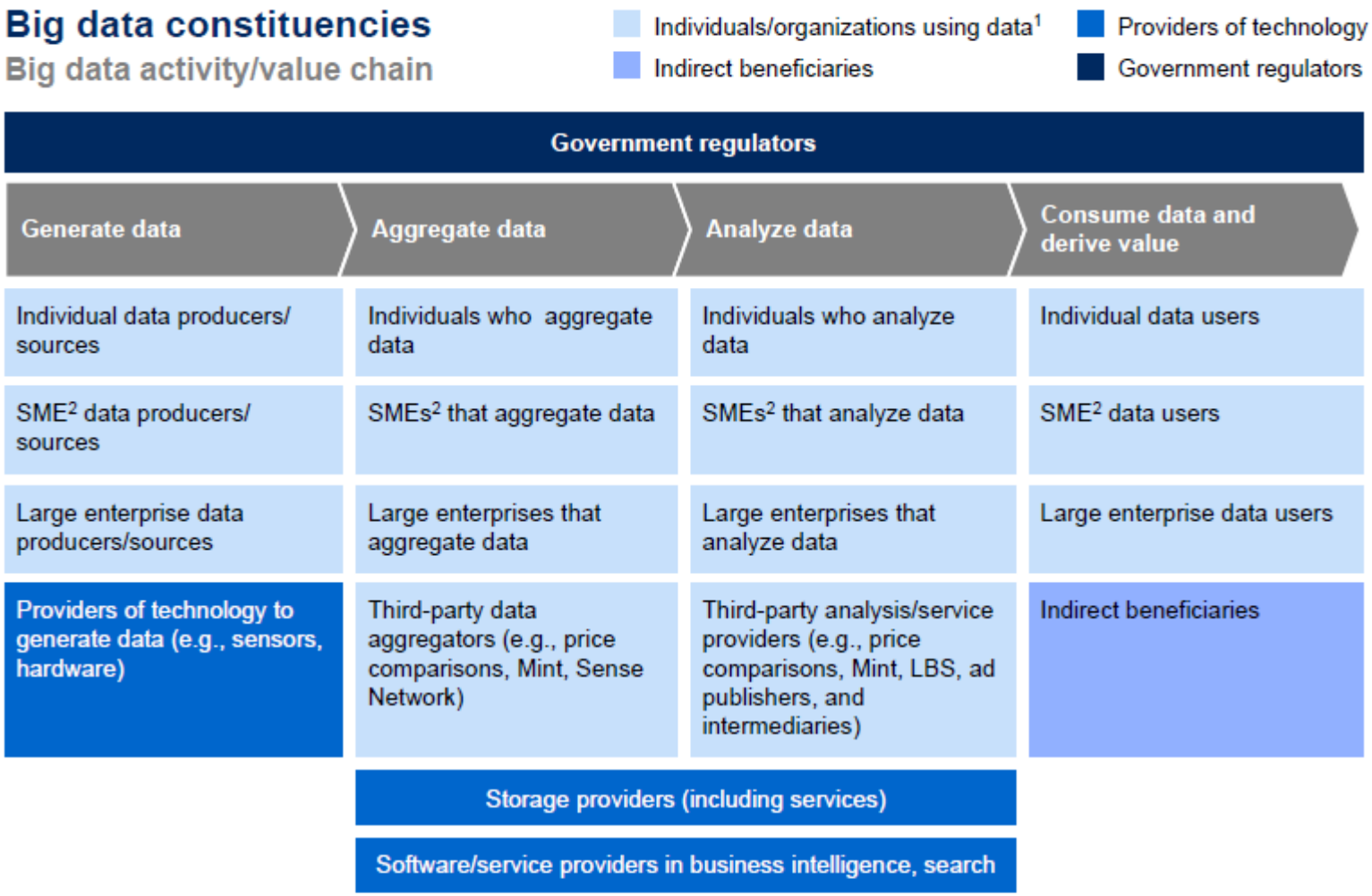
Data generated from the Internet of Things will grow exponentially as the number of connected nodes increases

Estimated number of connected nodes
Million



NOTE: Numbers may not sum due to rounding.
SOURCE: Analyst interviews; McKinsey Global Institute analysis

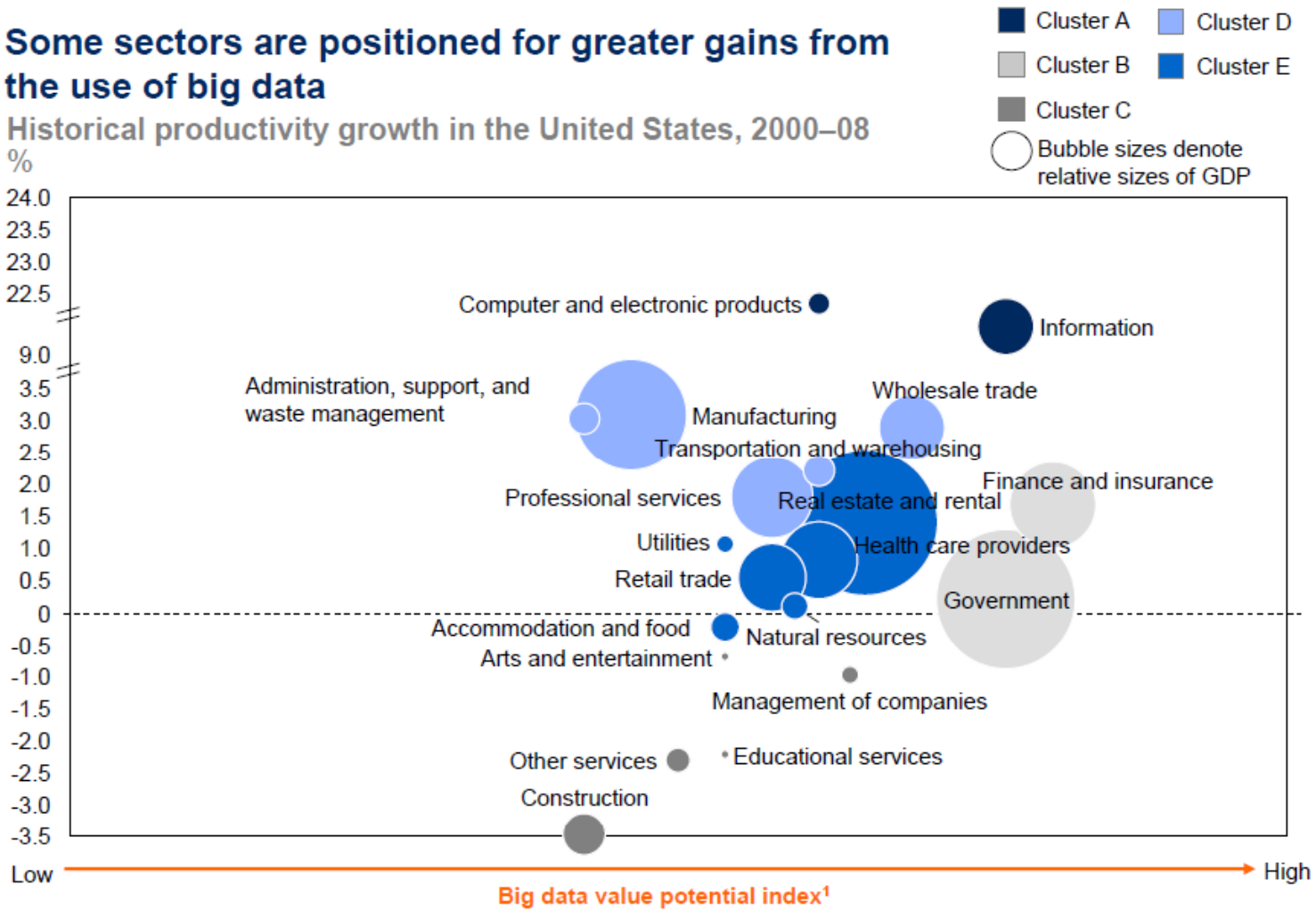
Big-data value chain



1 Individuals/organizations generating, aggregating, analyzing, or consuming data.
2 Small and medium-sized enterprises.

SOURCE: McKinsey Global Institute analysis

Gains from Big-Data per sector

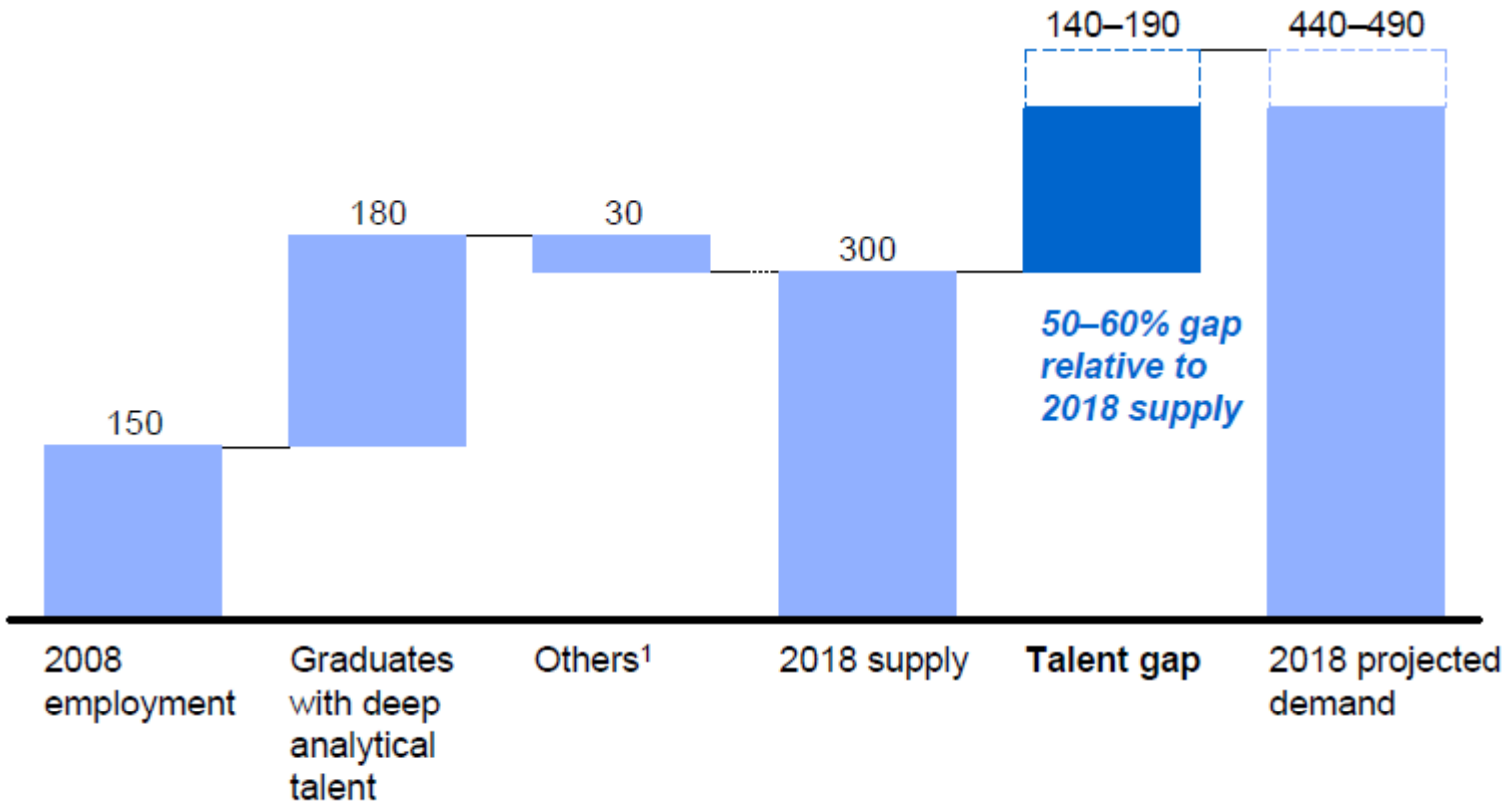


1 See appendix for detailed definitions and metrics used for value potential index.
SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

Predicted lack of talent for Big-Data related technologies

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018
Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

Tools

Tools typically used in Big-Data scenarios

- NoSQL
  - DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- MapReduce
  - Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- Storage
  - S3, Hadoop Distributed File System
- Servers
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- Processing
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

Techniques

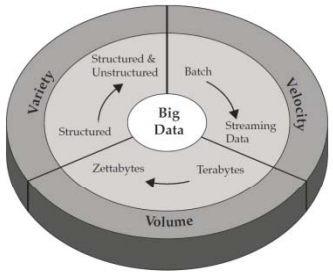
When Big-Data is really a hard problem?

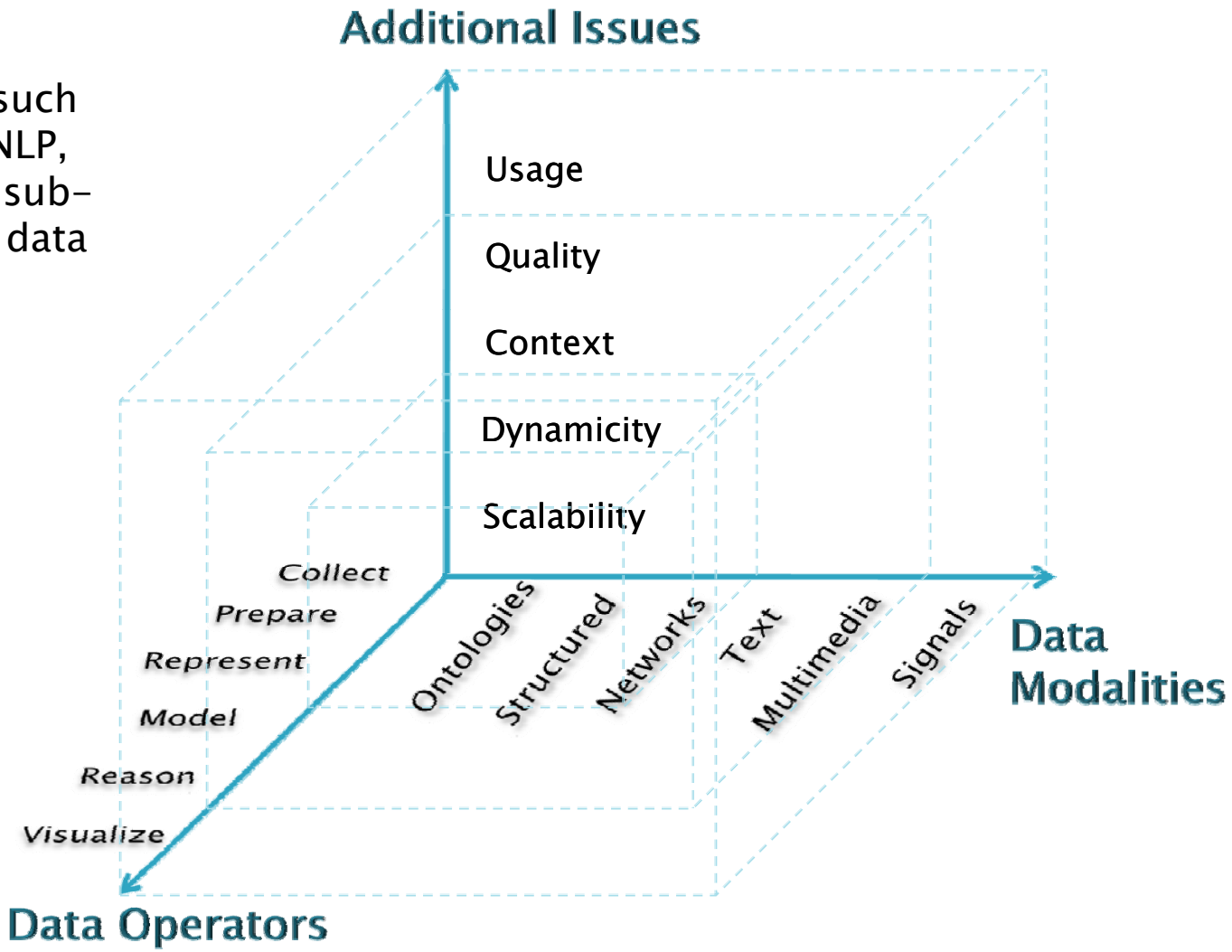
- …when the operations on data are complex:
  - …e.g. simple counting is not a complex problem
  - Modeling and reasoning with data of different kinds can get extremely complex

- Good news about big-data:
  - Often, because of vast amount of data, modeling techniques can get simpler (e.g. smart counting can replace complex model based analytics)…
  - …as long as we deal with the scale

What matters when dealing with data?

- Research areas (such as IR, KDD, ML, NLP, SemWeb, …) are sub-cubes within the data cube

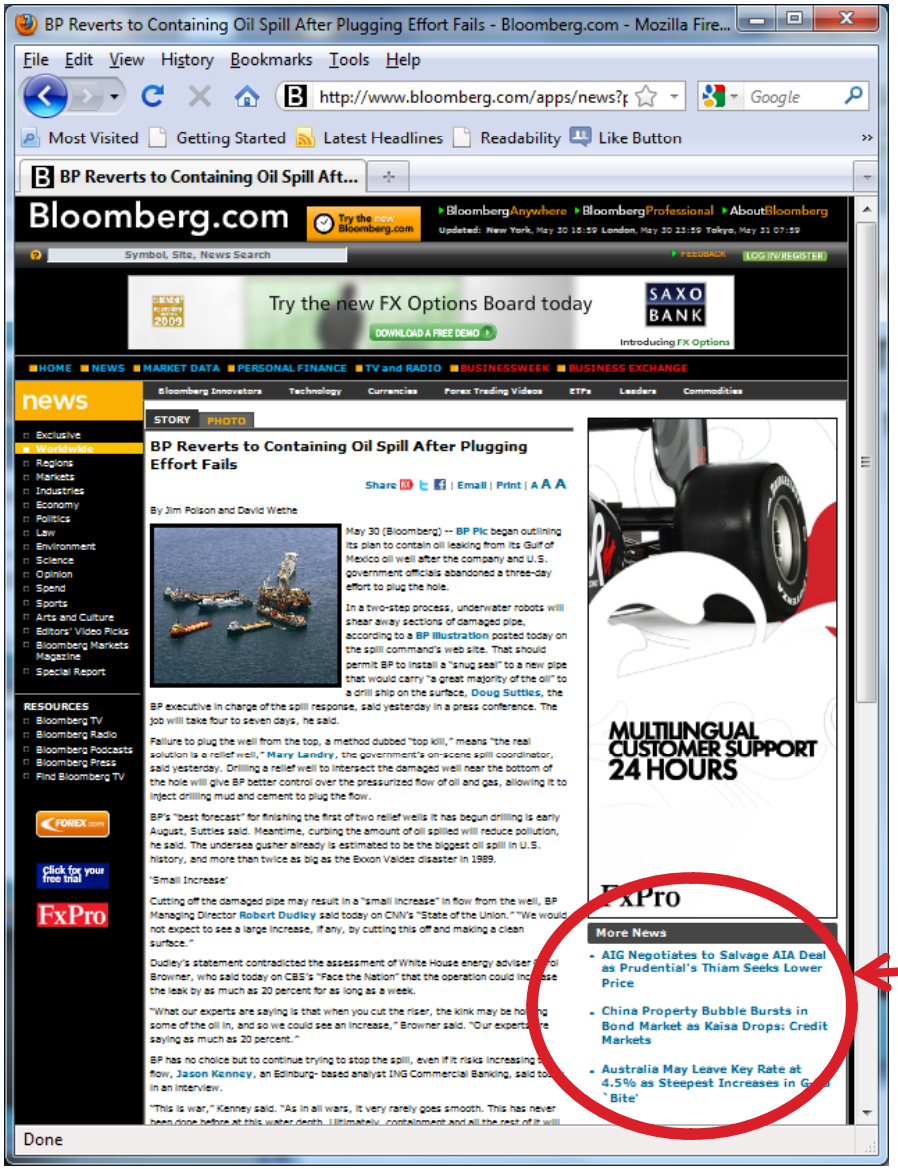




Applications

Recommendation

…an example: recommendation @Bloomberg.com



- Good recommendations can make a big difference when keeping a user on a web site
  - …the key is how rich context model a system is using to select information for a user
  - Bad recommendations <1% users, good ones >5% users click

Contextual personalized recommendations generated in ~20ms

Each click on the web site is enriched and indexed using:

- Domain
- Sub-domain
- Page URL
- URL sub-directories

- Page Meta Tags
- Page Title
- Page Content
- Named Entities

- Has Query
- Referrer Query

- Referring Domain
- Referring URL
- Outgoing URL

- GeoIP Country
- GeoIP State
- GeoIP City

- Absolute Date
- Day of the Week
- Day period
- Hour of the day
- User Agent

- Zip Code
- State
- Income
- Age
- Gender
- Country
- Job Title
- Job Industry

Application: Online Advertising for NYTimes (microtrends detection)

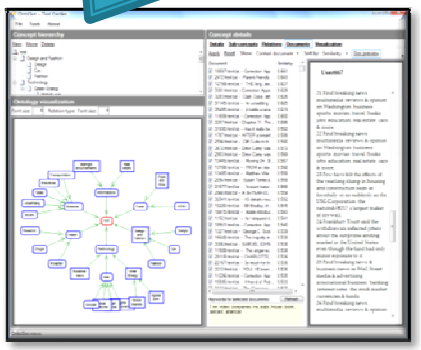
Log Files (~100M page clicks per day)

Stream of clicks

User profiles

Stream of profiles

Trend Detection System



Content

NYT articles

Trends and updated segments

Segment	Keywords
Stock Market	Stock Market, mortgage, banking, investors, Wall Street, turmoil, New York Stock Exchange
Health	diabetes, heart disease, disease, heart, illness
Green Energy	Hybrid cars, energy, power, model, carbonated, fuel, bulbs,
Hybrid cars	Hybrid cars, vehicles, model, engines, diesel
Travel	travel, wine, opening, tickets, hotel, sites, cars, search, restaurant
…	…

Segments

Sales

Campaign to sell segments

$

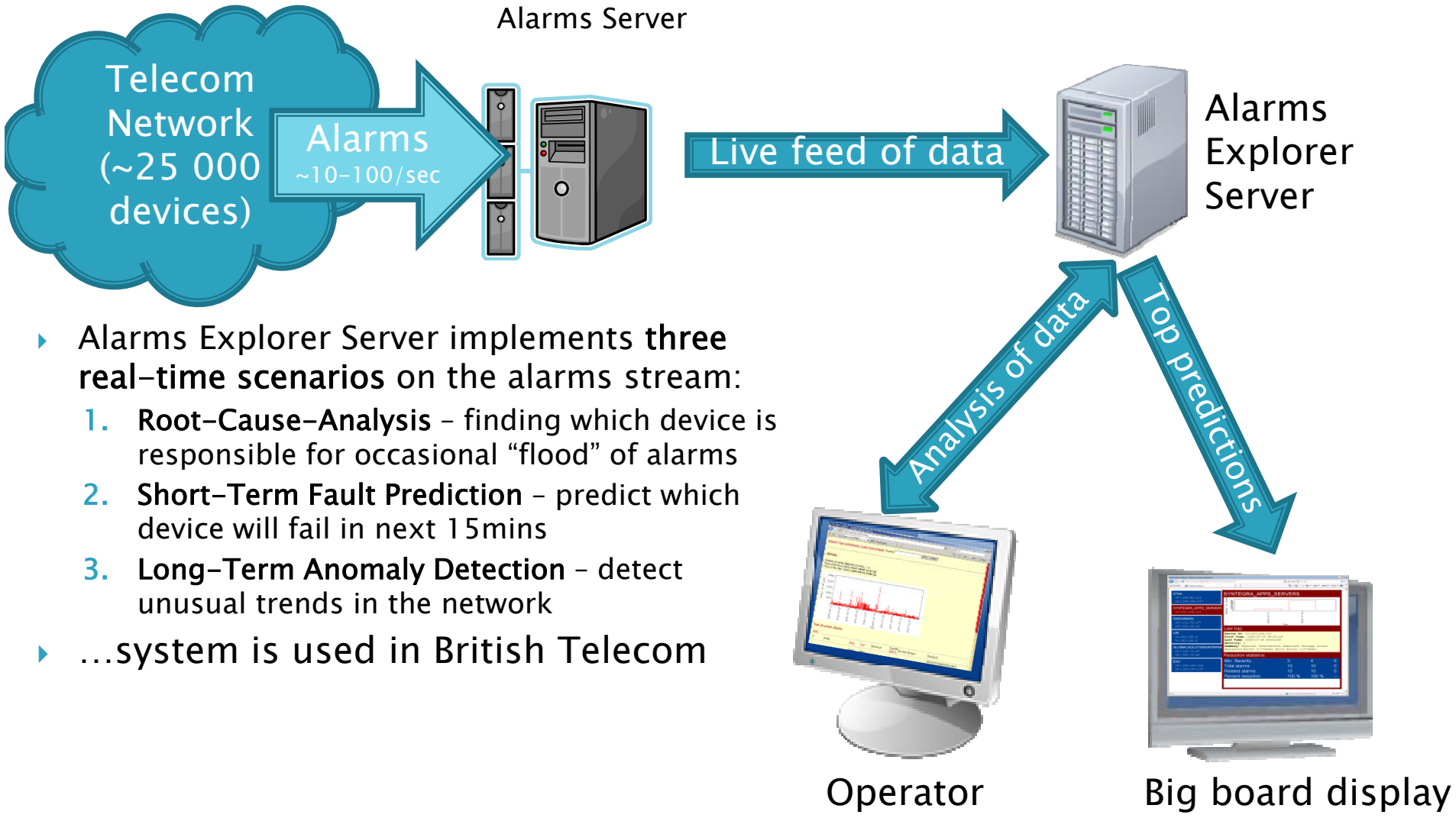
Advertisers

Figures for one day of NYTimes

- 50Gb of uncompressed log files
- 10Gb of compressed log files
- 0.5Gb of processed log files
- 50-100M clicks
- 4-6M unique users
- 7000 unique pages with more then 100 hits
- Index size 2Gb
- Pre-processing & indexing time
  - ~10min on workstation (4 cores & 32Gb)
  - ~1hour on EC2 (2 cores & 16Gb)

Root-cause analysis

Applications: Telecommunication Network Monitoring



- Alarms Explorer Server implements **three real-time scenarios** on the alarms stream:
  1. **Root-Cause-Analysis** – finding which device is responsible for occasional "flood" of alarms
  2. **Short-Term Fault Prediction** – predict which device will fail in next 15mins
  3. **Long-Term Anomaly Detection** – detect unusual trends in the network
- …system is used in British Telecom

Analysis of MSN-Messenger Social-network

- Presented in “Planetary-Scale Views on a Large Instant-Messaging Network” by Jure Leskovec and Eric Horvitz WWW2008

Instant Messenger – Phenomena at a planetary scale

- Observe social and communication phenomena at a *planetary* scale
- Largest social network analyzed to date

Research questions:

- How does communication change with user demographics (age, sex, language, country)?
- How does geography affect communication?
- What is the structure of the communication **network**?

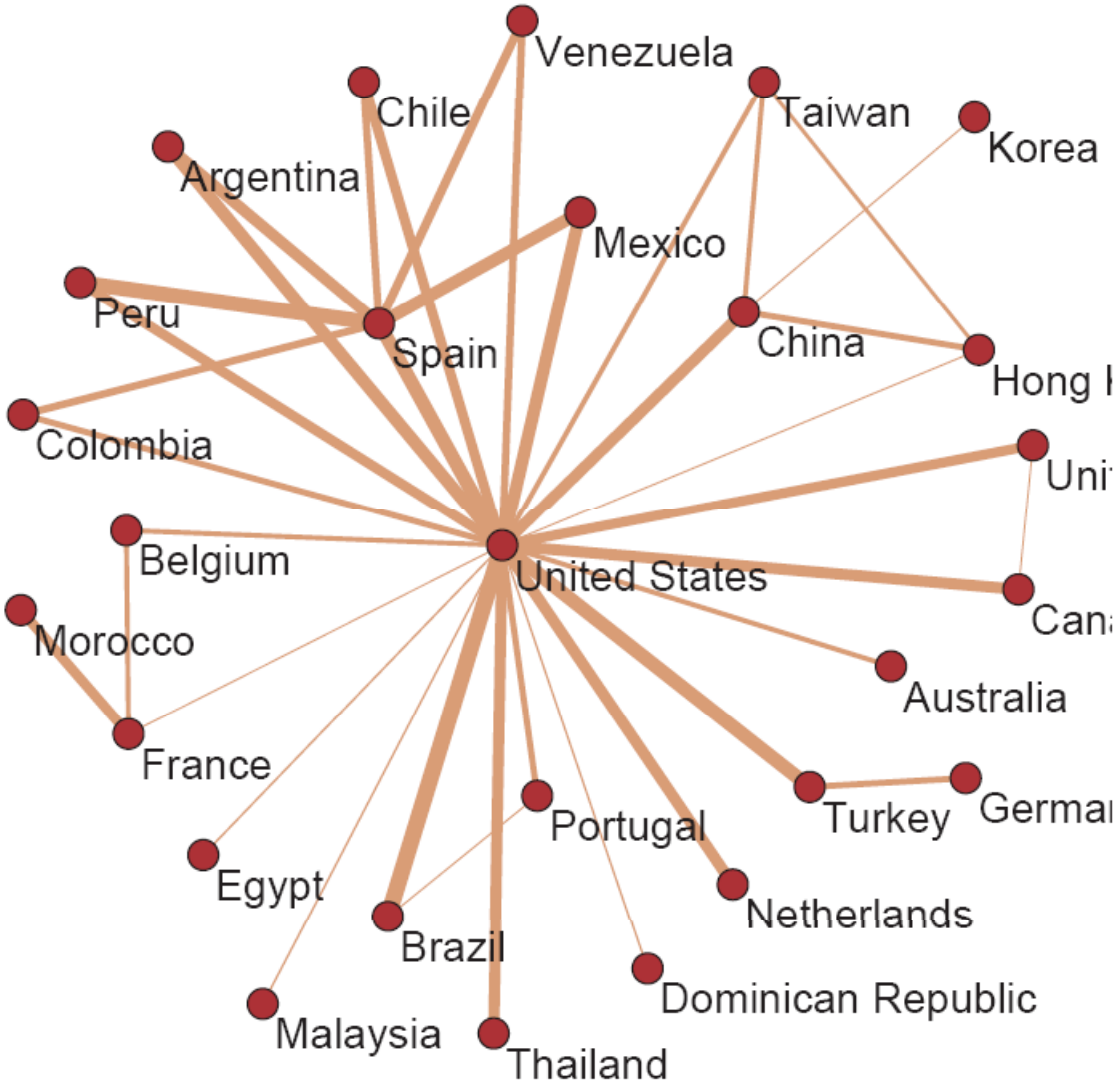
Data statistics: Total activity

- We collected the data for June 2006
- Log size:

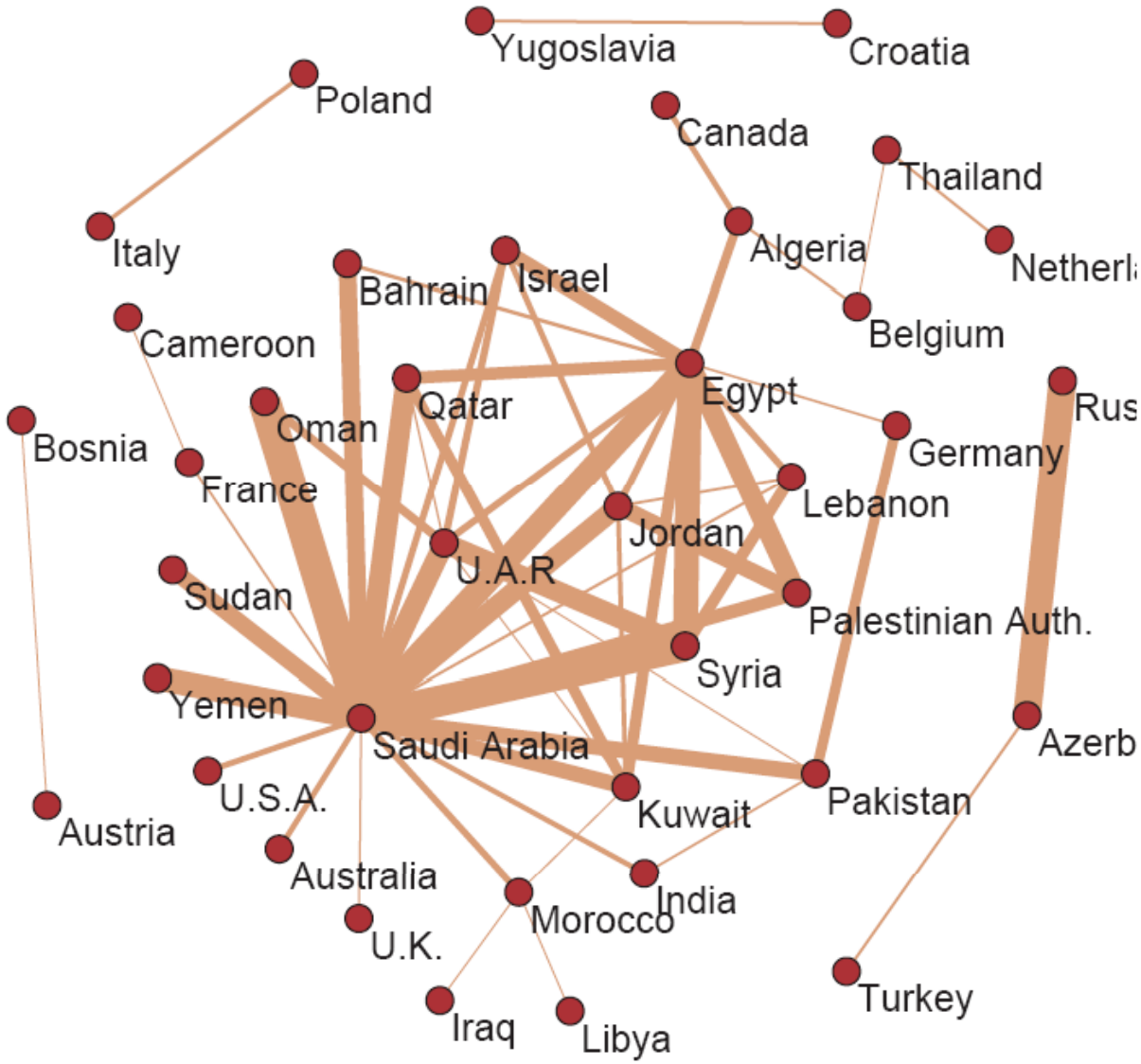
  150Gb/day (compressed)
- Total: 1 month of communication data:

  4.5Tb of compressed data
- Activity over June 2006 (30 days)
  - 245 million users logged in
  - 180 million users engaged in conversations
  - 17,5 million new accounts activated
  - More than 30 billion conversations
  - More than 255 billion exchanged messages

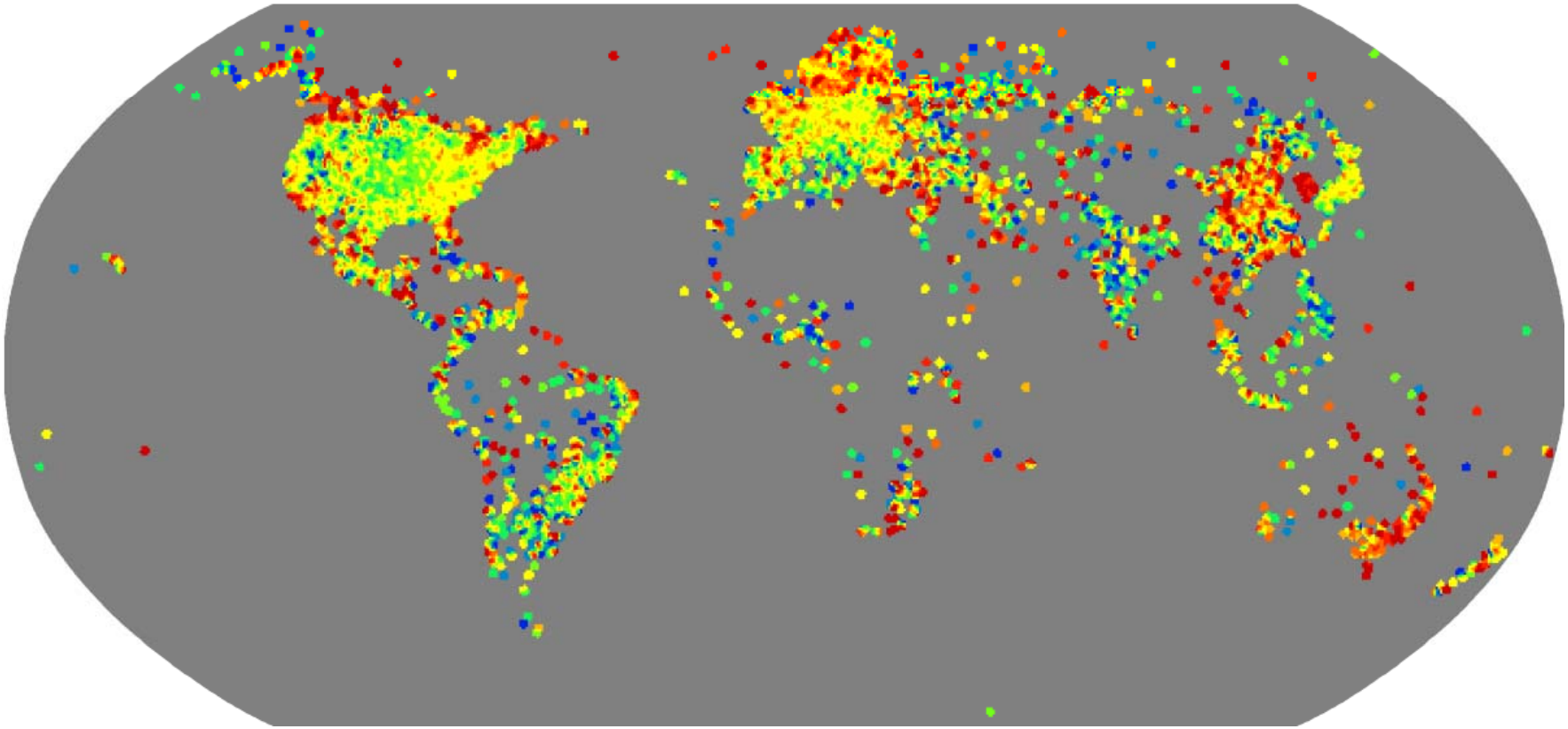
Who talks to whom: Number of conversations



Who talks to whom: Conversation duration



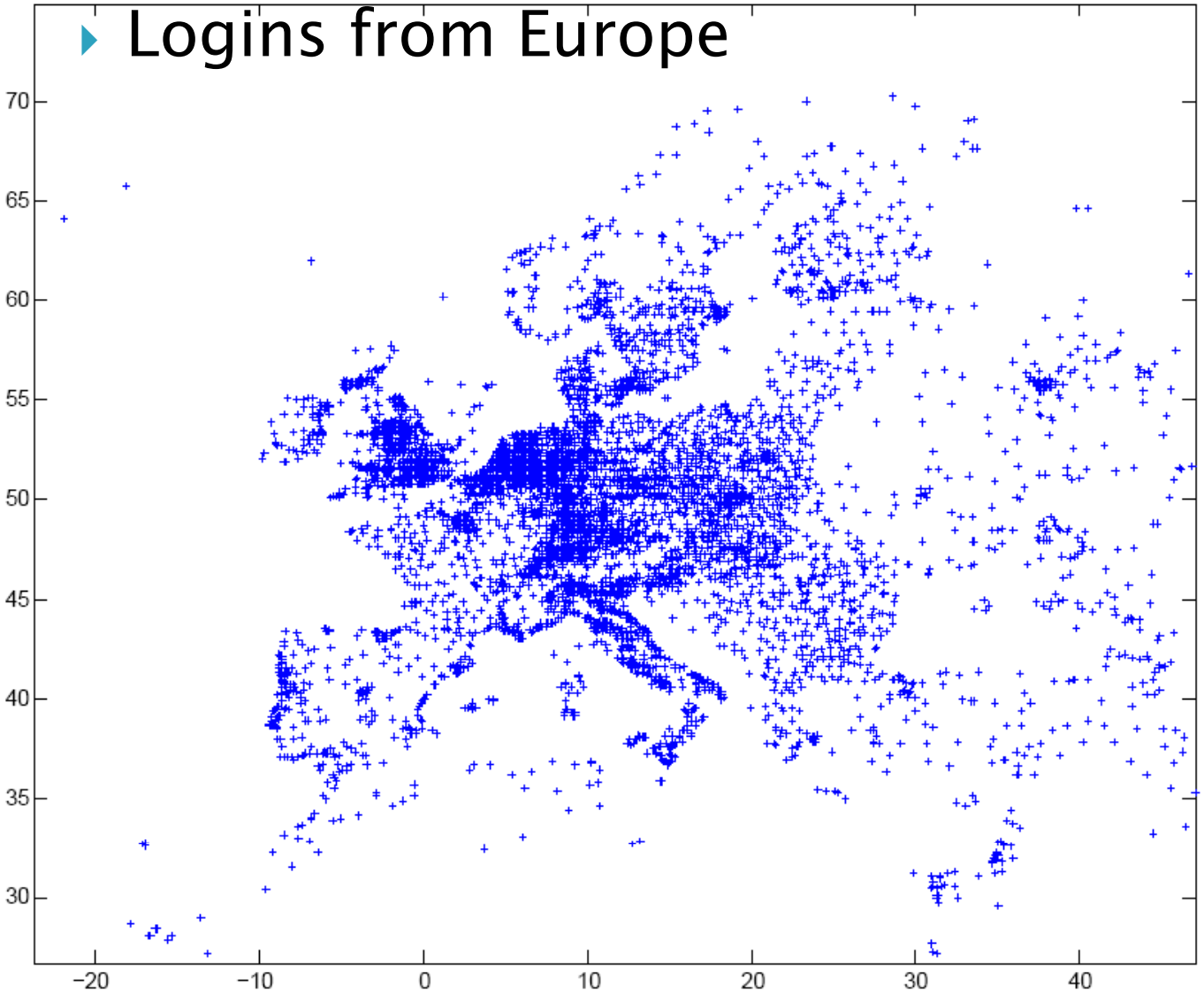
Geography and communication



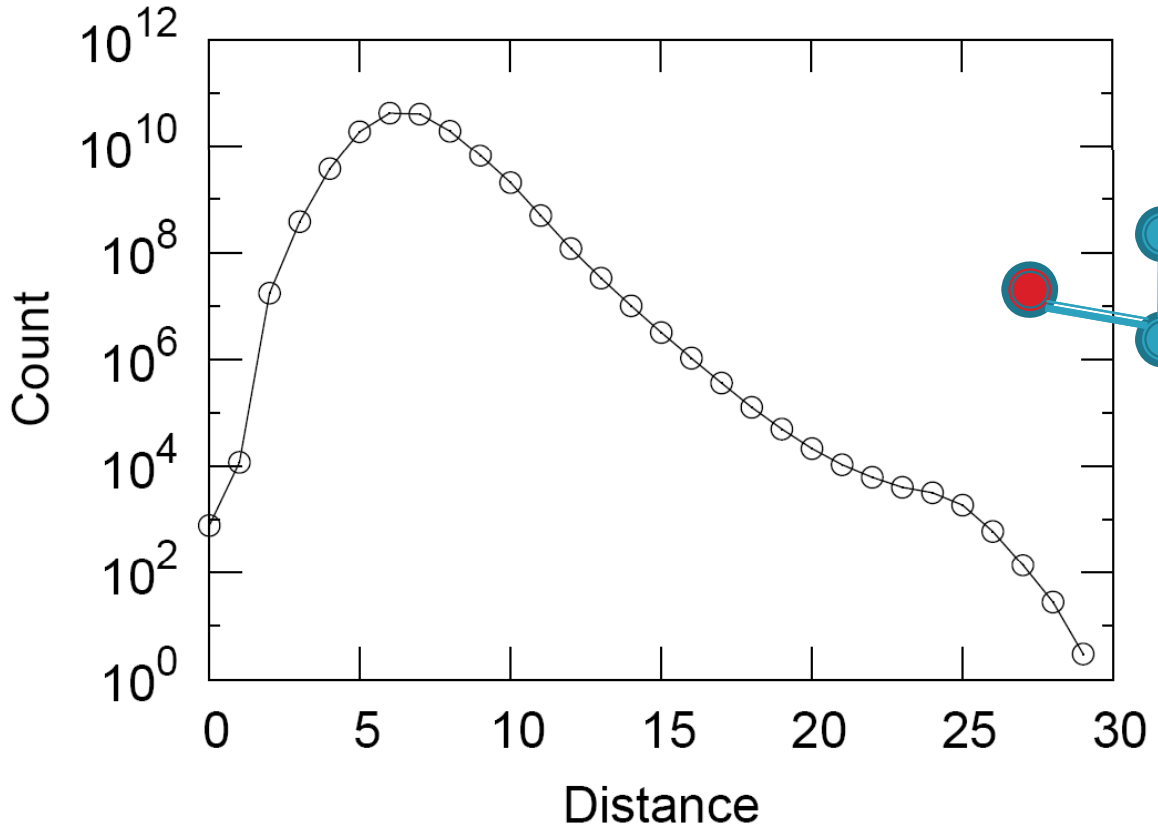
- Count the number of users logging in from particular location on the earth

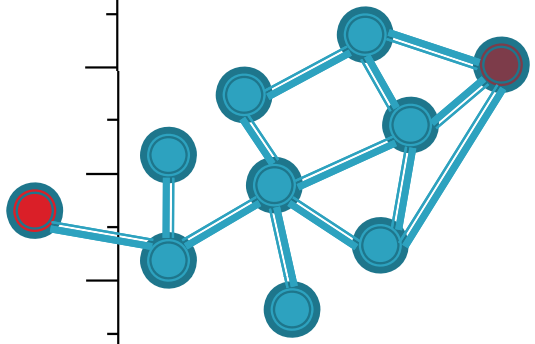
How is Europe talking

- Logins from Europe



Network: Small-world





Hops	Nodes
1	10
2	78
3	396
4	8648
5	3299252
6	28395849
7	79059497
8	52995778
9	10321008
10	1955007
11	518410
12	149945
13	44616
14	13740
15	4476
16	1542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

- 6 degrees of separation [Milgram '60s]
- Average distance between two random users is 6.6
- 90% of nodes can be reached in < 8 hops

Web-of-Things

Literature on Big-Data