Multi-Structured Data Sources

Outline

- Entity-centered view of the World
- Data Sources & Formats
- Motivating Example
- Is more data always better than better algorithms?

Entity View of the World

- Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

(Christine Borgmann 2014)

Entity View of the World

- Data is prevalent
  - Business Data:
    - Company filings to regulatory bodies
    - Security market (e.g., stock, fund, option) trading data
    - News articles, analyst reports,
  - Government Data:
    - US federal government spending data, earmarks data
    - Congress data (voting, members,)
- Users and applications prefer an entity view of the underlying data
  - Entities (Companies, People, Securities,)
  - Relationships (Employment, Investment, Ownership,)
  - Events (Mergers, Acquisitions, Bankruptcy, Appointment,)

What is Data Integration?

Data integration is the process of consolidating data from a set of heterogeneous data sources into a single uniform data set.

The integrated data set should:

1. Correctly and completely represent the content of all data sources.

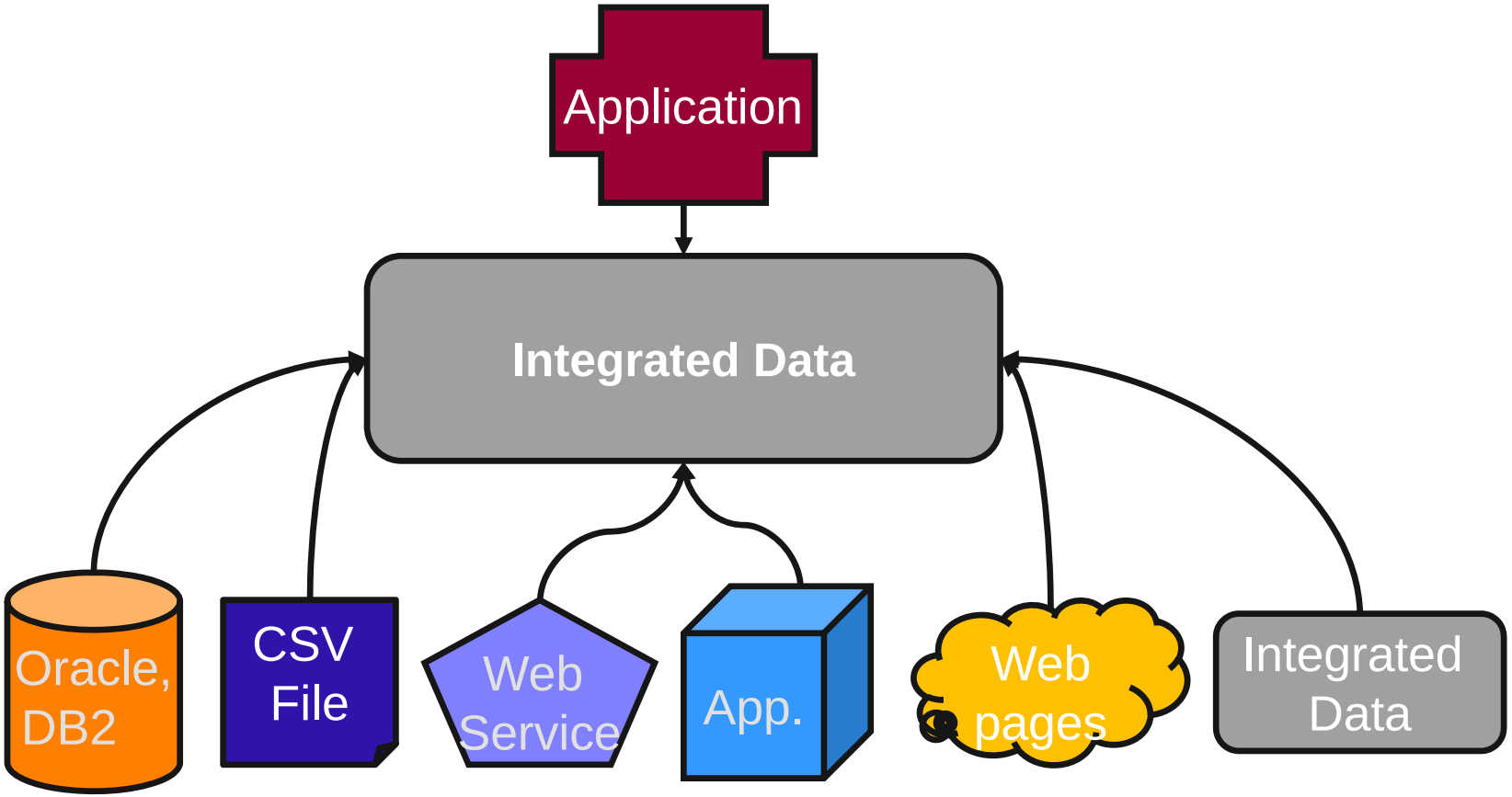
2. Use a single data model and a single schema.

3. Only contain a single representation of every real-world entity.

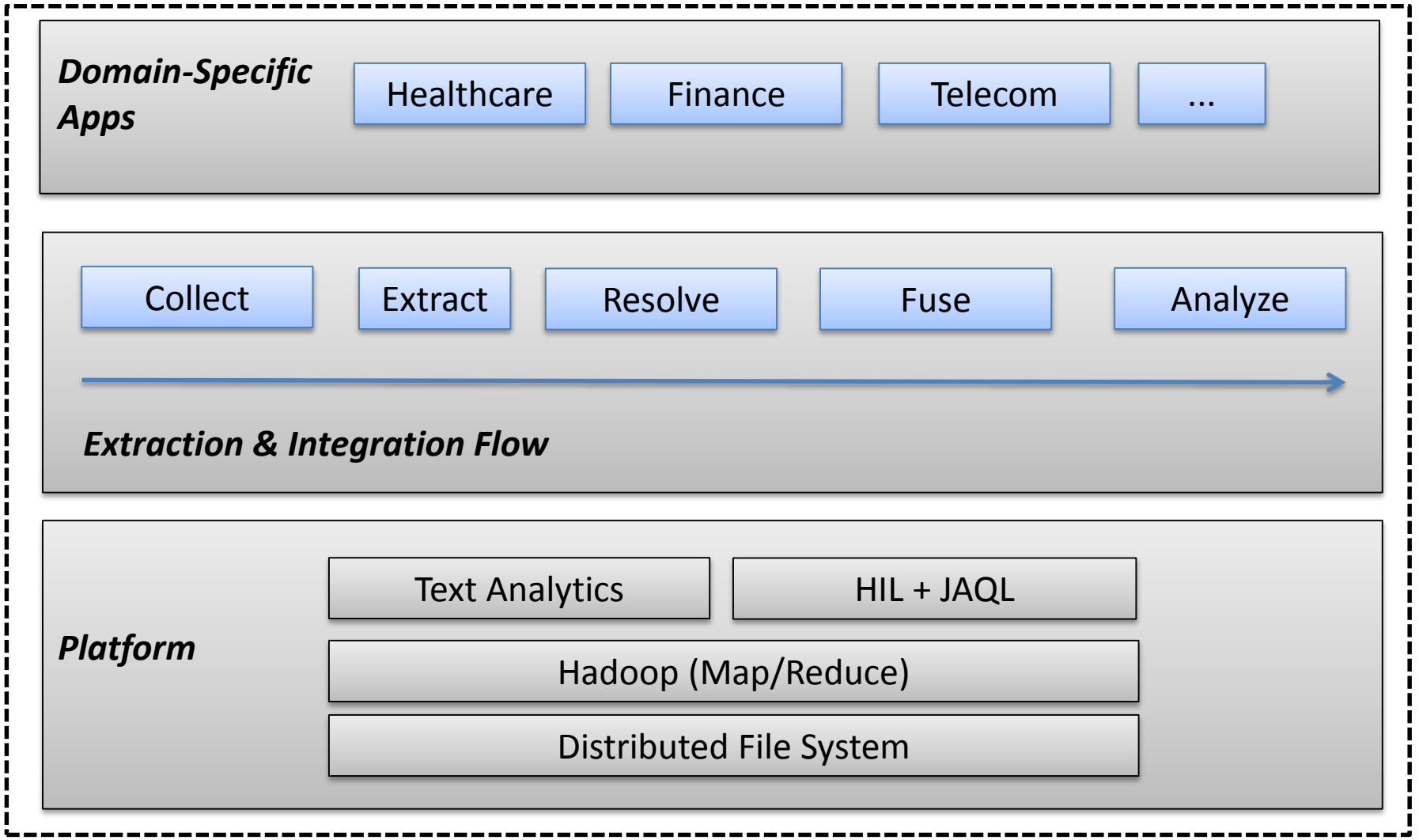
4. Not contain any conflicting data about single entities.

 To achieve this data integration needs to resolve different types of heterogeneity that exist between data sources.

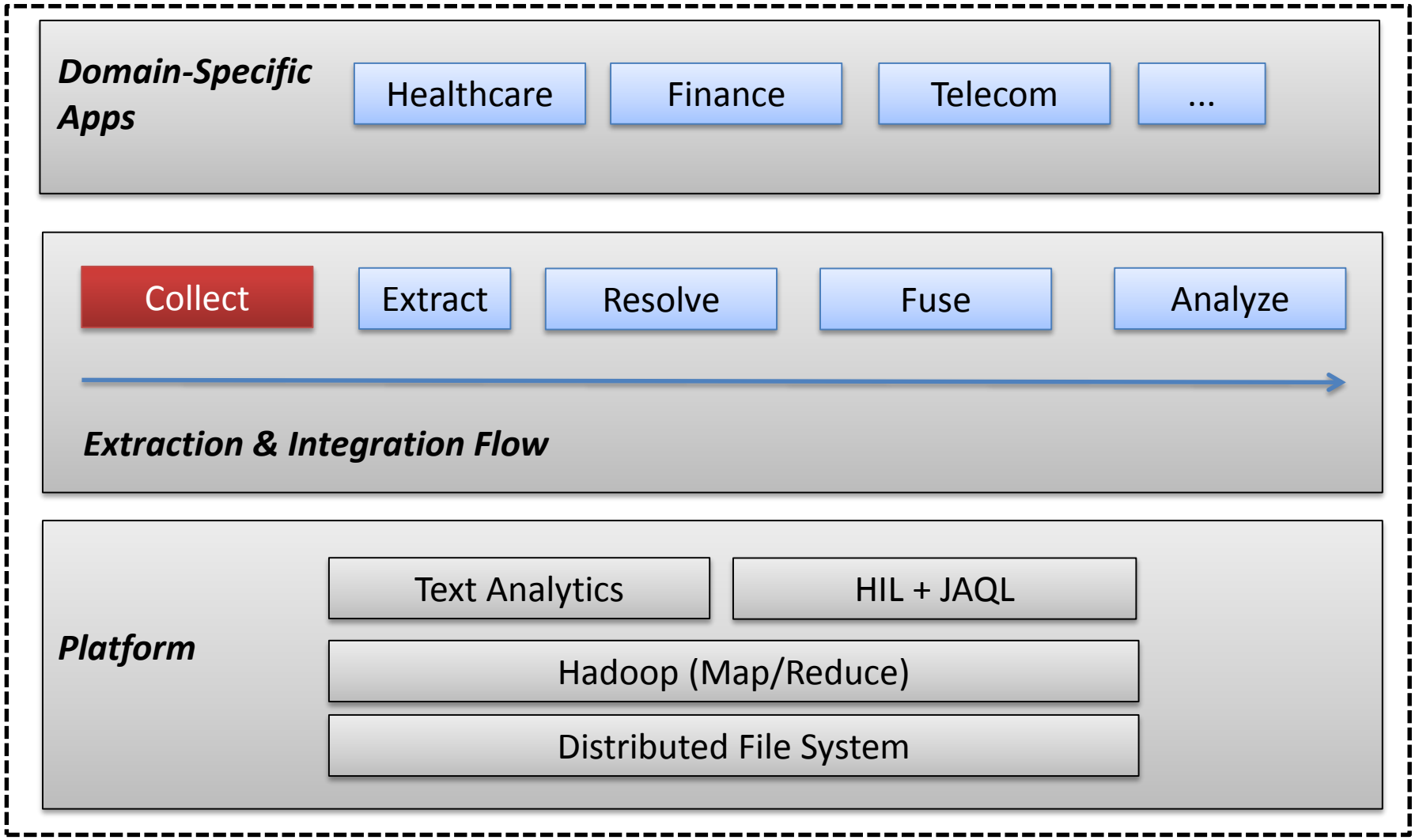
Data Integration



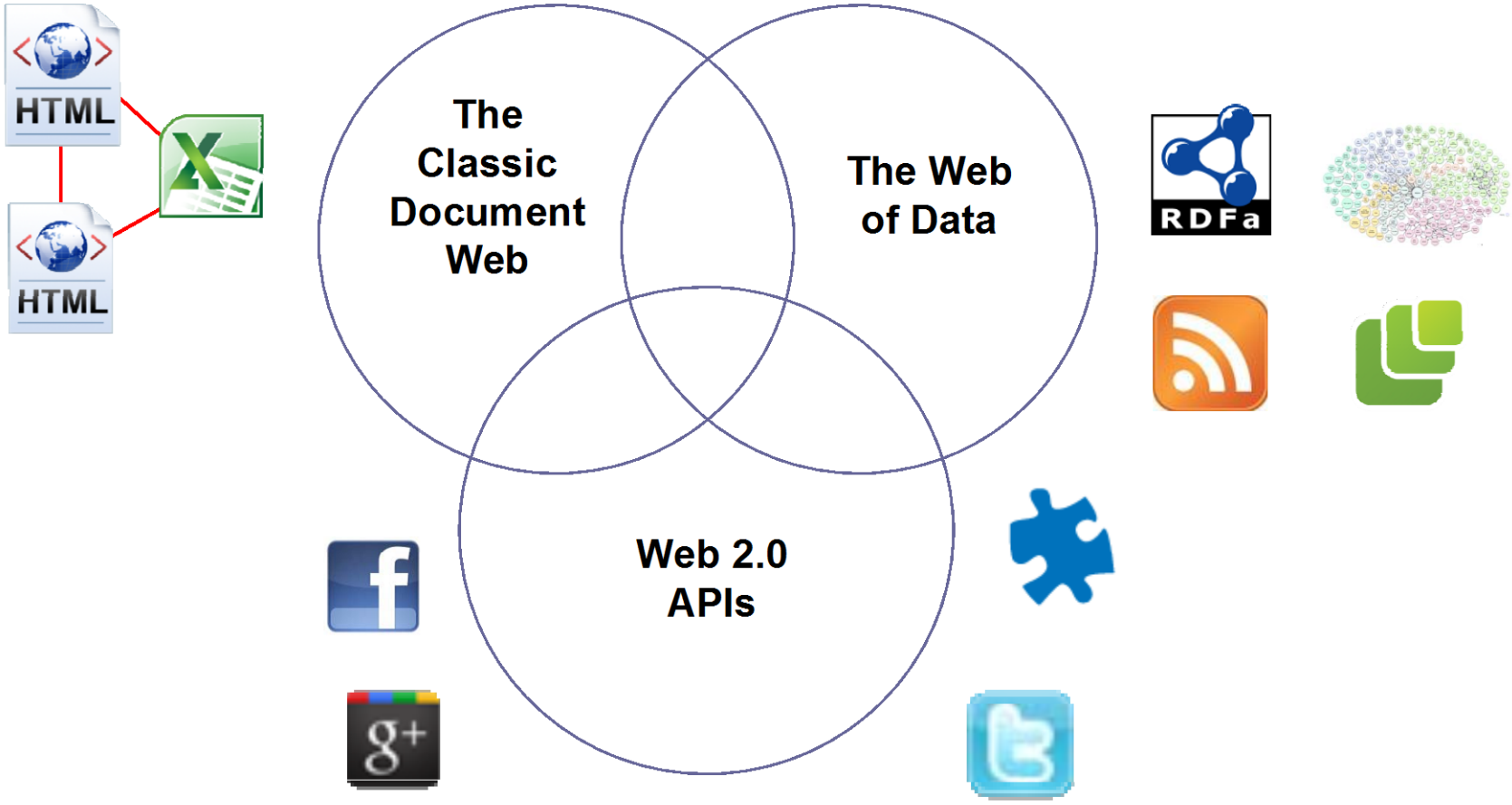
Data inside platform



Data inside platform



Topology of the Web Today



Data Catalogs and Marketplaces

- The Document Web traditionally contains structured data in various formats:
  - CSV Files, Excel Worksheets
  - XML Documents, SQL Dumps
- Data Catalogs and Data Market Places
  - collect and host data sets plus metadata
  - provide free or payment-based access to the data sets
- Examples
  - The Data Hub: data catalog containing 6,800 open-lisence data sets
  - Data.gov.uk, Data.gov.us: Thousands of public sector data sets
  - Infochimps, Azure Data Marketplace, Factual: commercial market places
  - data.mos.ru, hubofdata.ru
- List of Data Catalogs and Market Places
  - http://www.kdnuggets.com/datasets/api-hub-marketplace-platform.html

Организации Очистить Все

- Global (873)
- Economics Datasets (42)
- Linking Open Data C... (27)
- OpenSpending (25)
- Climate Data (21)
- Where Does My Money... (16)
- Canada (14)
- Bulgaria (14)
- Civil Society (13)
- occupy (10)

Показать больше Организации

Группы Очистить Все

Добавить пакет

Country

1 219 пакеты данных найдены для "Country"

Сортировать по: Актуальность

Country

country-level datasets

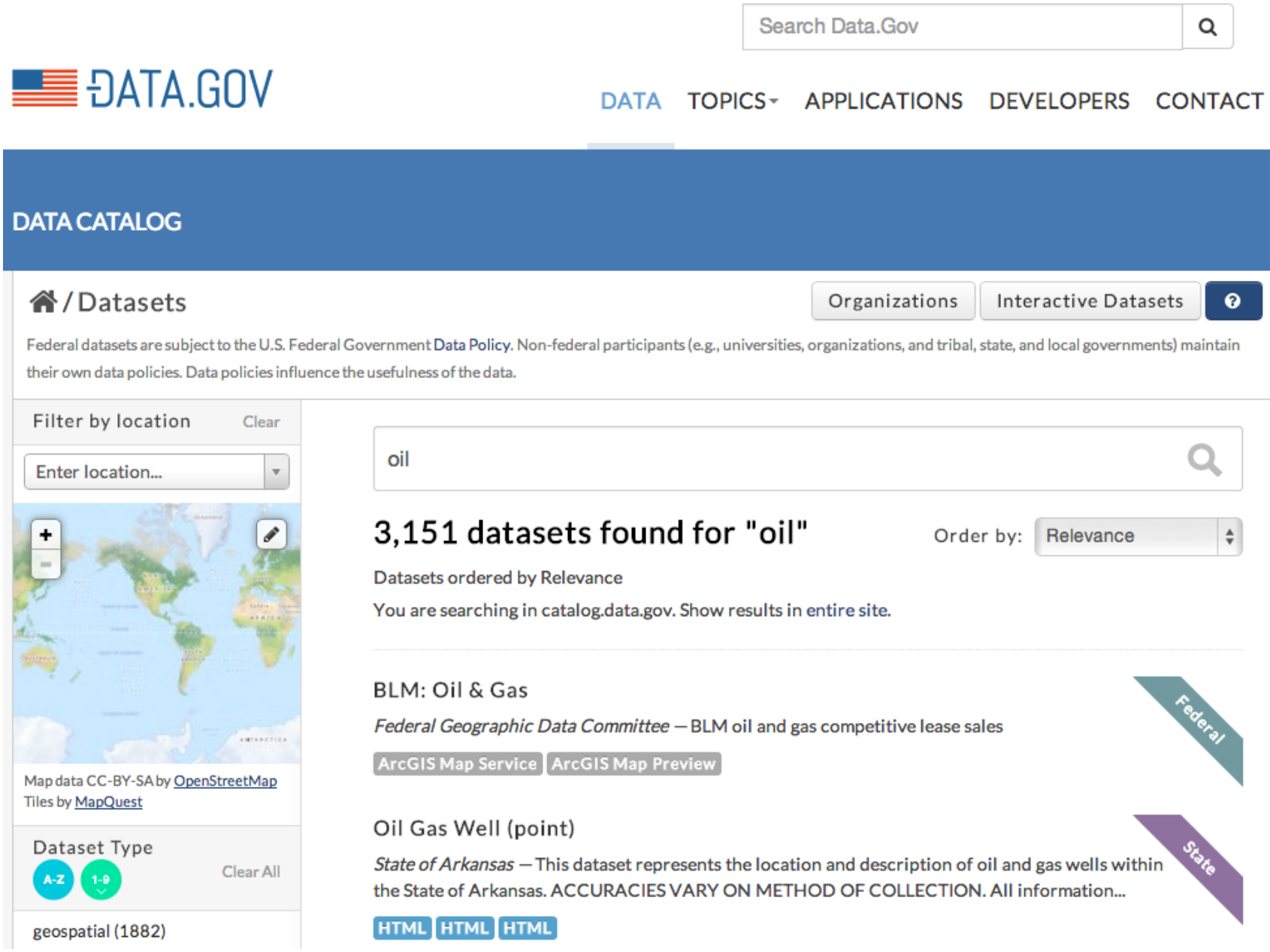
Countries Continents

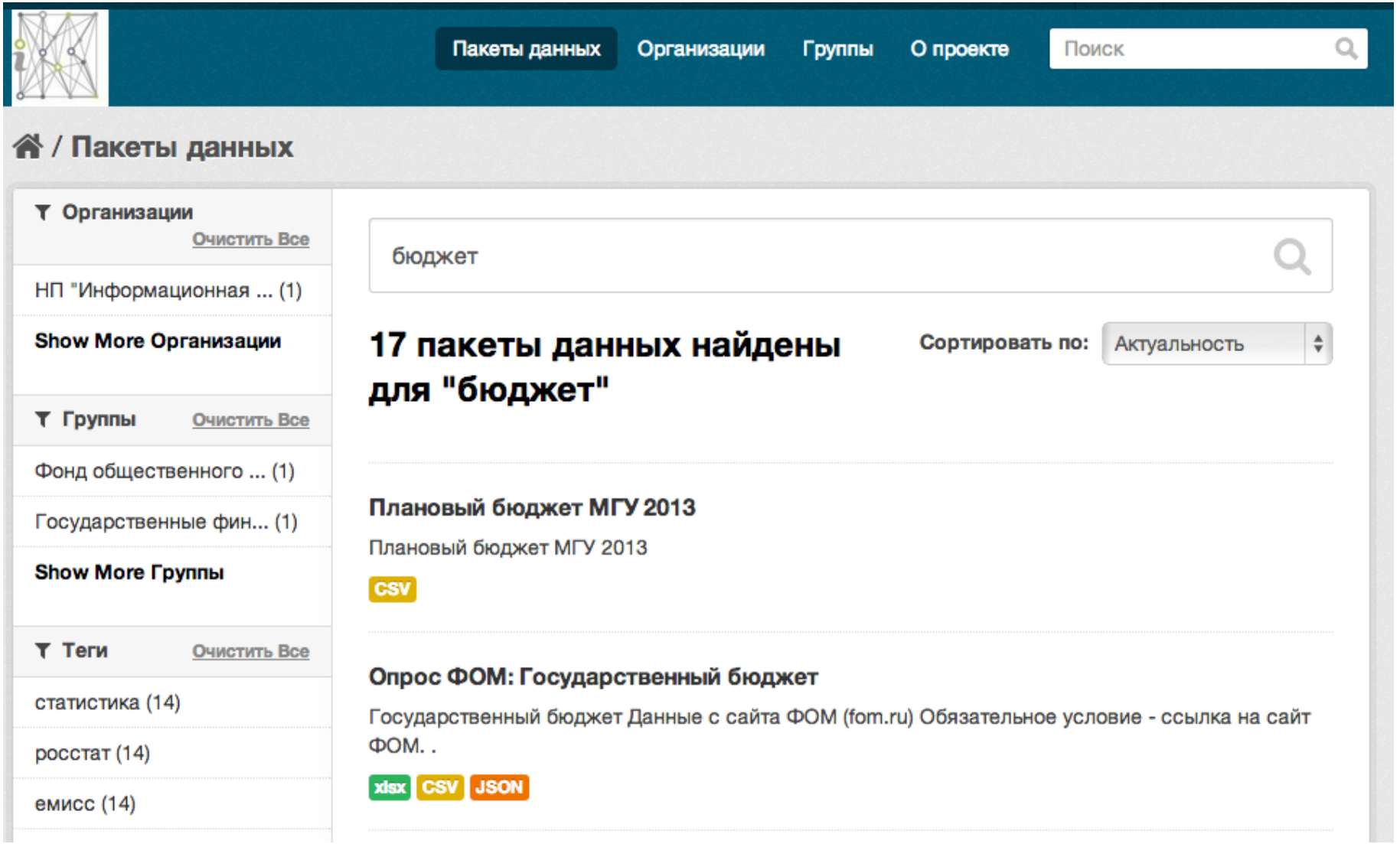
List matching countries to continents

CSV

Country Statistics

Over 80,000 stats covering 230+ countries. The spreadsheet is complementary to researchers/scholars.


Search Data.Gov
DATA.GOV
DATA
TOPICS
APPLICATIONS
DEVELOPERS
CONTACT
DATA CATALOG
/ Datasets
Organizations
Interactive Datasets
Federal datasets are subject to the U.S. Federal Government Data Policy. Non-federal participants (e.g., universities, organizations, and tribal, state, and local governments) maintain their own data policies. Data policies influence the usefulness of the data.
Filter by location
Clear
Enter location...
Map data CC-BY-SA by OpenStreetMap
Tiles by MapQuest
Dataset Type
A-Z
1-9
Clear All
geospatial (1882)
oil
3,151 datasets found for "oil"
Order by:
Relevance
Datasets ordered by Relevance
You are searching in catalog.data.gov. Show results in entire site.
BLM: Oil & Gas
Federal Geographic Data Committee — BLM oil and gas competitive lease sales
ArcGIS Map Service
ArcGIS Map Preview
Federal
Oil Gas Well (point)
State of Arkansas — This dataset represents the location and description of oil and gas wells within the State of Arkansas. ACCURACIES VARY ON METHOD OF COLLECTION. All information...
HTML
HTML
HTML
State


Пакеты данных
Организации
Группы
О проекте
Поиск
/ Пакеты данных
Организации
Очистить Все
НП "Информационная ... (1)
Show More Организации
Группы
Очистить Все
Фонд общественного ... (1)
Государственные фин... (1)
Show More Группы
Теги
Очистить Все
статистика (14)
росстат (14)
емисс (14)
бюджет
17 пакеты данных найдены для "бюджет"
Сортировать по:
Актуальность
Плановый бюджет МГУ 2013
Плановый бюджет МГУ 2013
CSV
Опрос ФОМ: Государственный бюджет
Государственный бюджет Данные с сайта ФОМ (fom.ru) Обязательное условие - ссылка на сайт ФОМ. .
xlsx
CSV
JSON

Web 2.0 Applications and Web APIs

- A multitude of Web-based applications has sprung up which enable users to share information.
- These applications form separate data spaces that are only partly accessible via the Web
  - HTML interfaces
  - Web APIs

Example: Facebook

Users (September 2012)

- 1 billion monthly active users
- including 600 million mobile users
- 140.3 billion friend connections
- 1.13 trillion likes since launch in February 2009
- 219 billion photos uploaded
- 17 billion location-tagged posts, including check-ins

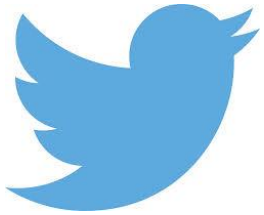
Data Volume

- over 100 Petabyte
- inluding profile data, communication, usage logs, ...

Sources

- https://s3.amazonaws.com/OneBillionFB/Facebook+1+Billion+Stats.docx
- http://www.technologyreview.com/featuredstory/428150/what-facebook-knows

Example: Twitter

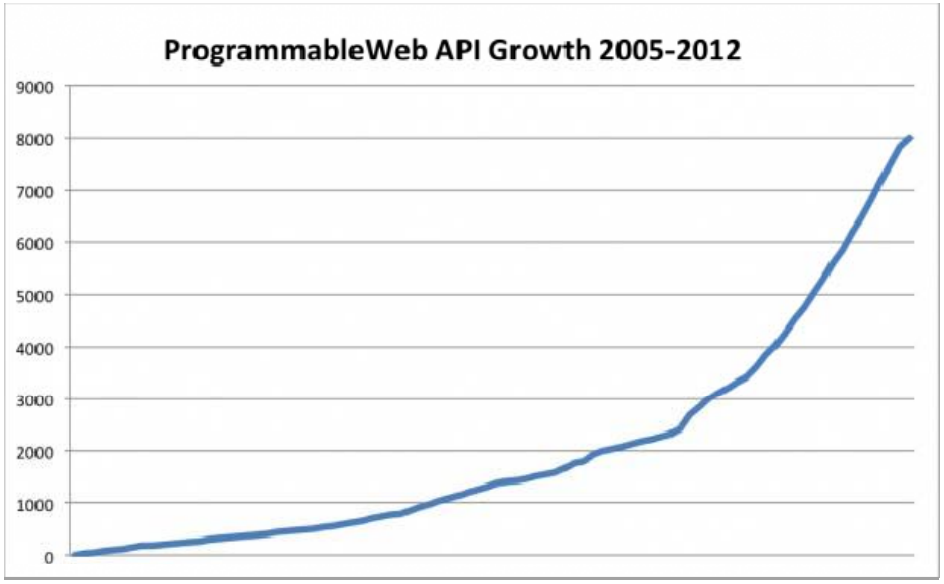


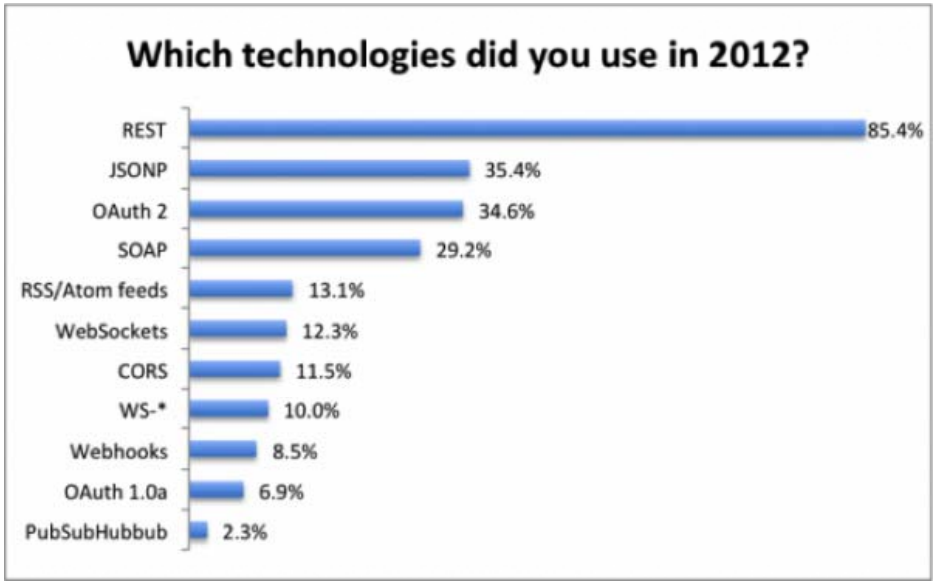
- 200 million – Monthly active users on Twitter, passed in December.
- 819,000+ – Number of retweets of Barack Obama's tweet "Four more years", the most retweets ever.
- 327,452 – Number of tweets per minute when Barack Obama was re-elected, the most ever.
- 729,571 – Number of messages per minute when the Chinese microblogging service Sina Weibo saw 2012 finish and 2013 start.
- 9.66 million – Number of tweets during the opening ceremony of the London 2012 olympics.
- 175 million – Average number of tweets sent every day throughout 2012.
- 37.3 years – Average age of a Twitter user.
- 307 – Number of tweets by the average Twitter user.
- 51 – Average number of followers per Twitter user.
- 163 billion – the number of tweets since Twitter started, passed in July.
- 123 – Number of heads of state that have a Twitter account.

http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/

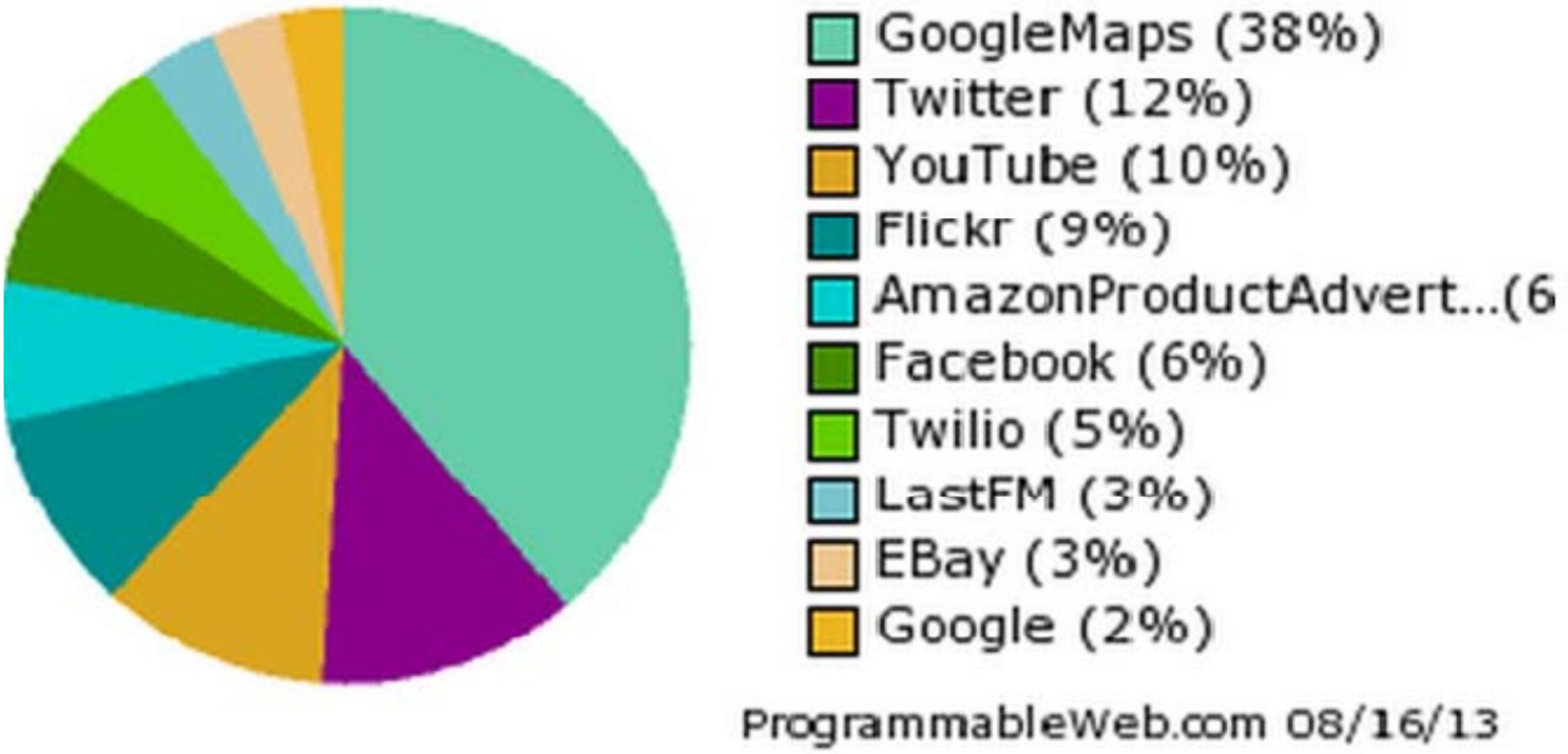
Web APIs

- Provide limited access to the collected data
  - restricted to specific queries (canned queries)
  - restrictred number of queries
- ProgrammableWebAPICatalog
  - lists over 9000 Web APIs
  - list over 6800 Mashups

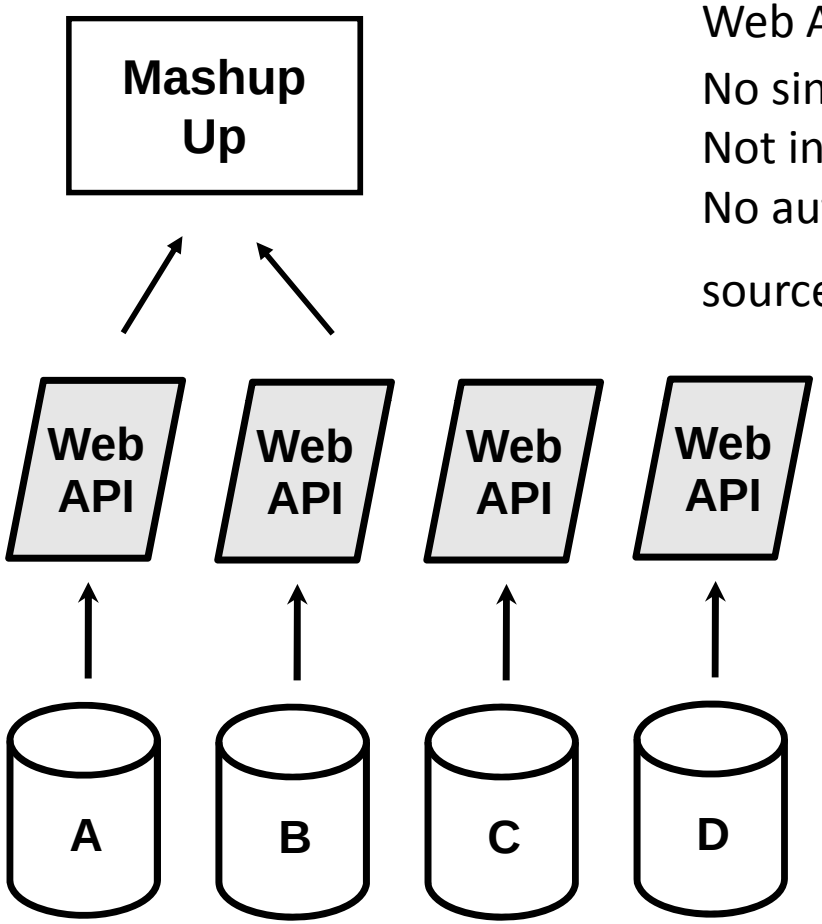




Most Popular Web API



Mashups

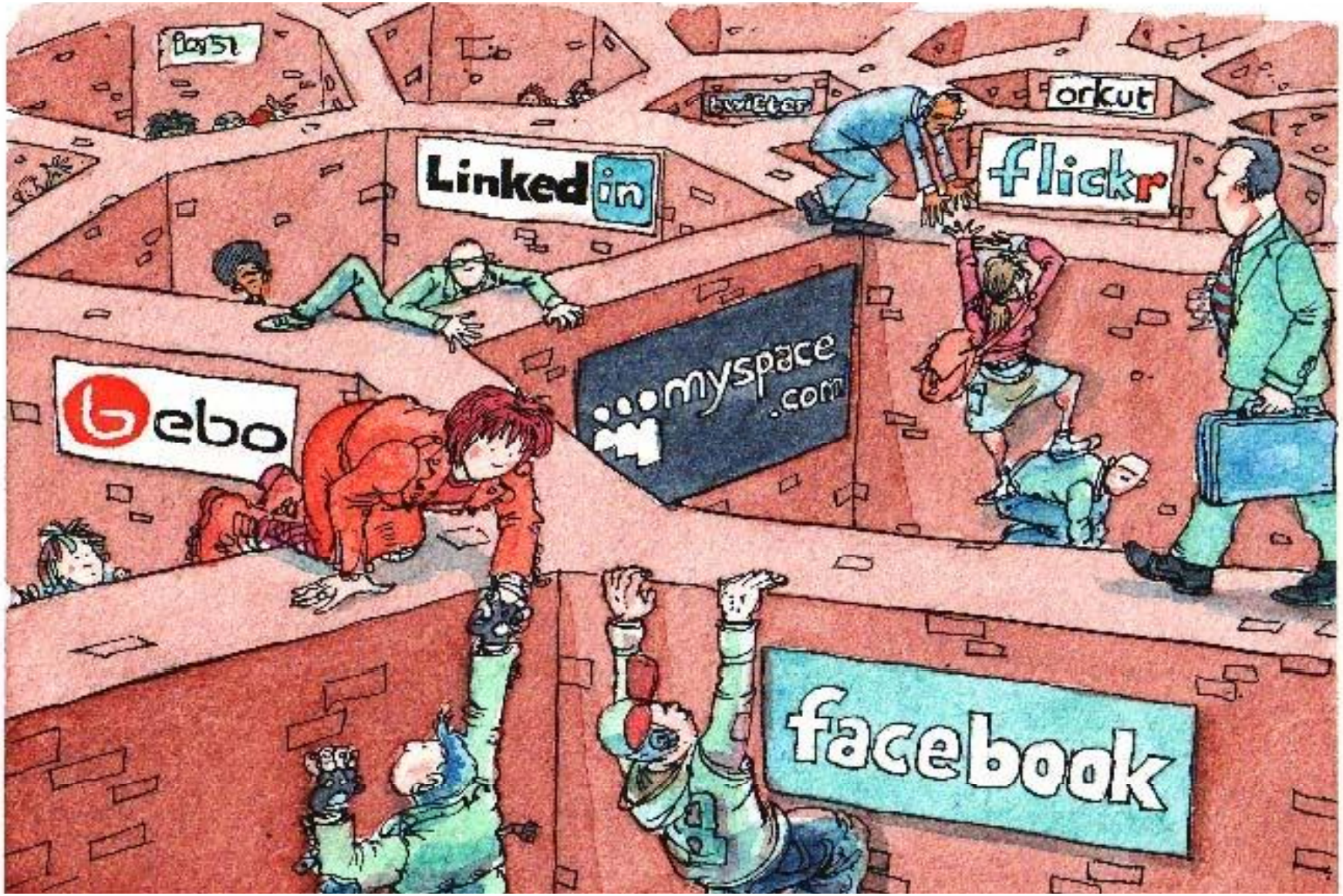


Web APIs expose proprietary interfaces

No single global data space

Not index-able by generic crawlers

No automatic discovery of additional data sources


LinkedIn
twitter
orkut
flickr
bebo
myspace
.com
facebook

Twitter API

View | What links here

Updated on Tue, 2014-01-21 23:42

API version 1.1

These libraries, while not necessarily tested by Twitter, should support Twitter API v1.1.

Want your library to be included in this index or need to update the details we have? Submit your library for inclusion!

Libraries built and maintained by Twitter

Java

- hbc — A Java HTTP client for consuming Twitter's Streaming API

Libraries built for the Twitter Platform

Multi-platform

- Temboo — *by @temboo* — Framework for working with Twitter via many platforms including iOS, Android, Java, PHP, Python, Ruby, and Node.js

ASP

- asptwitter *by @timacheson* — "the simplest possible way to implement Twitter within a classic ASP website" -- now supports API v1.1

Twitter API

C++

- twitcurl *by @swatkatsrants* — Twitcurl is a C++ twitter API library based on cURL. Twitcurl supports v1.1 twitter APIs and SSL.

Clojure

- twitter-api *by @adamjwynne and @peat* (announcement)

ColdFusion

- MonkehTweet Twitter API *by @coldfumonkeh*

.NET

- LINQ2Twitter by *@joemayo* (examples)
- Spring.NET Social extension for Twitter *by SpringSource* — A Spring.NET Social extension with connection support and an API binding for Twitter.
- TweetSharp *by @danielcrenna* — A .net library for Twitter API access
- Tweetinvi *maintained by Linvi* — a Twitter .Net C# API which has for mission to simplify the development of application for Twitter in C#. The streaming API has been used on research projects and collected around 3.2 million tweets a day. The twitter API has been created to be easy to implement new functionality and currently provide access to most of the REST 1.1 functionalities. (documentation<a>)
- Crafted,Twitter by *@martbrow* — A caching v1.1 API compatible solution - with implementations for both ASP.Net Web Forms and MVC. Making it easy to include tweets in your website.

Go

- twittergo *by @kurrik* (examples) — a library for accessing Twitter's REST API. Supports v1.1 and app-only auth.
- Anaconda *by @chimeracoder* — a simple, transparent Go package for accessing version 1.1 of the Twitter API. API queries are provided as methods returning native Go structs which can be used immediately, with no need for type assertions.

Twitter Streaming

- We used hbc library to collect tweets

```
/** Declare the host you want to connect to, the endpoint, and authentication (basic auth or oauth)
Hosts hosebirdHosts = new HttpHosts(Constants.STREAM_HOST);
StreamingEndpoint endpoint = new StatusesFilterEndpoint();
// Optional: set up some followings and track terms
List<Long> followings = Lists.newArrayList(1234L, 566788L);
List<String> terms = Lists.newArrayList("twitter", "api");
endpoint.followings(followings);
endpoint.trackTerms(terms);

// These secrets should be read from a config file
Authentication hosebirdAuth = new OAuth1("consumerKey", "consumerSecret", "token", "secret");
```

Creating a client

```
ClientBuilder builder = new ClientBuilder()
  .name("Hosebird-Client-01")
  .hosts(hosebirdHosts)
  .authentication(hosebirdAuth)
  .endpoint(hosebirdEndpoint)
  .processor(new StringDelimitedProcessor(msgQueue))
  .eventMessageQueue(eventQueue);

Client hosebirdClient = builder.build();
// Attempts to establish a connection.
hosebirdClient.connect();
```

Listening to a message queue

```
// on a different thread, or multiple different threads....
while (!client.isDone()) {
  String msg = msgQueue.take();
  something(msg);
  profit();
}
```

REST API

Returns a collection of relevant Tweets matching a specified query.

Resource URL
https://api.twitter.com/1.1/search/tweets.json

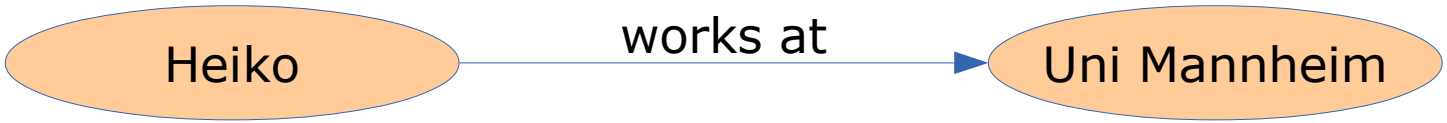
Example Request
GET
https://api.twitter.com/1.1/search/tweets.json?q=%23freebandnames&since_id=24012619984051000&max_id=250126199840518145&result_type=mixed&count=4

Resource Information

Rate Limited?	Yes
Requests per rate limit window	180/*user* 450/*app*
Authentication	Required
Response Formats	json
HTTP Methods	GET
Resource family	search
Response Object	Tweets
API Version	v1.1

Resource Description Framework (RDF)

- A W3C Standard (2004)
- Description of arbitrary data
- “Everything is a resource”
- View 1: Sentences in Subject-Predicate-Object form
  - „Heiko works at University of Mannheim.”
- View 2: Directed graphs with edge labels



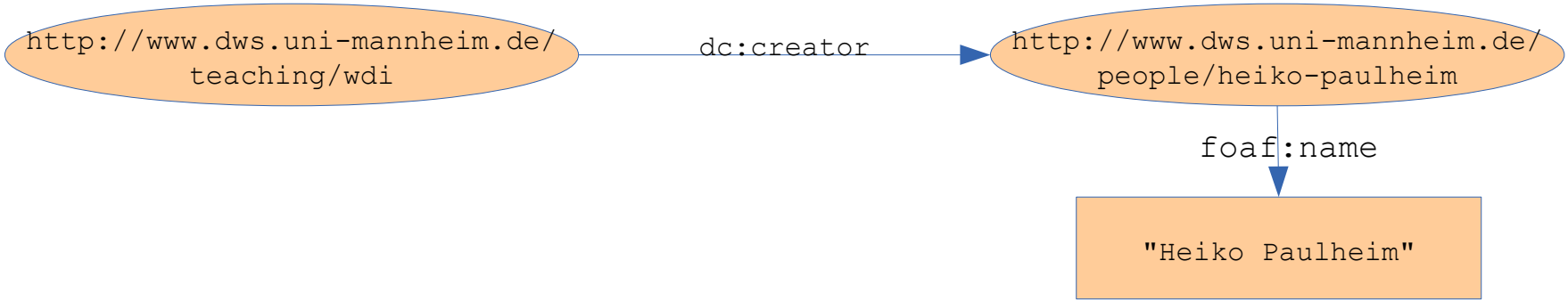
RDF Building Blocks

- Resources
  - in general, everything (a person, a place, a web site...) is a resource
  - identified by a URI
  - may have one or more types (e.g.: “Person”)
- Literals
  - are data values, e.g., strings and integers
  - may only be objects, not subjects (i.e., no outgoing edges)
  - may have a data type or a language tag (but not both)
- Properties (Predicates)
  - Connect resources to other resources
  - Connect resources to literals

Resource vs. Literal

- A literal is a simple value
  - cannot be a subject
  - i.e., at a literal, a graph always ends

- A resource may be the subject of another statement



Data Types in RDF

- Examples:
  - :Muenchen :hasName "München"@de .
    :Muenchen :hasName "Munich"@en .
    :Muenchen :hasPopulation "1356594 "^^xsd:integer .
    :Muenchen :hasFoundingYear "1158-01-01"^^xsd:date .
- Be careful: there are no default data types
- i.e., the following three literals are different:
  - "München"
  - "München"@de
  - "München"^^xsd:string .

RDF Triple Notation

- A W3C Standard (2004)
- Triples have a subject, a predicate, and an object
- All triples in a document are *unordered*
- Simple triple:

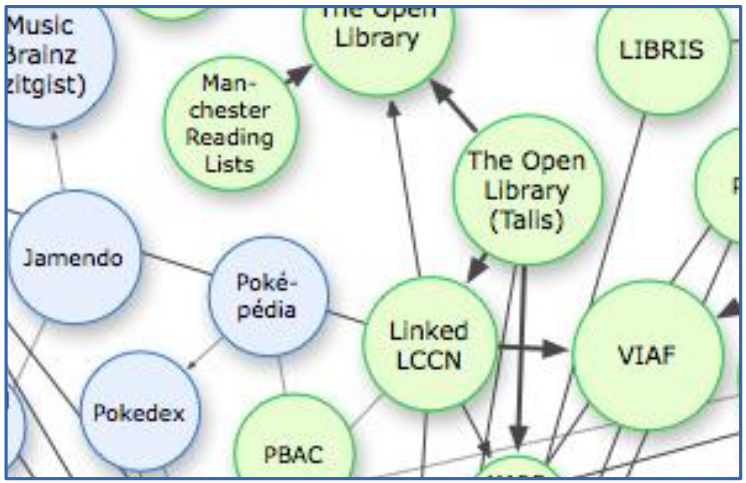
  <http://www.dws.uni-mannheim.de/teaching/wdi>
  <http://purl.org/dc/elements/1.1/relation>
  <http://www.w3.org/2001/sw/> .
- Literal with language tag:

  <http://www.dws.uni-mannheim.de/teaching/wdi>
  <http://purl.org/dc/elements/1.1/subject>
  "Web Data Integration"@en .
- Literal with type:

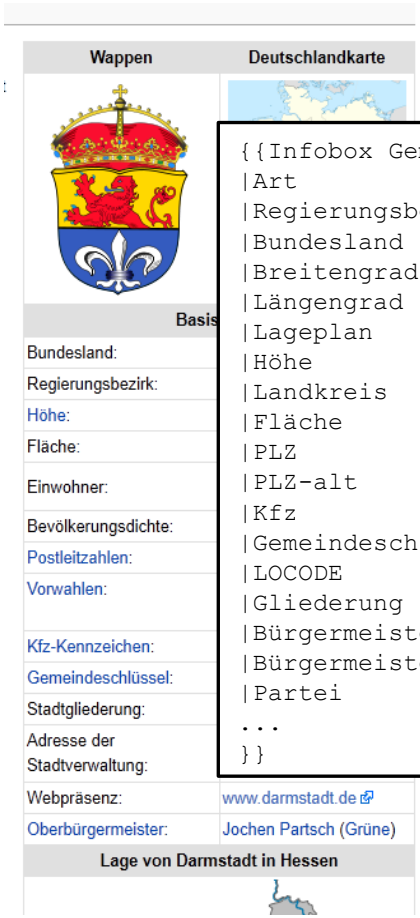
  <http://www.dws.uni-mannheim.de/teaching/wdi>
  <http://www.dws.uni-mannheim.de/teaching/credits>
  "6"^^<http://www.w3.org/2001/XMLSchema#integer> .

RDF Example: Dbpedia

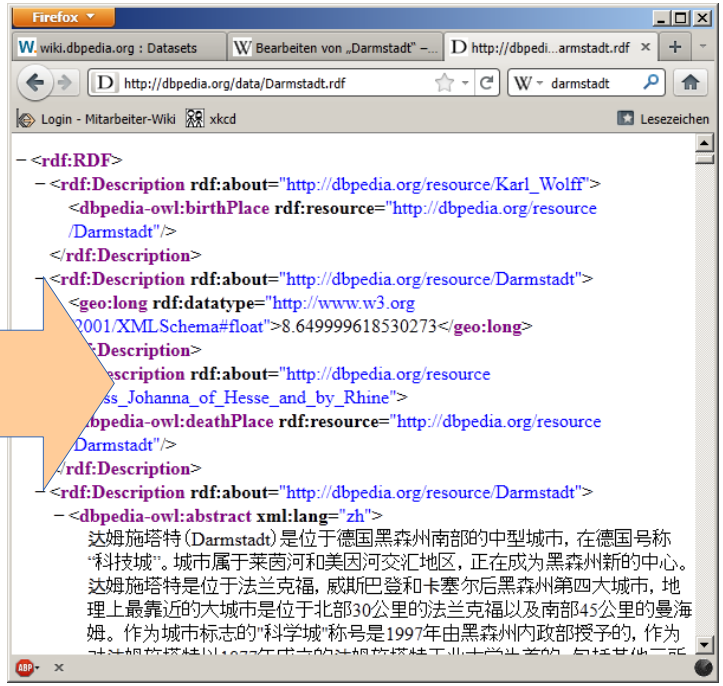
- Cross domain knowledge on millions of entities
- 500 million triples
- Linked to another 100 datasets
  - The most strongly linked data set in LOD



RDF Example: Dbpedia




```
{{Infobox Gemeinde in Deutschland
|Art                = Stadt
|Regierungsbezirk   = Darmstadt
|Bundesland         = Hessen
|Breitengrad        = 49/52//N
|Längengrad         = 08/39//E
|Lageplan           = Hesse DA(city).svg
|Höhe               = 144
|Landkreis          =
|Fläche             = 122.24
|PLZ                = 64283-64297
|PLZ-alt            = 6100
|Kfz                = DA
|Gemeindeschlüssel  = 06411000
|LOCODE             = DE DAR
|Gliederung         = 9 [[Stadtteil]]e
|Bürgermeister      = [[Jochen Partsch]]
|Bürgermeistertitel= Oberbürgermeister
|Partei             = Grüne
...
}}
```

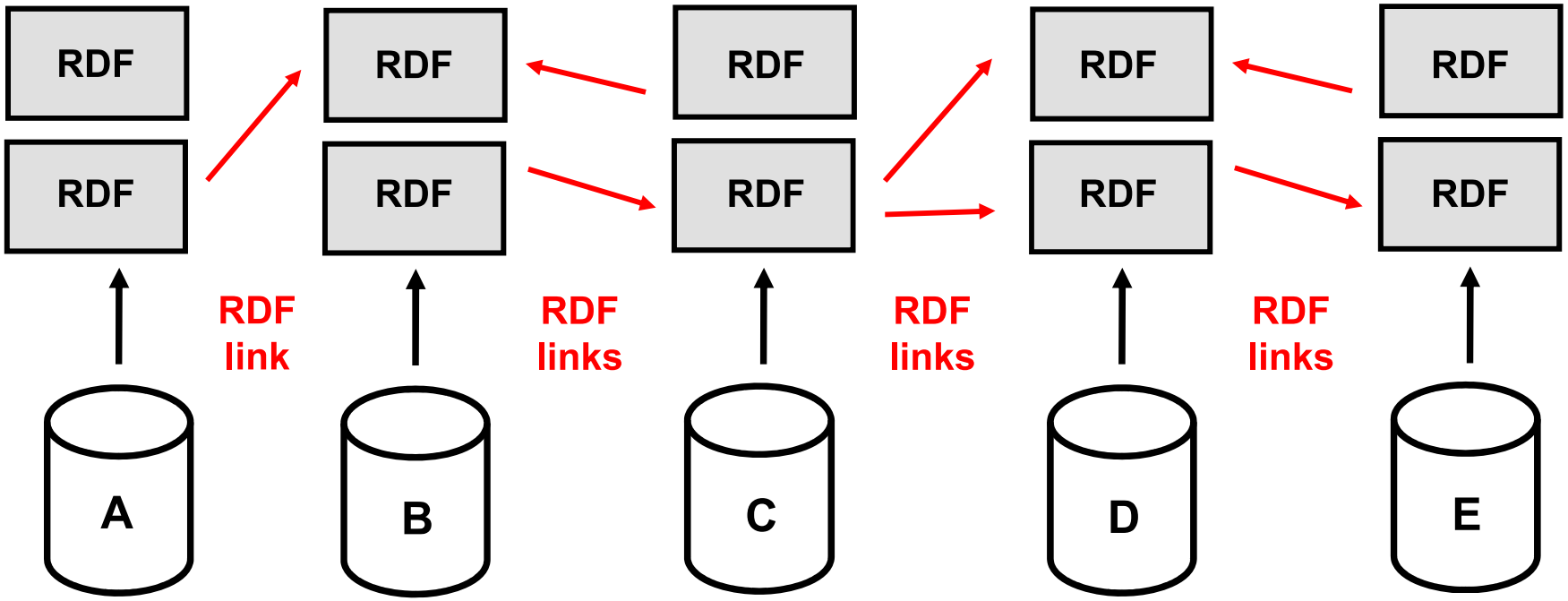


RDF Example: Dbpedia

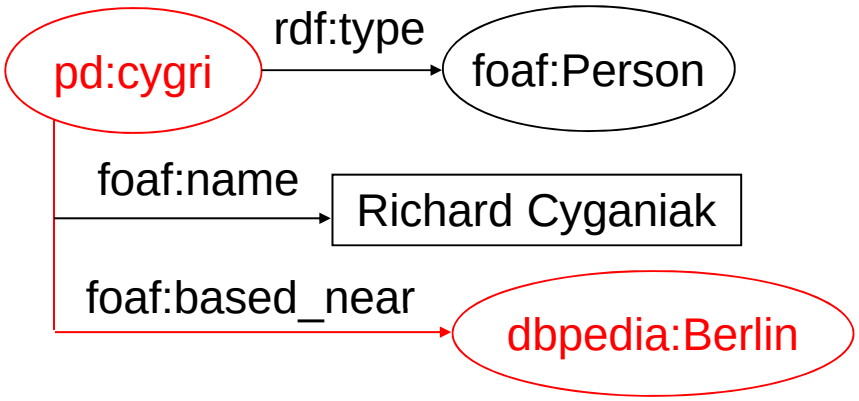
- Data from various infoboxes
- Redirects and disambiguations
- Cross-language links
- Links to other web sites
- Abstracts in various languages
- Type information according to various schemas
  - yet to come

Linked Data

- Extend the Web with a single global data graph
  - by using RDF to publish structured data on the Web
  - by setting links between data items within different data sources.



Entities are identified with URIs

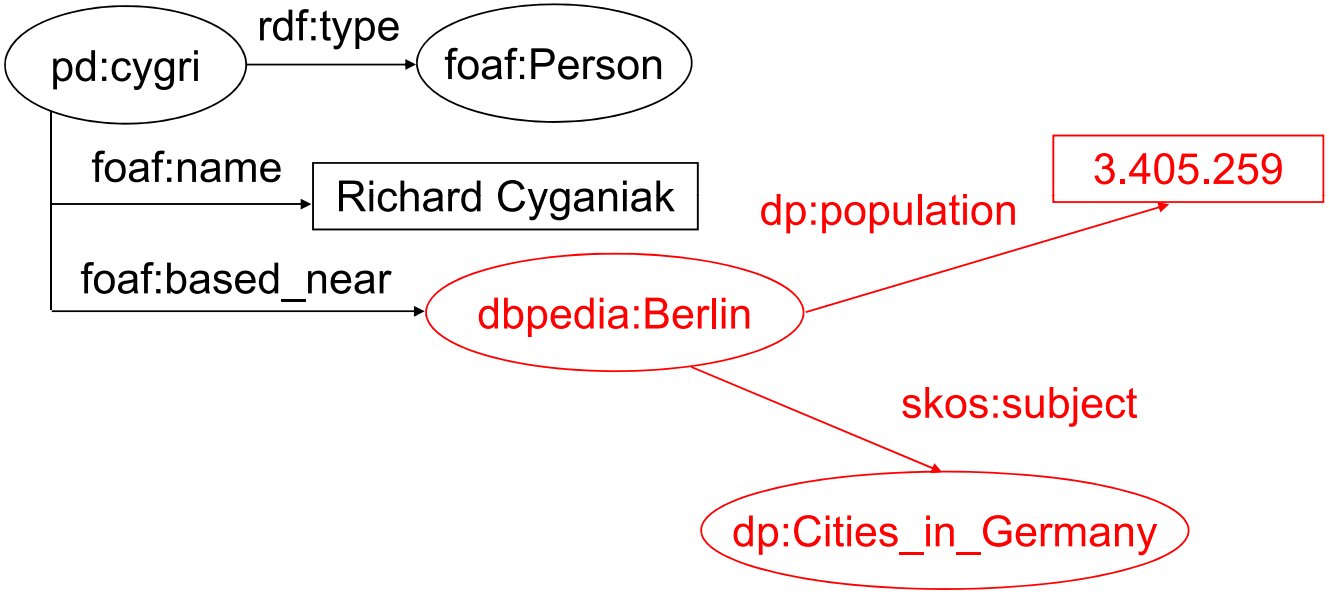


HTTP URIs take the role of global primary keys.

pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri

dbpedia:Berlin = http://dbpedia.org/resource/Berlin

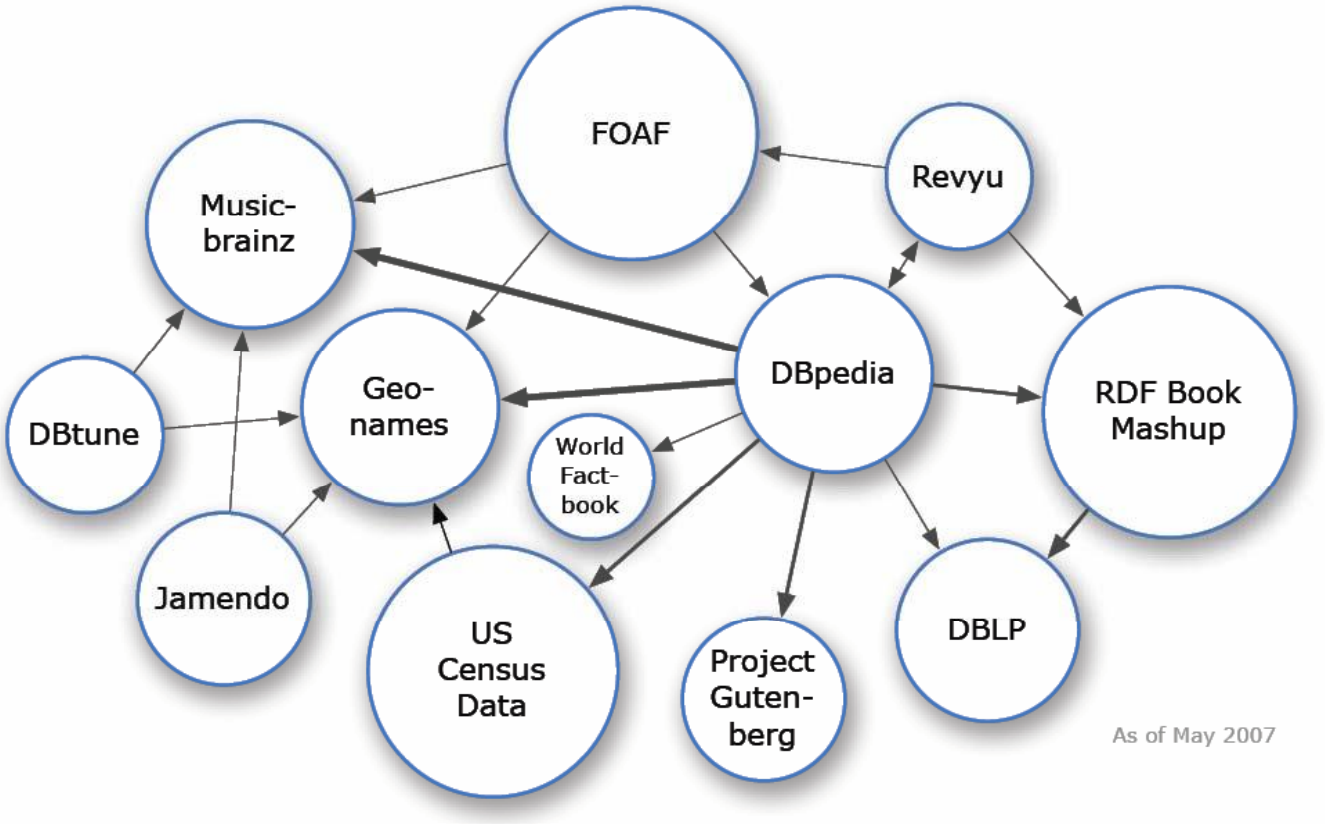
URIs can be looked up on the Web



By following RDF links application scan

- navigate the global data graph
- discover new data sources

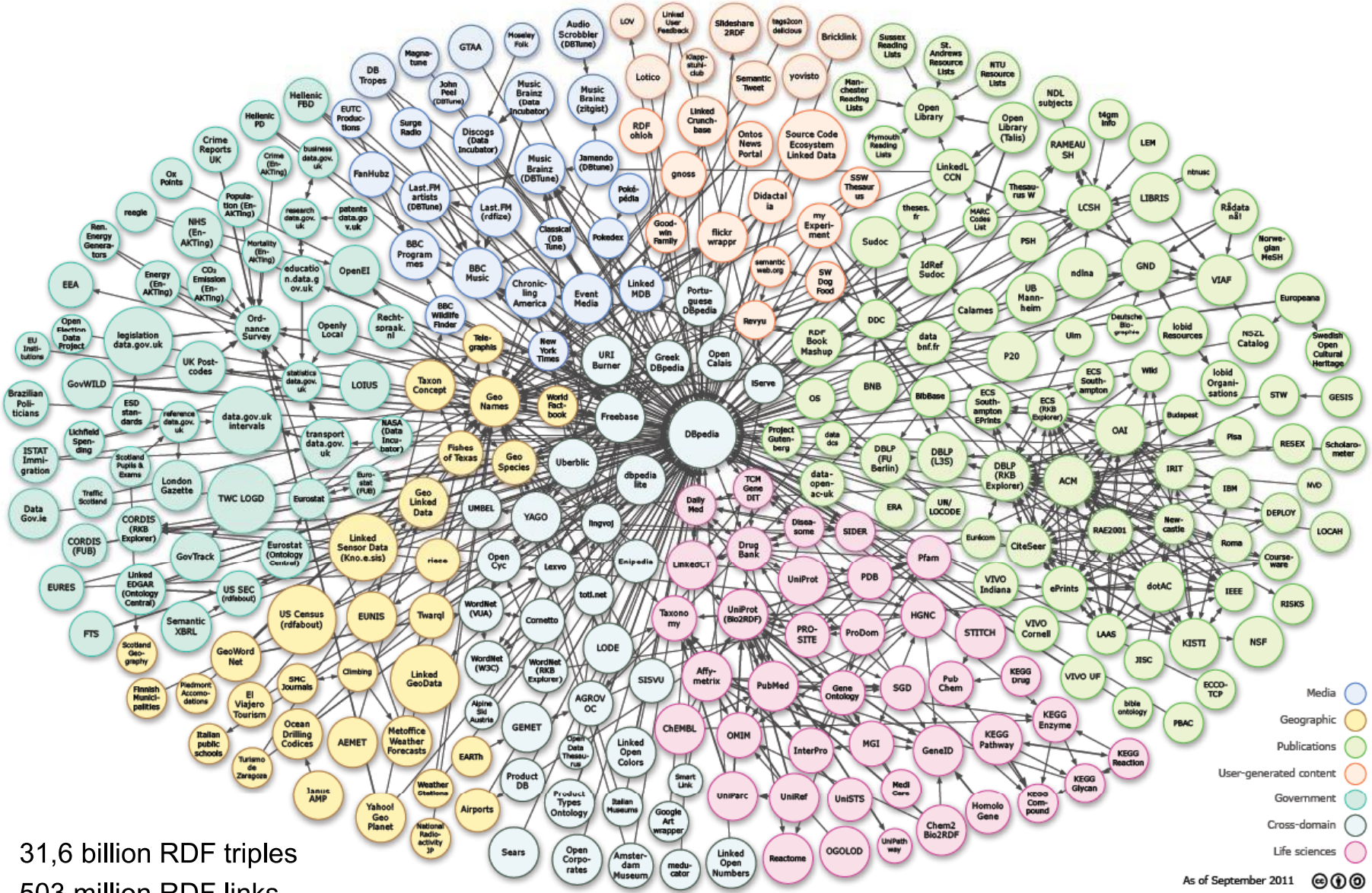
LOD Datasets: May 2007



Over 500 million RDF triples

 Around 120,000 RDF links between data sources

LOD Datasets: September 2011



- 31,6 billion RDF triples
- 503 million RDF links

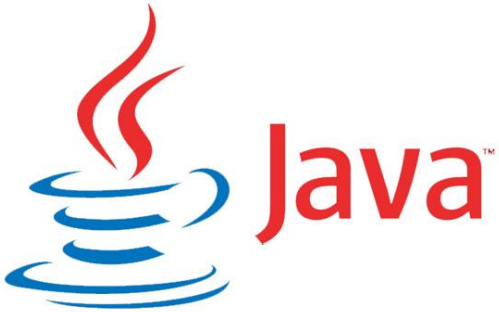
Distribution by topic

Domain	Data Sets	Triples	Percent	RDF Links	Percent
Media	25	1,841,852,061	5.82 %	50,440,705	10.01 %
Geographic	31	6,145,532,484	19.43 %	35,812,328	7.11 %
Government	49	13,315,009,400	**42.09 %**	19,343,519	3.84 %
Library	87	2,950,720,693	9.33 %	139,925,218	**27.76 %**
Cross-domain	41	4,184,635,715	13.23 %	63,183,065	12.54 %
Life sciences	41	3,036,336,004	9.60 %	191,844,090	**38.06 %**
User content	20	134,127,413	0.42 %	3,449,143	0.68 %
SUM	**295**	**31,634,213,770**		**503,998,829**	

Get Linked Data

- Download the Billion Triples Challenge Dataset
  - 1.4 billion triples (17 GB gzipped)
  - crawled from the public Web of Linked Data in May/June 2012
  - http://km.aifb.kit.edu/projects/btc-2012/
- Download the Sindice Dump
  - 12 billion triples (164GB gzipped, ~1,16TB uncompressed)
  - Linked Data, RDFa, Microdata, Microformat crawled 2009-2011
  - http://data.sindice.com/trec2011/download.html

Web Data Formats



Simple Tables – CSV files

- Not particularly a web data format
- But quite widely used (also on the web)
- Data exported from RDBMs and spreadsheet applications
- A CSV (comma separated values) file encodes a table
- First line is often used as header

Example:

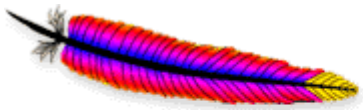
```
firstname,lastname,matriculation,birthday
thomas,meyer,3298742,15.07.1988
lisa,müller,43287342,21.06.1989
    ...
```

Processing CSV Files

- Apache Commons CSV
- Provides a simple API

```
Reader in = new FileReader("data/data.csv");
Iterable<CSVRecord> parser = CSVFormat.EXCEL.parse(in);
for (CSVRecord record : parser) {
  if (record.getRecordNumber()>1) {
    String firstname = record.get(0);
    String lastname  = record.get(1);
    ...
  }
}
```

skip headline



Apache Commons™
http://commons.apache.org/

Processing CSV Files

- There is no particular query language for CSV files
- But you can, e.g.,...
  - load a CSV file into a database table
  - and use SQL
- Example MySQL:

```
LOAD DATA LOCAL INFILE 'data.csv' INTO TABLE persons;
     SELECT * FROM persons WHERE lastname LIKE '%meyer%';
```

JavaScript Object Notation (JSON)

- JavaScript: a popular programming language on the web
- Embedded in HTML
- Originally:
  - Simple interactions (e.g., image exchange on mouse over)
- Nowadays:
  - Also for complex applications
  - Ajax (Asynchronous JavaScript and XML)

JavaScript Object Notation (JSON)

- Basics:
  - Objects as they are noted in JavaScript
  - Objects are enclosed in curly brackets { … }
  - Data is organized in key value pairs

this can only work because JavaScript is a dynamically typed programming language

- Example:

```
var obj = { "firstname" : "John" ,
            "lastname" : "Smith" ,
            "age" : 46
          }
```

- Simple processing in JavaScript:

```
var obj = eval(jsonString) ;
var name = obj.firstname +  " " + obj.lastname ;
```

JavaScript Object Notation (JSON)

- Nested objects are possible:

```
{ "firstname" : "John" ,
  "lastname" : "Smith" ,
  "age" : 46 ,
  "employer" {
    "name" : "Technology Inc.
    "address" : {
      "street" : "Main St." ,
      "number" : 14,
      "city" : Smalltown
    }
  }
}
```

```
<firstname>John</firstname>
<lastname>Smith</lastname>
<age>46</age>
<employer>
   <name>Technology Inc.</na
   <address>
     <street>Main St.</stree
     <number>14</number>
     <city>Smalltown</city>
   </address>
</employer>
```

JavaScript Object Notation (JSON)

- JSON is a lot like XML
  - Treestructure
  - Opening/closingtags/brackets
- Differences
  - JSON is *not* a standard (but widely used)
  - More compact notation than XML
  - No id/ref – JSON data is *strictly* tree shaped
  - Less data types (only strings and numbers)
  - No schema*
  - No query language*

*although people are working on that

Processing JSON in Java

- Things were easy in JavaScript:

```
var obj = eval(jsonString) ;
var name = obj.firstname + " " + obj.lastname ;
```

- But that only works in dynamically typed programming languages
- Java uses static typing
  - thus, we have to define the classes in advance

- And it's not built in
  - we need a particular library
  - e.g., gson

Processing JSON in Java

- Class definition

```
public class Person {
  private String firstname;
  private String lastname;
  private int age;
}
```

```
{ “firstname” : “John” ,
  “lastname” : “Smith” ,
  “age” : 46
}
```

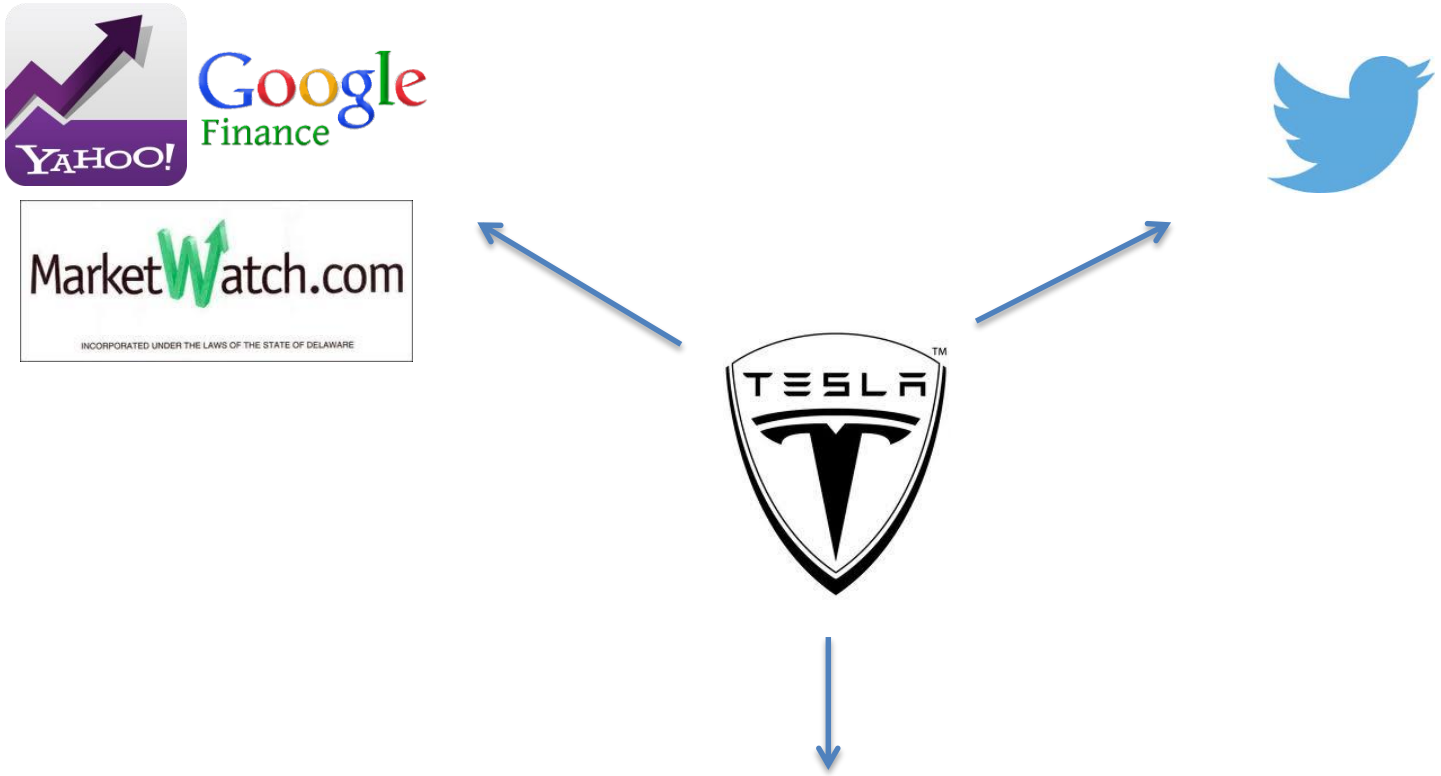
- Object serialization

```
Person person;
String json = gson.toJson(obj);
```

- Object deserialization

```
Person person = gson.fromJson(jsonString,
                              Person.class);
```

Motivating example



Motivating example

collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. ... We find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA

Johan Bollen et al paper Twitter mood predicts the stock market (Oct 2010)

Motivating example

Prices

Date	Open	High	Low	Close	Volume	Adj Close*
Mar 10, 2014	242.70	243.00	236.06	238.84	7,728,100	238.84
Mar 7, 2014	252.94	254.85	244.41	246.21	7,812,300	246.21
Mar 6, 2014	254.14	257.50	249.45	252.94	7,361,100	252.94
Mar 5, 2014	256.72	256.99	251.80	252.66	5,935,700	252.66
Mar 4, 2014	258.48	260.00	252.83	254.84	8,745,600	254.84
Mar 3, 2014	237.26	251.65	234.99	250.56	13,089,300	250.56
Feb 28, 2014	249.65	252.68	242.55	244.81	14,589,800	244.81
Feb 27, 2014	261.25	261.90	248.33	252.54	17,945,800	252.54
Feb 26, 2014	258.58	265.00	247.50	253.00	24,604,600	253.00
Feb 25, 2014	230.00	259.20	228.45	248.00	32,681,700	248.00

lol, had $TSLA june 65's at $2.6 - sold $21, now $42 - incredible to look back

$TSLA is probably an awesome buy right here at $110.

Here's What Sent Tesla's Shares to Record Highs t.co/6ASSzAA3o5#Tesla $TSLA

Elon Musk

picture: 

[1]  [20]

given name: Elon [3]

family name: Musk [3]

comment: Voir toute l'actualité de Elon Musk [1]

Elon Musk is the illustrious entrepreneur and originator of Tesla Motors. Tesla Motors is the electric automobile company that intends to execute a $178 million IPO. However, data floated up in recent times that Elon Musk had run out [11]

lol, had $TSLA june 65's **at $2.6 - sold $21, now $42** - **incredible to look back**

$TSLA is **probably** an **awesome buy** right here at $110.

Here's What Sent Tesla's Shares to Record Highs t.co/6ASSzAA3o5#Tesla $TSLA

Integrated entity

company: {

name : “ Tesla Motors”;

owner: “Elon Musk”;

CEO: “Elon Musk”;

financial_stats: { sharpe_ratio: 0.17;

beta: 3.9;

..}

tweets: [..];

positive tweets : 0.7;

recommendation: strong_buy}

Analysis example

- Collect data on thousands of companies
- Find the diversified portfolios of securities;
- Choose the one to invest in according to financial stats and twitter recommendations;

Essentially all models are wrong, but some are useful

That if the model is going to be wrong anyway, why not see if you can get the computer to quickly learn a model from the data, rather than have a human laboriously derive a model from a lot of thought.

[Banko and Brill, 2001]

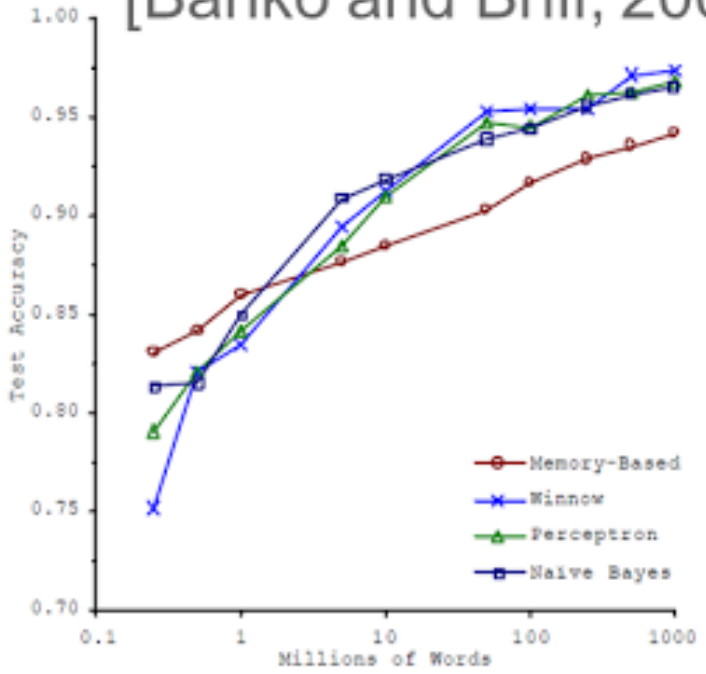


Figure 1. Learning Curves for Confusion Set Disambiguation

The Unreasonable Effectiveness of Data

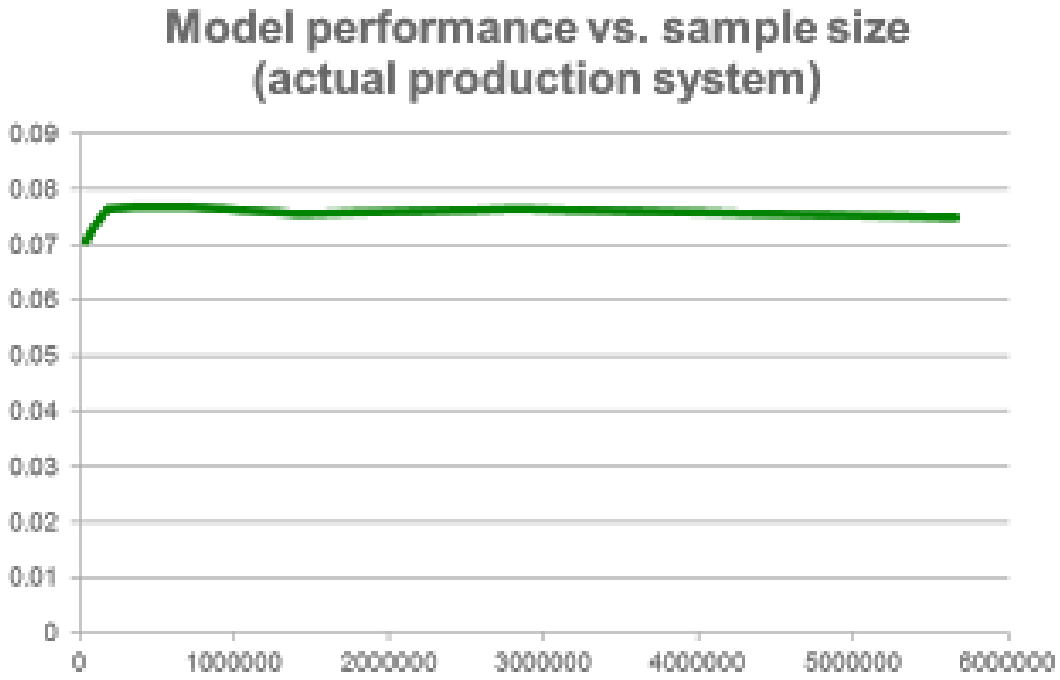
Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Variance or Bias?

- model that is too complicated for the amount of data we have - *high variance* (leads to model overfitting)
  - High variance problems can be addressed by reducing the number of features, and by increasing the number of data points.

Variance or Bias?

- We might have a model that is too simple to explain the data we have - *high bias* (adding more data will not help)



Theory has not ended, it is expanding into new forms. Sure, we all love succinct theories like F = m a. But social science domains and even biology appear to be inherently more complex than physics. Let's stop expecting to find a simple theory, and instead embrace complexity, and use as much data as well as we can to help define (or estimate) the complex models we need for these complex domains.

P. Norvig