

# Architectures for massive data management

**Ioana Manolescu**

INRIA Saclay & Ecole Polytechnique

[ioana.manolescu@inria.fr](mailto:ioana.manolescu@inria.fr)

<http://pages.saclay.inria.fr/ioana.manolescu/>

M2 Data and Knowledge  
Université de Paris Saclay

# Who am I

- Senior researcher (DR) at INRIA Saclay
- Part-time professor at Ecole Polytechnique
- I work on **algorithms and systems** for **efficient management** of **complex data**
  - **Complex data**: complex structure (trees, graphs...) and/or complex semantics (meaning)
  - **Data management**: queries, updates
  - **Efficiency**: fast evaluation; low resource usage
- PhD: INRIA Rocquencourt, France, 2001
- Master: Ecole Normale Supérieure and U. Paris 6
- Undergraduate: Polytechnic U. Bucharest, Romania

# What is your background in data management?

Please fill in the **questionnaire**

# Course goal

1. Discuss the main **characteristics (dimensions)** of massive data management platforms
  - « Big Data »
2. Present the main **classes** of such systems, according to the above dimensions
3. Analyze **advantage/disadvantage trade-offs**
4. Introduce some **open research issues**

# Architectures to be covered

- Distributed databases
- P2P systems
  - Structured
  - Unstructured
- Key-value stores
  - Redis, DynamoDB
- Data integration systems
  - Local-as-view, global-as-view
  - Mediator systems
  - Dataspaces
  - Data lakes
- MapReduce parallelism [covered in another course]
- Large distributed tables (Spanner, F1)
- Graph databases: Neo4J, Virtuoso
- Pregel
- [Spark covered in another course]

# Course organization

- **Instructors:**

Ioana Manolescu (INRIA)

Silviu Maniu (U. Paris Sud)



- **Location:** U. Paris Sud, PUIO

- Lecture: **B109**

- Lab: **D104**

30%

- **Evaluation:** exam + lab work

# Practical matters

- You will have access to the course material in PDF format
  - At least, we will send them on the list:  
[dk-students-2017@googlegroups.com](mailto:dk-students-2017@googlegroups.com)
- Feel free to contact us
  - Ask questions during course days
  - Write (usable ☺) e-mail  
[ioana.manolescu@inria.fr](mailto:ioana.manolescu@inria.fr)

# Today's course plan

## 1. Motivation: Big Data

- Characteristics
- Applications

## 2. From databases to architectures for Big Data management

- Databases: quick recall or crash course
- What needs to change to handle Big Data?

# **MOTIVATION: BIG DATA**

# Defining Big Data: the V's

- Volume
  - Scale
- Velocity 速度
  - Speed of producing and consuming the data
- Variety
  - Very different sources and data types
- Veracity
  - Is the data correct / certain / true?

# Where does the data volume come from? (1)

- Human-produced data
  - **Web content**: Web pages, blogs, social networks, tweets...
  - **Twitter**: 7 Terabytes (1 tera =  $10^{18}$ ) per day
  - **Facebook**: 10 Terabytes per day

A blue rounded rectangular button containing the word 'facebook.' in white lowercase letters.

# Where does the data volume come from? (1)

- Human-produced data

- **Web content**: Web pages, blogs, social networks, tweets...



- **Twitter**: 7 Terabytes (1 tera =  $10^{18}$ ) per day
  - **Facebook**: 10 Terabytes per day



- Machine-produced data

- Log data from all kind of servers
  - Real world devices: payments, telecom, energy, weather, transportation, shipment...

Sensors, including on (US) highways  
and trains

Gazpar (GDF)

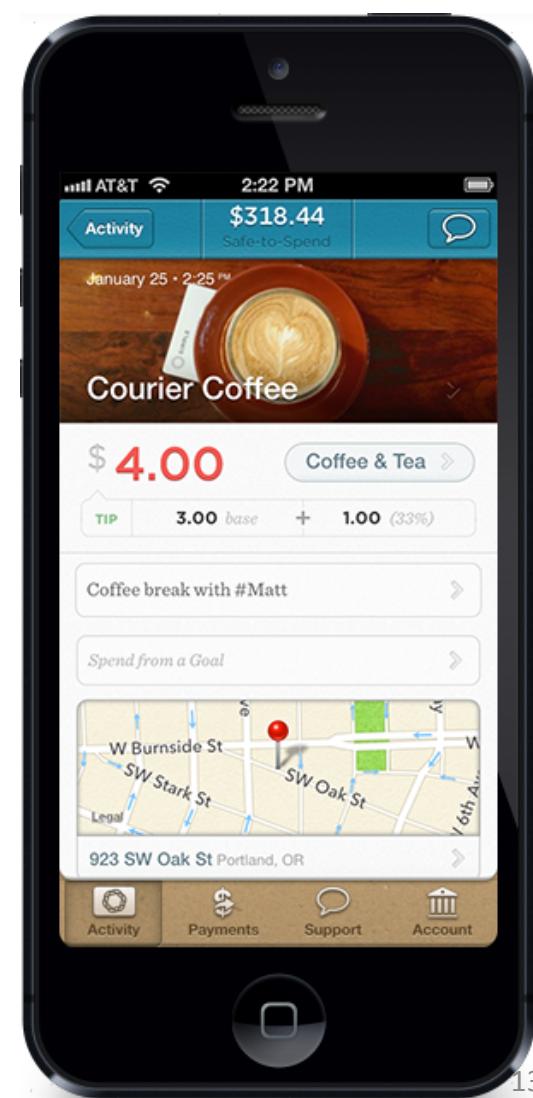


Linky (EDF)



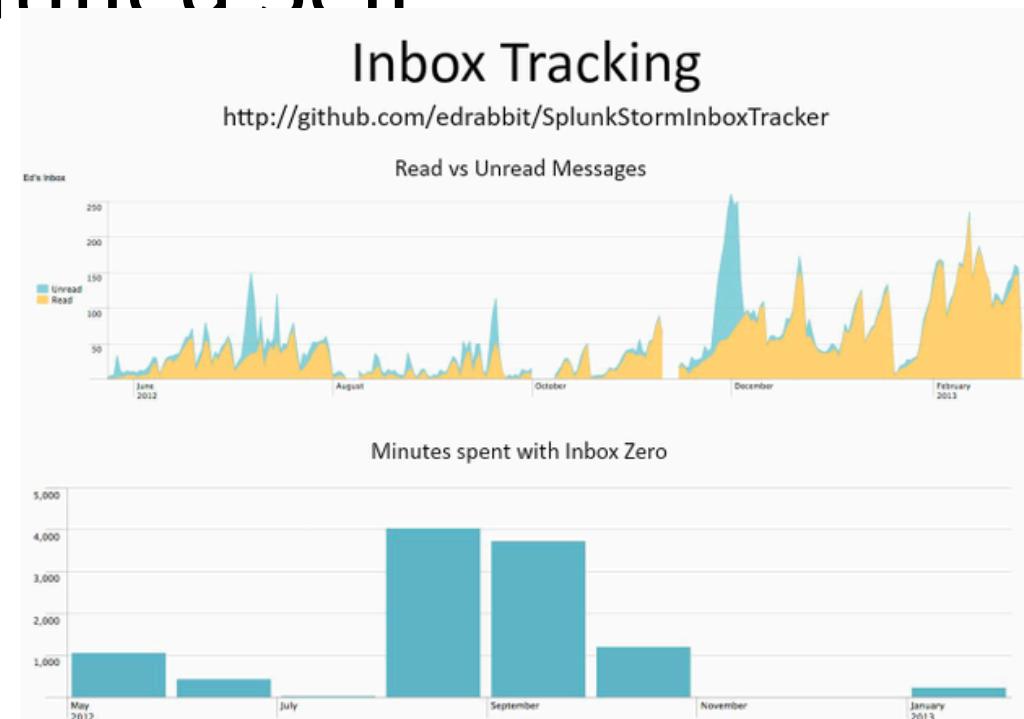
# More data than we thought: bank transactions

- The first historical **database benchmark** application
- (TPC benchmark: bank, branches, clients, accounts, tellers, transactions etc.)
  - 分行
  - 柜员
- Simple.com: bare-bones financial services + access to transaction data
  - 80 attributes per transaction
  - Banks typically use only 4 ("Date/Libellé/Euros")
    - 裸机金融服务
  - Applications to exploit the complete transaction data



# Human-produced data / Quantified Self

- “How big is your unread inbox?”
- “How did your heart rate change while exercising?”
- Social media content (including professional use), company intranets, photos etc.



# Capturing and storing human user activity

- Gordon Bell (Microsoft), 2009  
[http://edition.cnn.com/2009/TECH/09/25/total.recall.microsoft.bell/index.html?eref=rss latest](http://edition.cnn.com/2009/TECH/09/25/total.recall.microsoft.bell/index.html?eref=rss_latest)
- "video equipment, cameras and audio recorders to capture his conversations, commutes, trips and experiences. [...] SenseCam that would hang around a person's neck and automatically capture every detail of life in photo form.  
[...] **saves everything** -- from restaurant receipts (he takes pictures of them) to correspondence, bills and medical records. He makes PDF files out of every Web page he views.  
[...] more than 350 gigabytes worth, not including the streaming audio and video -- is a **replica of Bell's biological memory**.  
It's actually better, he says, because, if you back up your data in enough places, this digitized "e-memory" never forgets.

# Gordon Bell, Microsoft, 2009

"It's like having a **multimedia transcript** of your life.

By about 2020 [...] **our entire life histories will be online and searchable.**

Location-aware smartphones and inexpensive digital memory storage in the "cloud" of the Internet make the transition possible and inevitable.

**No one will have to fret about storing the details of their lives in their heads anymore. We'll have computers for that.**

And this revolution will "*change what it means to be human*"



# Where does the data volume come from? (2)

- Machine-produced data
  - Log data from all kind of servers
  - Real world devices: payments, telecom, energy, weather, transportation, shipment...

Sensors, including on (US) highways and trains

Gazpar (GDF)



Linky (EDF)



# Where does the data volume come from? (2)

- E.g. french railway system: surveillance trains for the normal and high-speed lines
- TGV specially equipped for measuring while circulating at 320 km/h:
  - rail geometry
  - train/rail interaction
  - rail signalization and communication devices
  - electric power availability etc.
  - 150 sensors, 20 cameras
  - complete tour of France's high speed network in 6 days



# Machine-produced data: « Internet of Things » (IoT)

“The **Internet of Things (IoT)** is the interconnection of uniquely identifiable devices over the Internet

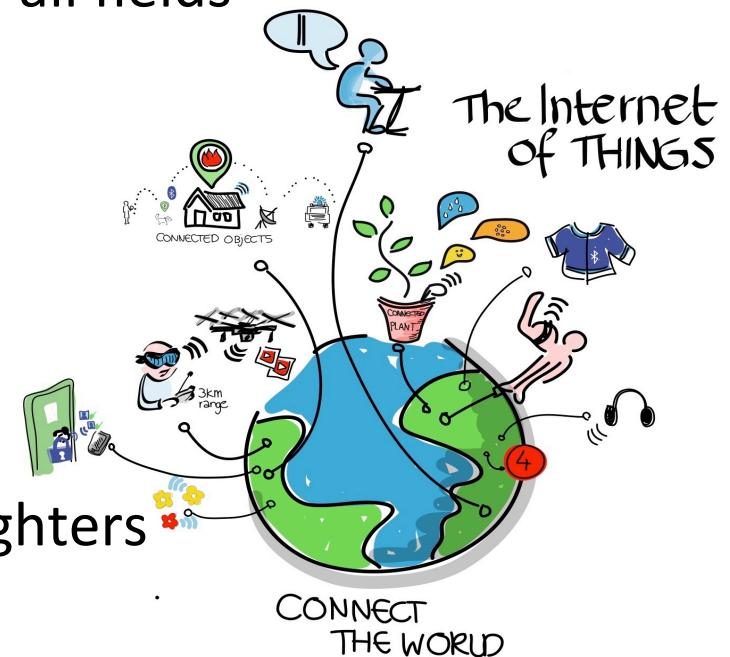
Expected to usher in **automation** in nearly all fields

Sample connected “things”:

- Fridge → supermarket 😊
- heart monitoring implants
- biochip transponders on farm animals
- automobiles with built-in sensors
- field operation devices that assist firefighters

Already there:

- smart thermostat systems
- washer/dryers using wifi for remote monitoring.



# Machine-produced data: « Internet of Things » (IoT)

“The **Internet of Things (IoT)** is the interconnection of uniquely identifiable devices over the Internet

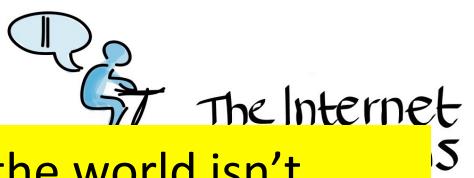
Expected to usher in **automation** in nearly all fields

Sample connected “things”:

- Fridge
  - heart
  - biochip
  - autor
  - field of
- “All the information gathered by all the sensors in the world isn’t worth very much if there isn’t **an infrastructure in place to analyze it in real time.**
- Cloud-based applications are the key to using leveraged data.** The Internet of Things doesn’t function without cloud-based applications to interpret and transmit the data coming from all these sensors.”
- <https://www.wired.com/2014/11/the-internet-of-things-bigger/>

Already there:

- smart thermostat systems
- washer/dryers using wifi for remote monitoring.



The Internet  
of Things

CONNECT  
THE WORLD

# Defining Big Data: the V's

- Volume
  - Scale
- Velocity
  - Speed of producing and consuming the data
- Variety
  - Very different sources and data types
- Veracity
  - Is the data correct / certain / true?

# Big Data velocity

- How much data is produced e.g. per second
- Data enters a **pipeline** consisting of storage and/or processing
  - **Store-then-process**: for off-line data analysis. Storage by itself is a challenge sometimes, e.g. data links to/from clouds are rather slow
  - **Process-then-store**: for data whose interest is maximized upon arrival (real-time processing)
  - **Process-then-discard**: sensor/monitoring (if nothing happens)

# Sample high-throughput data streams

- French IT company runs a data center of **2000** servers
- **5000** energy efficiency indicators (temperature, electricity consumption etc.) are measured every **20** seconds x **50** Kb per measure result = 170 Terabytes / year
- Currently unable to store all data → sample (measure more rarely)
- May miss important things when they happen



# The importance of current/recent data

- Real-time applications work only / mostly with **the latest data**
  - Embedded control mechanisms based on sensor data, e.g., "*this railway wagon component is breaking*" (now!)
  - Intrusion or malfunctioning detection...
- Keeping **humans engaged**
  - Customer relationship management *while the client is on the phone* with the customer service  
(see: “ordering a pizza in the future” video)
  - Recommending places where your friends are hanging out *now*



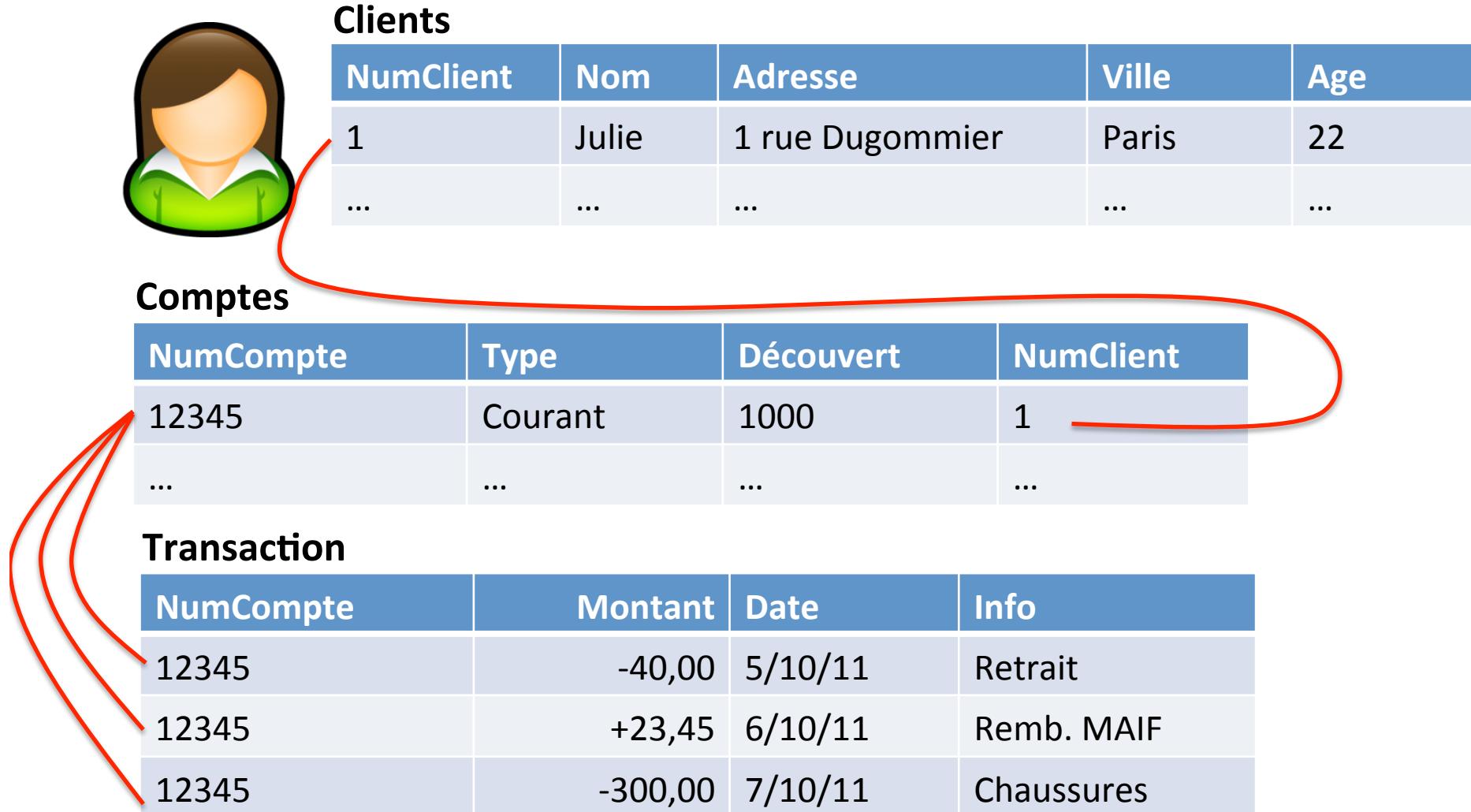
# Defining Big Data: the V's

- Volume
  - Scale
- Velocity
  - Speed of producing and consuming the data
- Variety
  - Different sources, data formats, data types
- Veracity
  - Is the data correct / certain / true?

# Big data heterogeneity (variety)

- Each new data type has added up on the old ones
  - Enterprise data typically has **high per-byte value** (\$/byte)
    - Hard to explain that "we will not need this database in the future"
  - In many areas, **legal obligation** to keep old data (e.g. railway sensors, telecom, commercial...)
- Data model & data management system soup
  - hierarchical, relational, object-oriented, XML, RDF, JSON, key-value pairs...

# Sample relational database

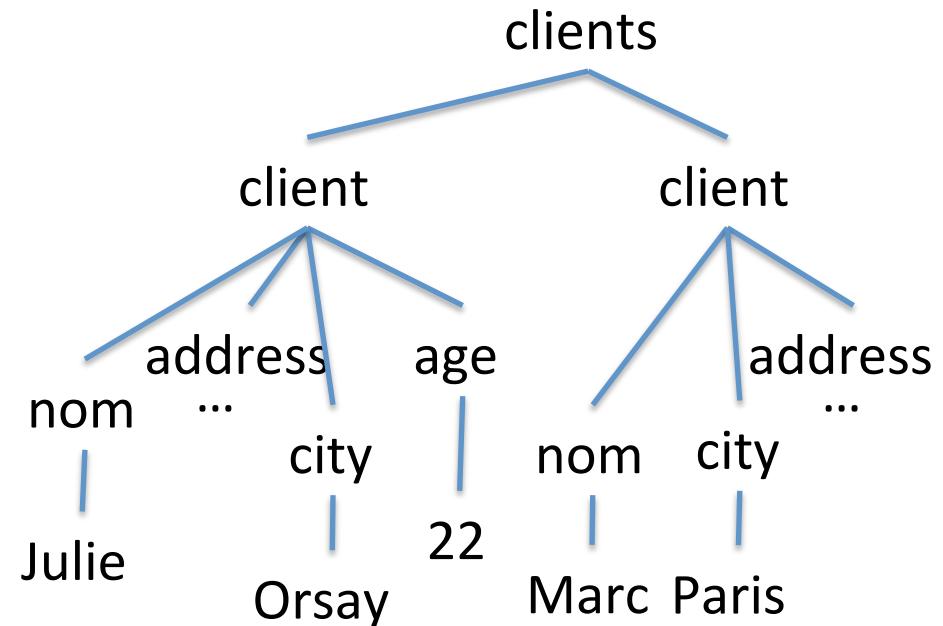


# XML: extensible markup language

W3C, 2008

clients.xml:

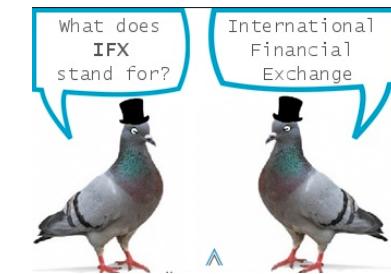
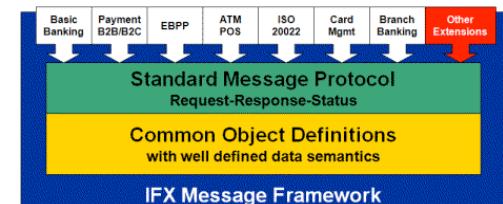
```
<clients>
<client><nom>Julie</nom>
<address>1, rue Dugommier</address>
<city>Paris</city><age>22</age>
</client>
<client><nom>Marc</nom>...
</client>
</clients>
```



Flexible  
Platform-independent  
Separate content from presentation  
Schema possible (not compulsory)

# XML applications

- Main language for the Web: **XHTML, XML Schema, SVG, RSS, ...**
- Web Services: **SOAP, WSDL, BP4WS**
- **MathML** (mathematical markup language)
- **CML** (chemical markup language)
- **SMILE** (synchronized multimedia integration language)
- Financial Exchange (**IFX**)
- The Text Encoding Initiative (**TEI**)



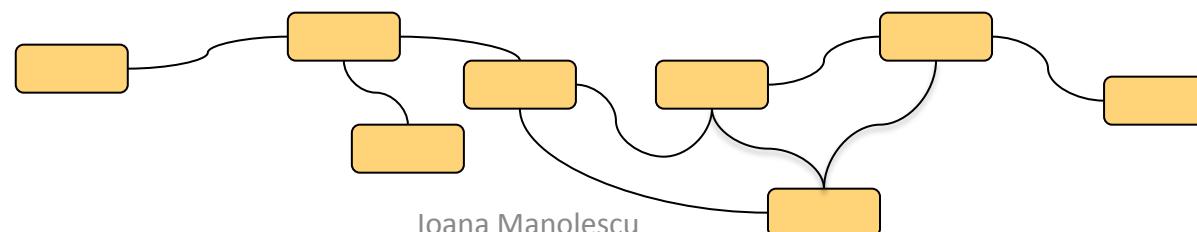
# Critique of XML

- Each information ends up in only one place
  - OK for "classification" applications, structured text
  - Fundamentally restrictive for **data = real world!**

Tim Berners-Lee, WWW proposal, CERN, 1998:

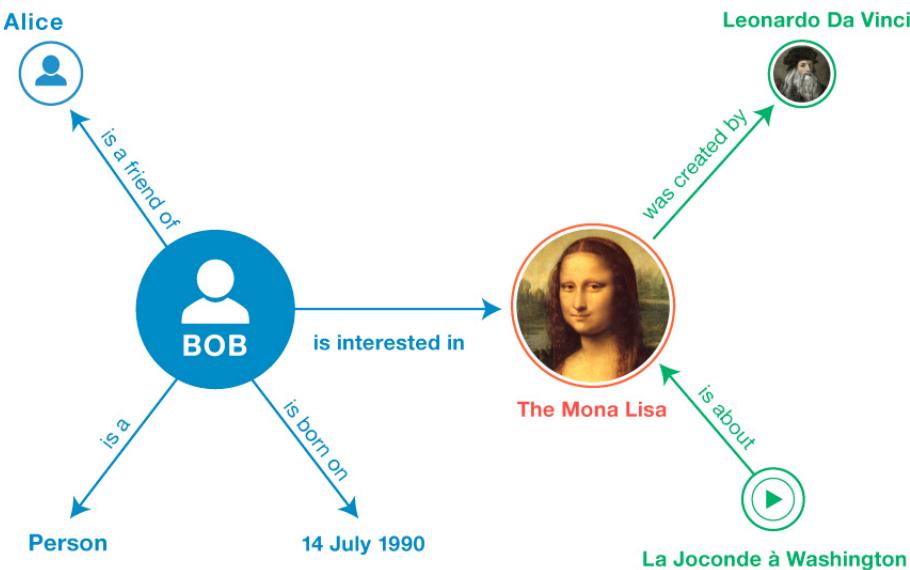
*"Many systems are organised hierarchically. A tree has the practical advantage of giving every node a unique name. However, it does not allow the system to model the real world."*

(On newsgroups): "Typically, a discussion under one newsgroup will develop into a different topic, at which point **it ought to be in a different part of the tree.**"



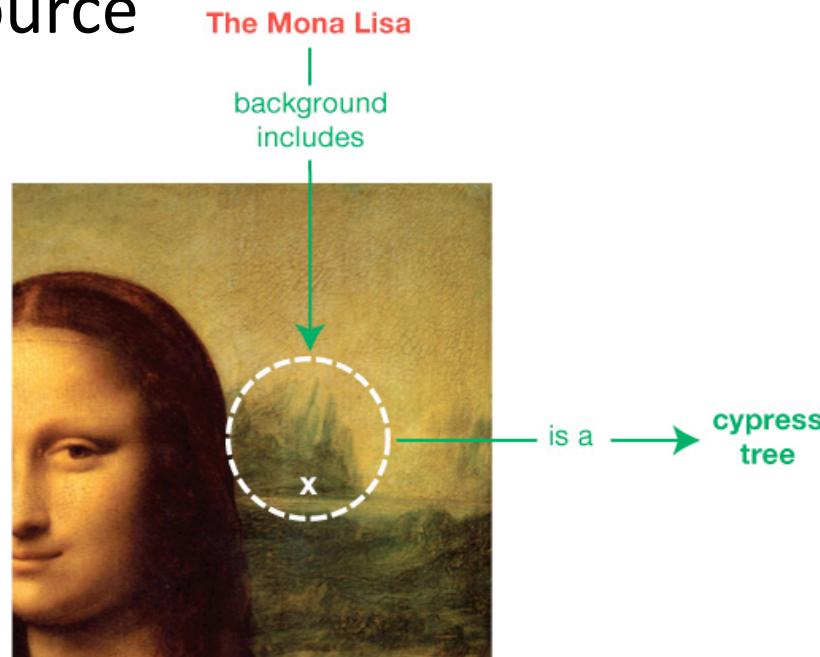
# Graph data format for the Web: RDF (see also pre-requisite RDF course!)

- Resource Description Format, W3C, 2003
- Resources have properties with values.
- URIs (Universal Resource Identifiers) identify resources
- Resources, properties, or values may be specified by an URI.
- Properties and values may be constants



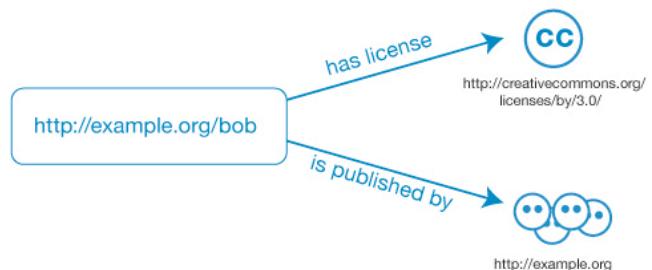
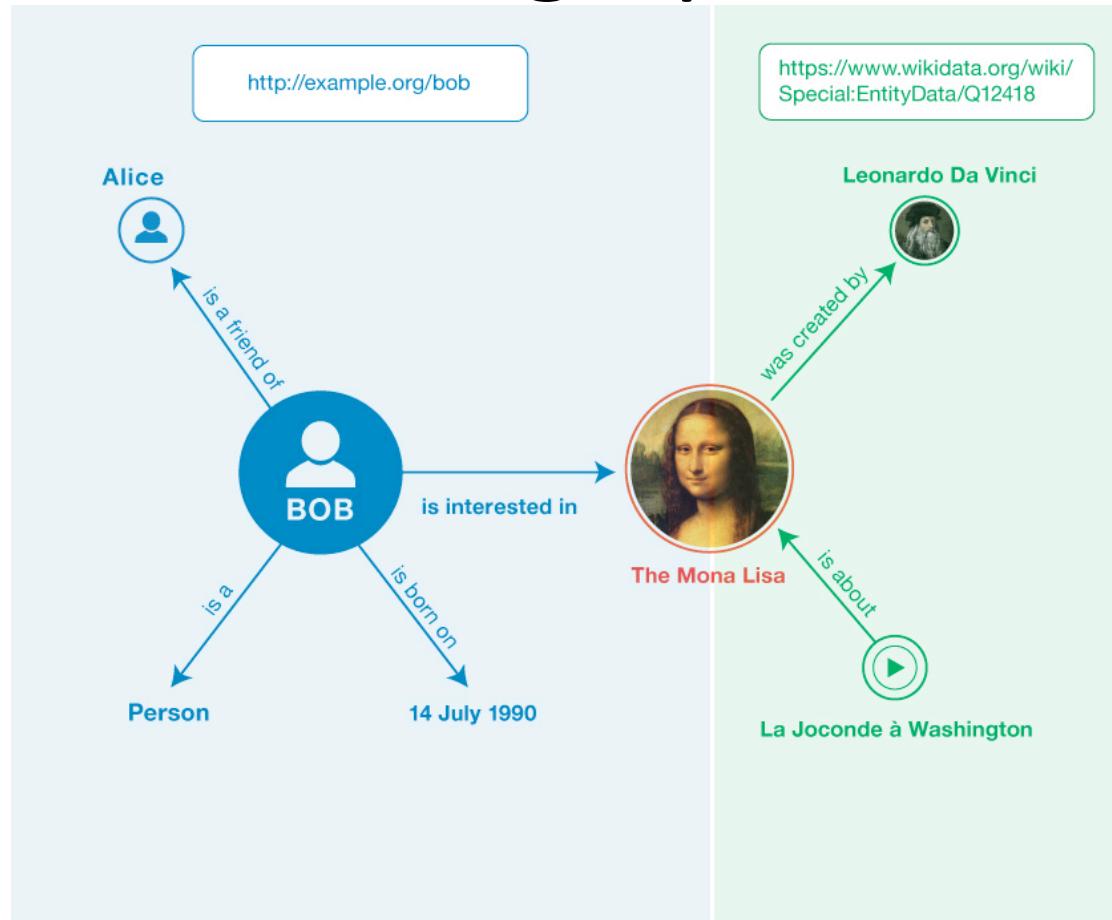
# RDF feature: blank nodes

- Unnamed resource



- « Labeled null »

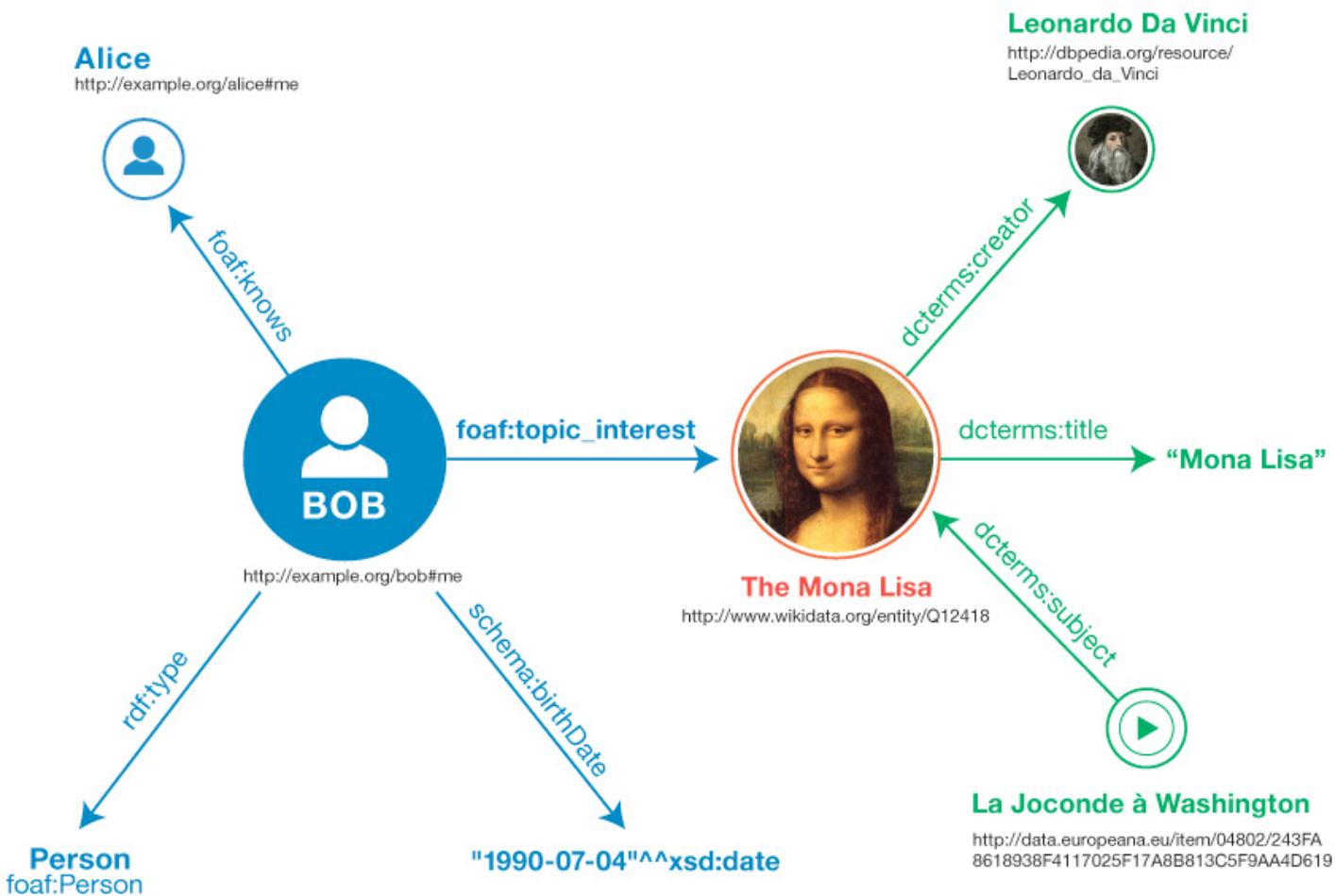
# RDF graphs



# RDF Schema constructs

Construct	Syntactic form	Description
<a href="#">Class</a> (a class)	<b>C</b> rdf:type rdfs:Class	<b>C</b> (a resource) is an RDF class
<a href="#">Property</a> (a class)	<b>P</b> rdf:type rdf:Property	<b>P</b> (a resource) is an RDF property
<a href="#">type</a> (a property)	<b>I</b> rdf:type <b>C</b>	<b>I</b> (a resource) is an instance of <b>C</b> (a class)
<a href="#">subClassOf</a> (a property)	<b>C1</b> rdfs:subClassOf <b>C2</b>	<b>C1</b> (a class) is a subclass of <b>C2</b> (a class)
<a href="#">subPropertyOf</a> (a property)	<b>P1</b> rdfs:subPropertyOf <b>P2</b>	<b>P1</b> (a property) is a sub-property of <b>P2</b> (a property)
<a href="#">domain</a> (a property)	<b>P</b> rdfs:domain <b>C</b>	domain of <b>P</b> (a property) is <b>C</b> (a class)
<a href="#">range</a> (a property)	<b>P</b> rdfs:range <b>C</b>	range of <b>P</b> (a property) is <b>C</b> (a class)

# Typed RDF graph

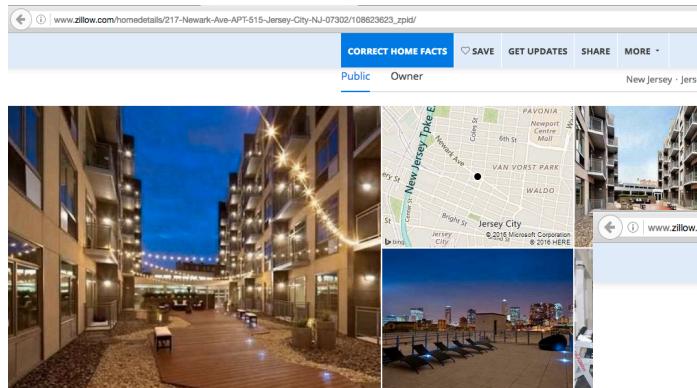


# RDF reasoning

- RDF allow expressing **data and knowledge**
- Example:
  - if X *teaches a class*
  - then X *is a person*, *is an instructor*, *belongs to the school giving the class*, and *works for the university which includes the school*
- Reasoning exploits knowledge to infer *implicit data*

# Varied Big Data has huge value potential

- Real estate ad from Zillow (US):



217 Newark Ave APT 5  
Jersey City, NJ 07302  
-- beds -- baths - 871 sqft [Edit](#)  
Edit home facts for a more accurate estimate.

Thinking About Selling?  
Find a local agent who can give you a professional estimate of your home value.  
[Find an Agent](#)

2 Bedroom 2 Bath w/ Garage Parking at The Saffron. modern design and details that include stainless steel appliances, Caesarstone counters, Bamboo floors an soaring ceiling heights. Master bedroom with walk in and master bathroom. Great storage space, W/D and dishwasher. Manhattan views from spacious roof deck lounge and sun deck or relax in common courtyard, fitness center. Low maint. & taxes. 2.5 blocks from the PATH.

### Price / Tax History

Home price	\$ 505,000
Down payment	\$ 101,000 20 %
Loan program	30-year fixed
Interest rate	See current rates 3.304 96
<input type="checkbox"/> Include taxes/ins. <a href="#">?</a>	

**Your payment \$1,770**

P&I \$1,770

[Get pre-qualified](#)

### Home Expenses

#### INTERNET, TV & PHONE

**\$50-\$100 / month**

[Learn More](#)

**Bundle Services and Save**  
Bundle your monthly services and consider ways to "cut the cord" with low-cost online alternatives to cable such as streaming and on-demand video.

#### SECURITY

**\$14.99 / month**

[Try It Today](#)

**Make Your Home a Fortress**

- 24/7 protection for just \$14.99/mo
- No long-term contracts, so you're free to cancel anytime

Powered by [SimpliSafe](#)

### Nearby Schools in Jersey City

SCHOOL RATING	GRADES	DISTANCE
Number 4 Middle (assigned) out of 10	6-8 & ungraded	0.3 mi
Dr. Ronald Mc Nair Academy High (assigned) out of 10	9-12 & ungraded	0.3 mi
Number 5 Elementary out of 10	PK-8 & ungraded	0.4 mi

[More schools in Jersey City](#)

Data by [GreatSchools.org](#)

cher, centre...

Comparez et économisez jusqu'à 55% sur votre nuit d'hôtel.

[trivago.fr](#)

### Nearby Similar Sales

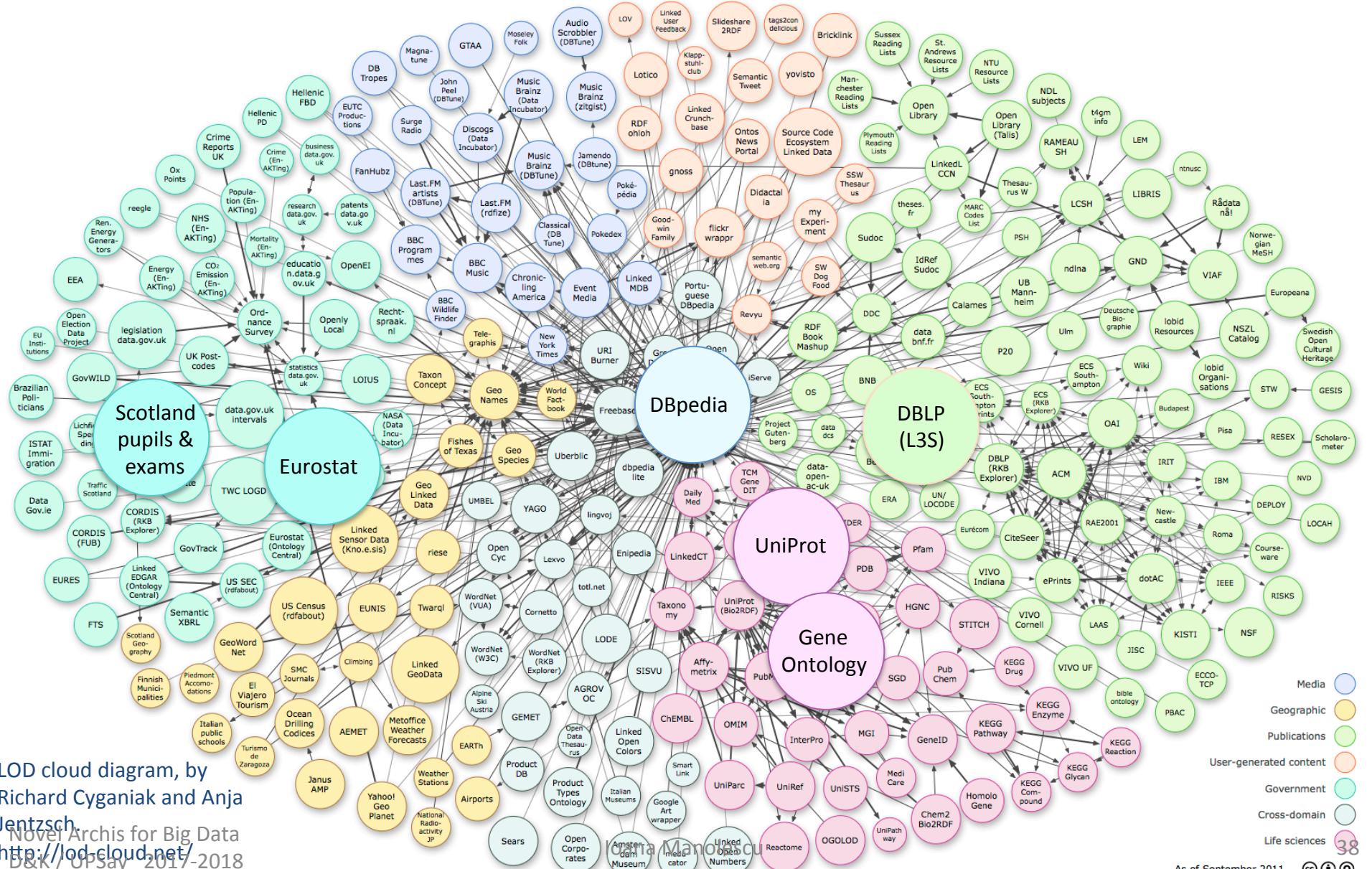
- SOLD: \$551,700**  
Sold on 4/4/2016  
2 beds, 1.0 baths, 800 sqft  
158 Wayne St APT 401A, Jersey City, NJ 07302
- SOLD: \$589,000**  
Sold on 10/15/2015  
2 beds, 1.0 baths, 987 sqft  
280 Monmouth St APT B, Jersey City, NJ 07302
- SOLD: \$594,000**  
Sold on 6/1/2016  
2 beds, 1.0 baths, 1013 sqft  
227 Christopher Columbus Dr APT 217B, Jersey City, NJ 07302
- SOLD: \$599,000**  
Sold on 5/27/2016  
2 beds, 1.0 baths, 978 sqft  
341 Monmouth St APT 311D, Jersey City, NJ 07302
- SOLD: \$600,000**  
Sold on 10/5/2015  
2 beds, 2.0 baths, 1016 sqft  
287 8th St APT 4B, Jersey City, NJ 07302

[See sales similar to 217 Newark Ave APT 515](#)

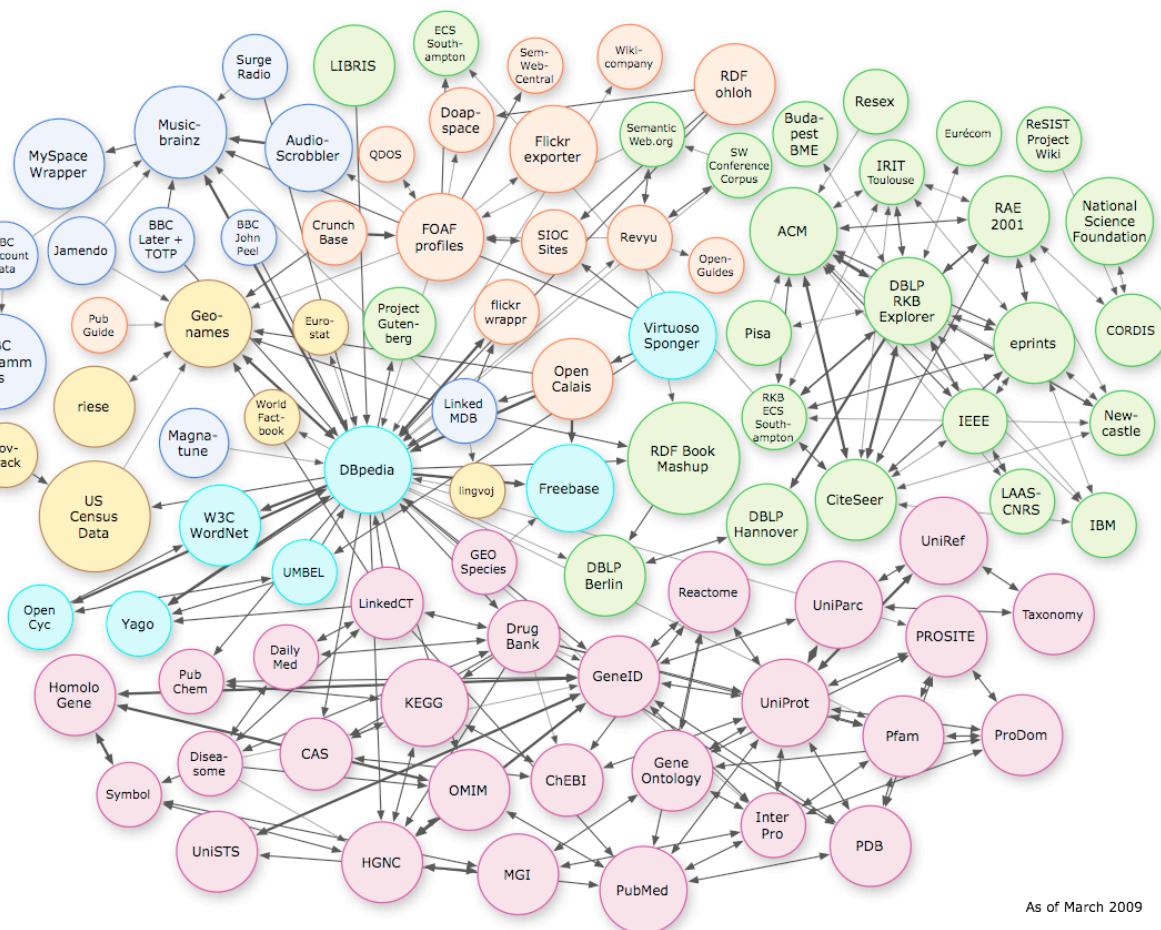
Did you own property near the Rocky Flats Nuclear Weapons Plant on June 7, 1989? Are you an heir or someone who did? Are you the successor of an entity that did? You could get money from a \$375 million Settlement.

# Linked Open Data (LOD) cloud

Linked Open Government Data project ([lod.rw.rpi.edu](http://lod.rw.rpi.edu)):  $10^{10}$  triples.



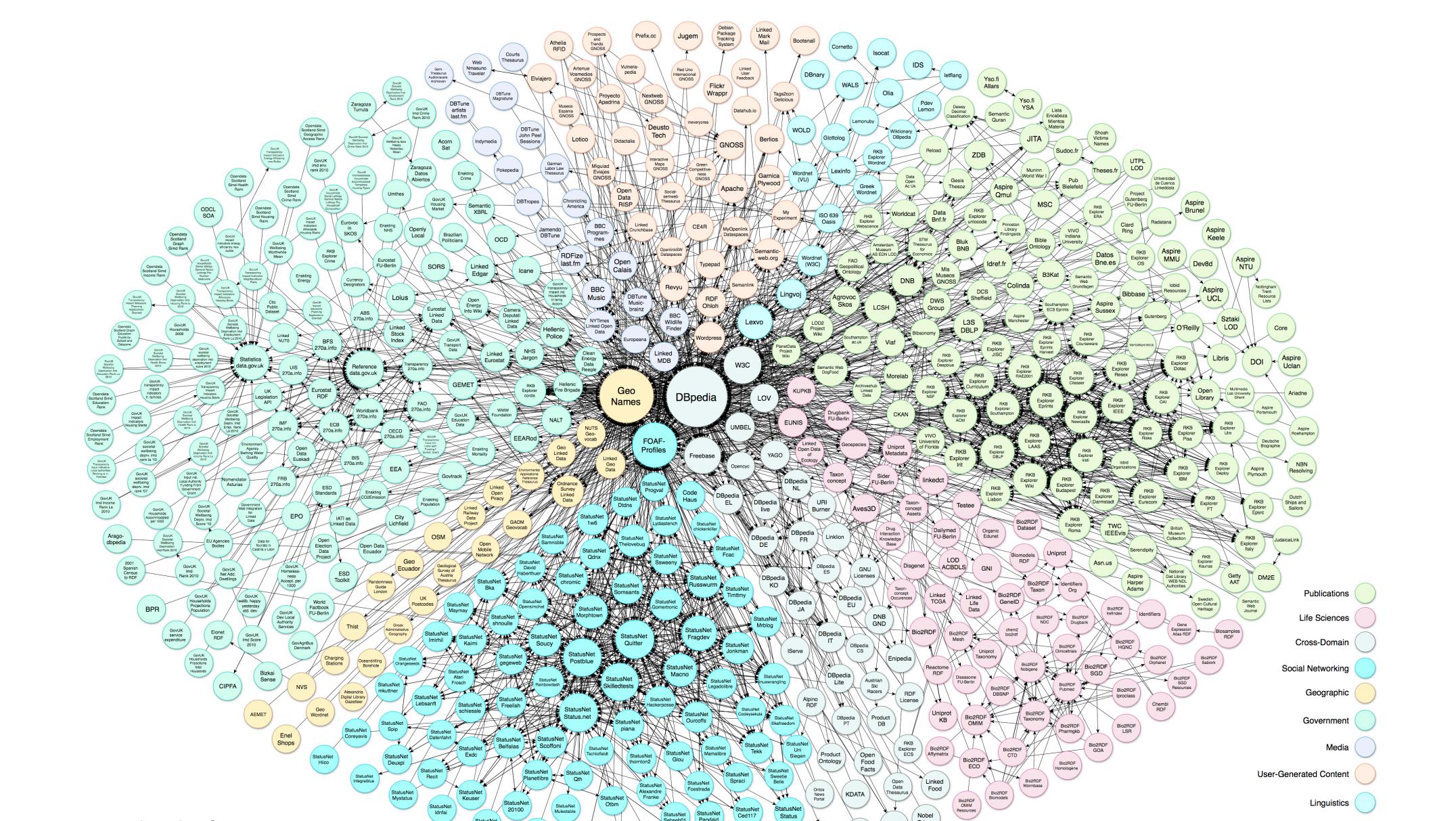
# LOD cloud 2009



Novel Archis for Big Data  
D&K / UPSay 2017-2018

Ioana Manolescu

# LOD cloud 2014

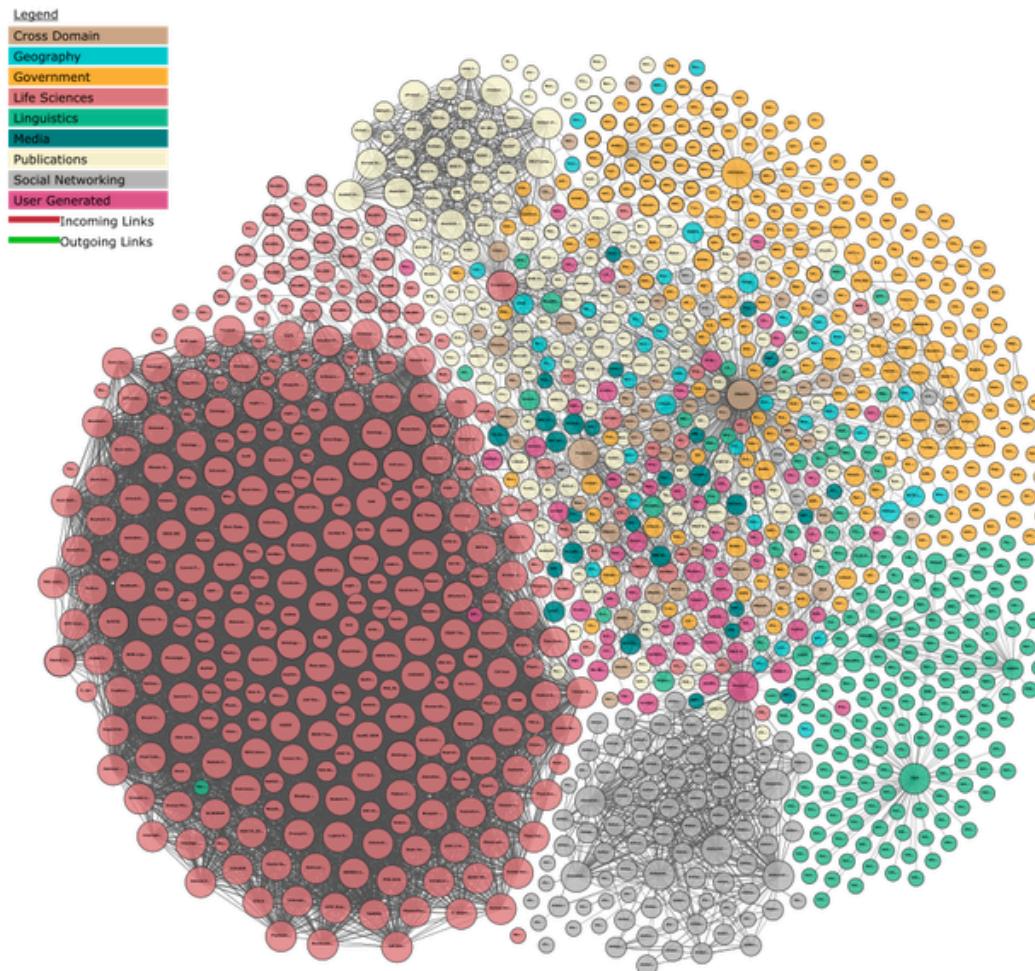


Novel Archis for Big Data  
D&K / UPSay 2017-2018

Linked Datasets as of August 2014



# LOD cloud 2017



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak.  
<http://lod-cloud.net/>

# Open vs. linked data

## 1. Linked Data:

"recommended **best practice** for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using **URIs** and **RDF**"

- (Tim Berners-Lee) vision for the Web

## 2. Open Data:

"**idea** that certain data should be **freely available** to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control"

- In principle, orthogonal to the Linked aspect
- In practice, Linked is a technical mean toward Open

# Open Data: data.gov (US)

The screenshot shows the Data.gov website. At the top, there's a navigation bar with links for HOME, DATA, TOOLS, COMMUNITY, METRICS, OPEN DATA SITES, GALLERY, and WHAT'S NEW. Below the navigation is a search bar with the placeholder "Search our catalogs..". A large blue banner spans the width of the page, titled "PREVIOUSLY HIGHLIGHTED DATASETS AND TOOLS".

Below is a gallery of datasets and tools that have been highlighted on the Data.gov home page. Click on "View More" to learn more about the data and link to the data itself.

This gallery displays just a tiny fraction of the datasets available to you on Data.gov. As we continue to add datasets, tools and highlights, we encourage you to explore all the valuable resources in our raw data, tools, and geodata catalogs.

\* Displaying 62 datasets and tools.



# Open Data: data.gov (US)

**FEATURED TOOL: US CENSUS BUREAU**  
DataFerrett

The DataFerrett is an online analytically oriented, self-service tool designed to deliver a wide variety of population, health, economic, geographic and housing information about the United States. It searches American Community Survey Public Use Microdata, Current Population Survey(CPS), CPS supplemental surveys, Survey of Income and Program Participation (SIPP), SIPP Topical Module surveys, Survey of Program Dynamics, the American Housing Survey, National Survey of Fishing, Hunting, and Wildlife Associated Recreation, The New York City Housing and Vacancy Survey, Local Employment Dynamics.

[VIEW THIS TOOL ▾](#)



Search our catalogs.. [SEARCH ▾](#)

[ITEMS](#) [GALLERY](#) [WHAT'S NEW UPDATED](#)

## DATASETS AND TOOLS

ghted on the Data.gov home page. Click on "View More" to  
ble to you on Data.gov. As we continue to add datasets, tools  
resources in our raw data, tools, and geodata catalogs.

**FEATURED TOOL:**  
**BUREAU OF TRANSPORTATION STATISTICS**  
Airline On-Time Performance and Causes of Flight Delays

The Airline On-Time Performance and Causes of Flight Delays table contains on-time arrival data for non-stop domestic flights by major air carriers, and provides such additional items as departure and arrival delays, origin and destination airports, flight numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, taxi-out and taxi-in times, air time, and non-stop distance.

[VIEW THIS TOOL ▾](#)



**FEATURED DATASET:**  
US PATENT AND TRADEMARK OFFICE  
Patent Grant Bibliographic Data (1976-)

[VIEW THIS DATASET ▾](#)

**SET:**  
HYDROLOGIC Remote Sensing Center (NRHASC) — Snow Water Equivalents

[VIEW THIS DATASET ▾](#)

**INTERESTED IN:**

[VIEW THIS DATASET ▾](#)

**Airline On-Time Performance and Causes of Flight Delays**

# Open Data: data.gov (US)

**FEATURED TOOL: US CENSUS BUREAU**  
DataFerrett

The DataFerrett is an online analytically oriented, self-service tool designed to deliver a wide variety of population, health, economic, geographic and housing information about the United States.

**FEATURED DATASET:**  
**ENERGY INFORMATION ADMINISTRATION (EIA)**  
Residential Energy Consumption Survey (RECS)

The Residential Energy Consumption Survey (RECS) provides information on the use of energy in residential housing units in the United States. This information includes the physical characteristics of the housing units, the appliances utilized including space heating and cooling equipment, demographic characteristics of the household, the types of fuels used, and other information that relates to energy use.

**COMPLETE DATASET** >  
**CONSUMPTION PORTION OF DATASET** >

**FEATURED DATASET:**  
US PATENT AND TRADEMARK OFFICE  
Patent Grant Bibliographic Data (1976-)

The Patent Grant Bibliographic Data is a collection of bibliographic records for patent grants issued by the U.S. Patent and Trademark Office in 1976, or subsequently with the addition of new data.

**VIEW THIS TOOL** >  
**VIEW THIS DATASET** >

**AIRLINE ON-TIME PERFORMANCE AND CAUSES OF FLIGHT DELAYS**

The Airline On-Time Performance and Causes of Flight Delays table contains on-time arrival data for non-stop domestic flights by major air carriers, and provides such additional items as departure and arrival delays, origin and destination airports, flight numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, taxi-out and taxi-in times, air time, and non-stop distance.

**VIEW THIS TOOL** >  
**VIEW THIS DATASET** >

**FEATURED DATASET:**  
NATIONAL WEATHER SERVICE (NWS)  
Hydrologic Remote Sensing Center (NRHSC) — Snow Water Equivalents

The National Weather Service (NWS) Hydrologic Remote Sensing Center (NRHSC) provides hydrologic remote sensing products for snow water equivalent (SWE) estimation. These products include a wide variety of operational and private vendor derived SWE products, including SWE maps, SWE time series, SWE time series, and SWE forecasting products.

**VIEW THIS DATASET** >

Search our catalogs.. **SEARCH** >

ITEMS GALLERY WHAT'S NEW UPDATER

## DATASETS AND TOOLS

View home page. Click on "View More" to view more datasets, tools, data, and geodata catalogs.

# Open Data: data.gov (US)

**FEATURED TOOL: US CENSUS BUREAU**  
DataFerrett

The DataFerrett is an online analytically oriented, self-service tool designed to deliver a wide variety of population, health, economic, geographic and housing information about the United States. It includes data from the American Community Survey, Population Survey, Survey of Income and Program Dynamics, Wildlife Associate, Employment Dynamics, and more.

[VIEW THIS TOOL >](#)



Search our catalogs..

ITEMS GALLERY WHAT'S NEW UPDATED

## DATASETS AND TOOLS

View home page. Click on "View More" to see more datasets and tools.

As we continue to add datasets, tools and geodata catalogs.

**FEATURED DATASET:**  
**ENERGY INFORMATION ADMINISTRATION (EIA)**  
Residential Energy Consumption Survey (RECS)



The Residential Energy Consumption Survey (RECS) is a survey conducted by the U.S. Energy Information Administration (EIA) to collect data on energy consumption in the residential sector. The survey covers a wide range of topics, including energy use, energy prices, and energy efficiency measures.

**FEATURED TOOL:**  
**RECREATION INFORMATION DATABASE (RIDB)**



The Recreation Information Database (RIDB) is a warehouse of information about Federal recreation sites. This web service has the ability to export the data to state tourism portals, recreation-related businesses, etc. It is also the "back end" supplying data to the Recreation.gov portal for trip planning information regarding more than 3,000 Federal recreation sites.

[VIEW THIS TOOL >](#)

FEATURED DATA: US PATENT AND TRADEMARK OFFICE Patent Grant Filings



Airline On-Time Performance and Causes of Flight Delays

# Open Data: data.gov (US)

**FEATURED TOOL: US CENSUS BUREAU**  
DataFerrett

The DataFerrett is an online analytically oriented, self-service tool designed to deliver a wide variety of population, health, economic, geographic and housing information about the United States.

Community Survey  
Population Survey  
Survey of Income  
Program Dynamics  
Wildlife Associate  
Employment Dyn

[VIEW THIS](#)



Search our catalogs.. [SEARCH ▶](#)

ITEMS GALLERY WHAT'S NEW UPDATED

## DATASETS AND TOOLS

**FEATURED DATASET:**  
**ENERGY INFORMATION ADMINISTRATION (EIA)**  
Residential Energy Consumption Survey (RECS)



The EIA provides energy information to address pressing issues in the nation's energy markets, including the supply and distribution of oil, natural gas, coal, electricity, and biofuels.

[VIEW THIS DATASET](#)



**FEATURED TOOL:**  
**RECREATION INFORMATION DATABASE (RIDB)**



The RIDB is a central clearinghouse for recreation information, including service hubs, portals, reports, "back end" trip planners, Federal recreation areas, and more.

[VIEW THIS TOOL](#)

Airline On-Time Performance and

**FEATURED DATASET:**  
**NATIONAL WEATHER SERVICE (NWS)**  
National Operational Hydrologic Remote Sensing Center (NOHRSC)  
— Snow Water Equivalents

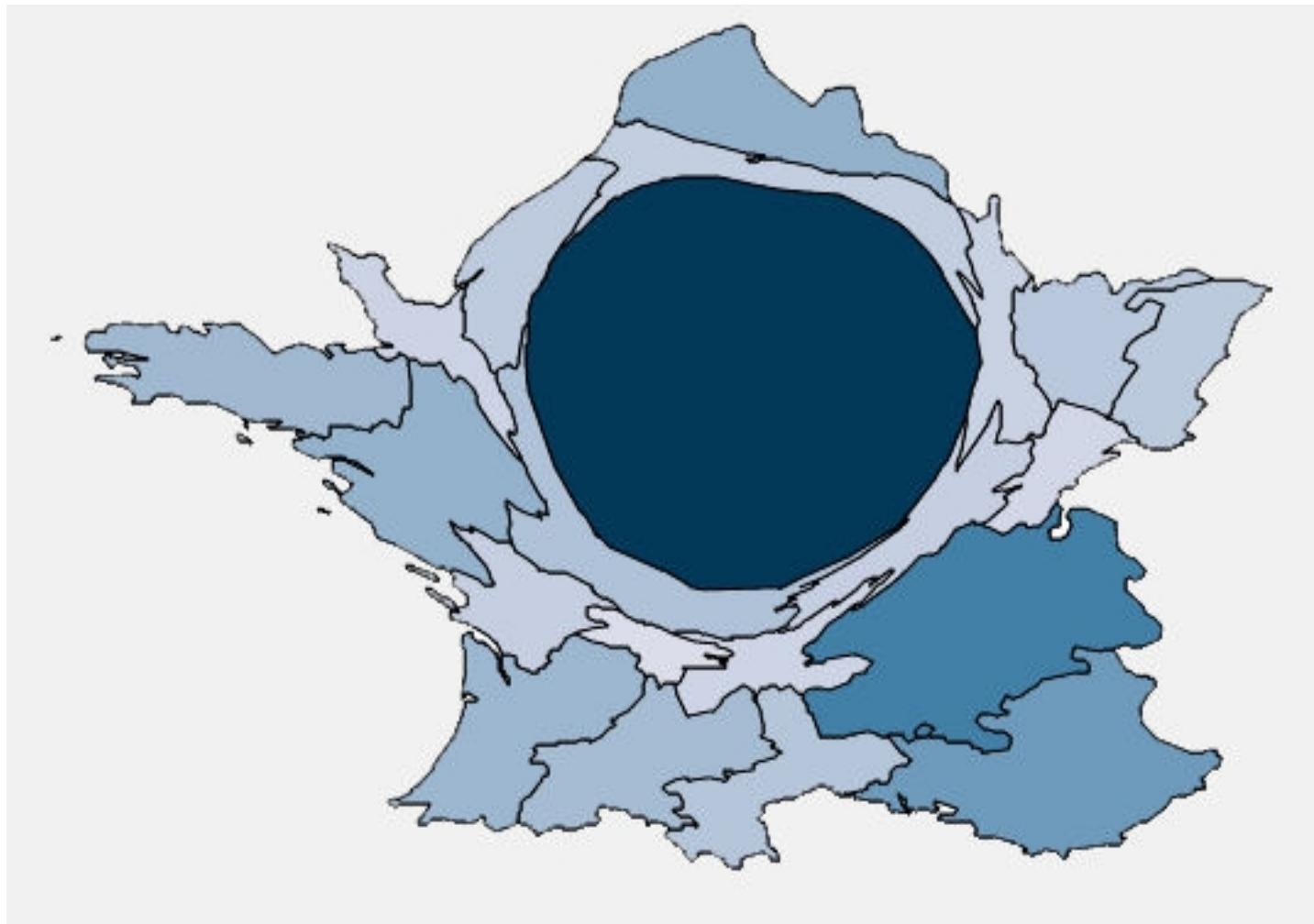


The National Weather Service (NWS) National Operational Hydrologic Remote Sensing Center (NOHRSC) provides comprehensive snow observations, analyses, data sets and map products. Available to all, these products specifically support a wide variety of government and private-sector applications in water resource management, disaster and emergency preparedness, weather and flood forecasting, agriculture, transportation, and commerce.

[VIEW THIS DATASET ▶](#)

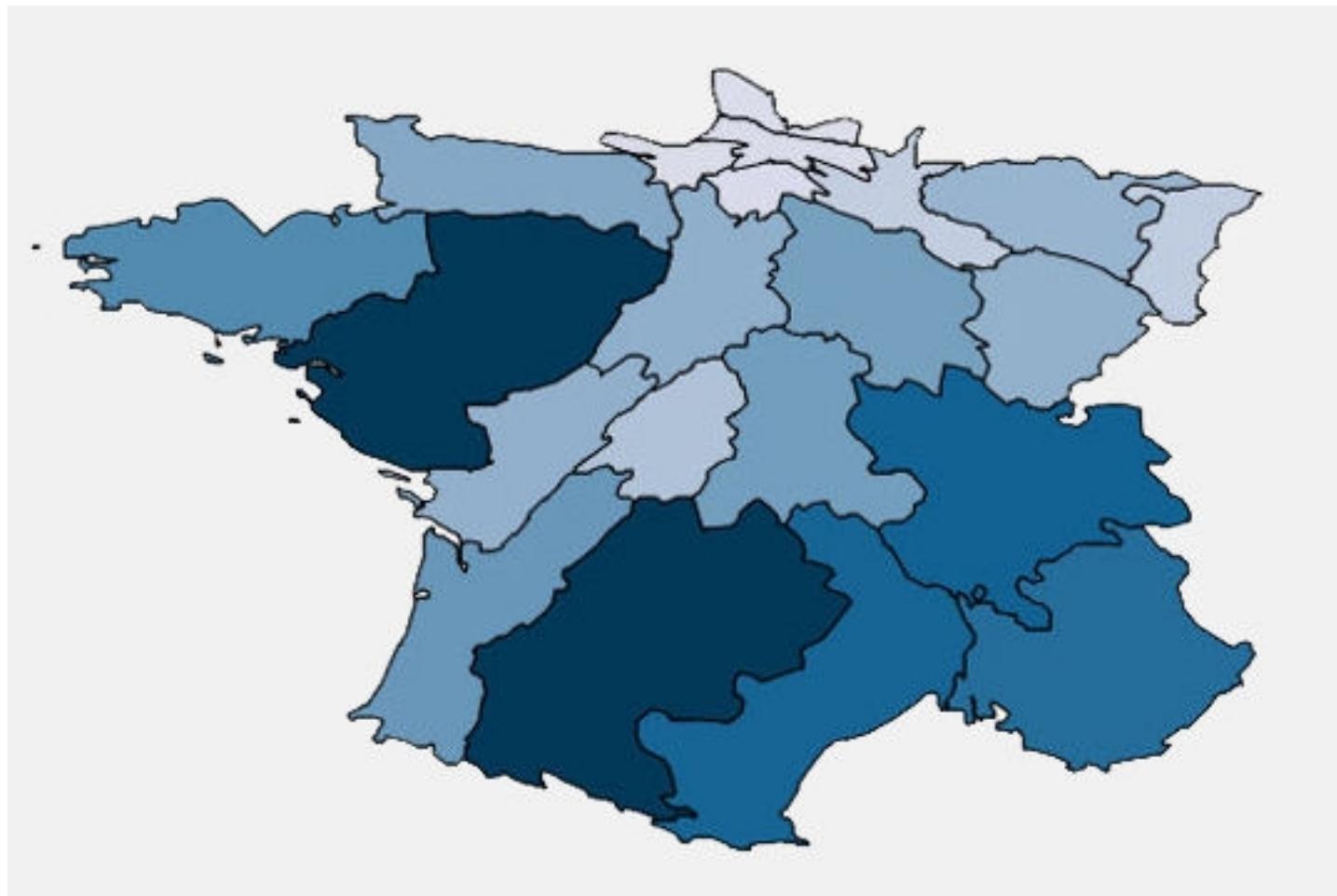
# Open Data from Etalab (FR)

GDP per French region (Le Journal Du Net)



# Open Data from EtaLab (FR)

Organic agriculture per French region



# Big data heterogeneity (variety)

Hierarchical, relational, object-oriented, XML, RDF,  
**JSON**, key-value pairs...

Data model & data management system soup

- hierarchical, relational, object-oriented, XML, RDF,  
JSON, key-value pairs...

Traditionally this has been solved (time and \$ permitting) with **data migration / ETL** (extract-transform-load)

- Heterogeneous data and high throughput may make ETL **impractical for Big Data**

# JavaScript Object Notation

## Human-readable XML

1. Object = set of (attribute, value) pairs
2. Array = list of values.
3. Value = string | number | true | false | null | object | Array

```
{"menu": {  
    "header": "SVG Viewer",  
    "items": [  
        {"id": "Open"},  
        {"id": "OpenNew", "label": "Open New"},  
        {"id": "ZoomIn", "label": "Zoom In"},  
        {"id": "ZoomOut", "label": "Zoom Out"},  
        {"id": "OriginalView", "label": "Original View"},  
        null,  
        {"id": "Quality"},  
        {"id": "Pause"},  
        {"id": "Mute"},  
        {"id": "Help"},  
        {"id": "About", "label": "About CVG Viewer..."}  

```

# Defining Big Data: the V's

- Volume
  - Scale
- Velocity
  - Speed of producing and consuming the data
- Variety
  - Very different sources and data types
- Veracity
  - Is the data correct / certain / true?

# Big Data veracity

- Is this **true**? (What is the **probability**?)
- Contradictory sources (1 vs. 2 clocks)
- Errors in the data
  - Humans introduce many errors
  - Sensors may have failures or erroneous readings
    - Light or heating sensors in a building
    - Wear and tear
- Tackled by **data curation / cleaning / quality** tools for regular (or at least homogeneous) data

# Big Data veracity

- Data reconciliation / entity extraction
- Large-scale **ontologies** such as YAGO, DBpedia, Google Knowledge Base (> Freebase)
  - Reference database of core facts
    - Places (city/country etc.), people (public figures, scientists, artists etc.), events (born, died, emigrated, was created...), time
    - Ontologies automatically extracted from Web and other specific sources → need for **reconciliation**



# Big picture on big data

1. **Volume, velocity, variety, veracity**
2. Probably not all the data has the same **value \$/B**
  - This is why existing databases will stay!
3. Very large, unstructured, uncertain-value data may fit in scalable, relatively slow systems (e.g. MapReduce/Hadoop)
  - Massive parallelism for dummies (to be seen)
  - Mining, batch, iterative processing
4. Large SW companies develop "database-style" layers on Map/Reduce, sometimes in Open Source
5. Many ways to distribute and parallelize the work; many data models

