

# Matteo Palmonari

[matteo.palmonari@disco.unimib.it](mailto:matteo.palmonari@disco.unimib.it)

## ***Making Sense of Big Data & Data Linking***

***The web of data and the open linked data cloud***



**ITIS Lab** – Innovative Technologies for Interaction and Services

*Dipartimento di Informatica, Sistemistica e Comunicazione  
Università degli Studi di Milano-Bicocca*



# Outline

- Big Data
- Linking Web Data
- Complex Questions & Decision Making
- Open Data
- Linked Data

# Big Data

## Volume Velocity Variety

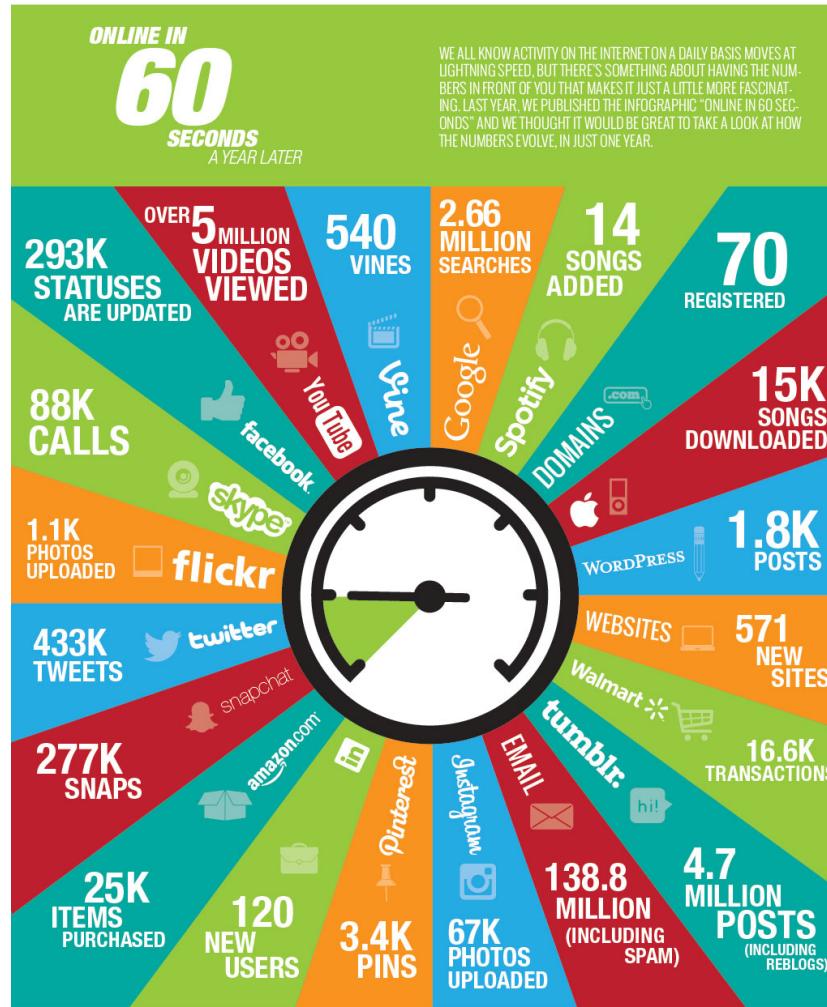
In 2010

## THE “BIG DATA” PHENOMENON



# Every 60 secs in 2014

**Qmee.com**

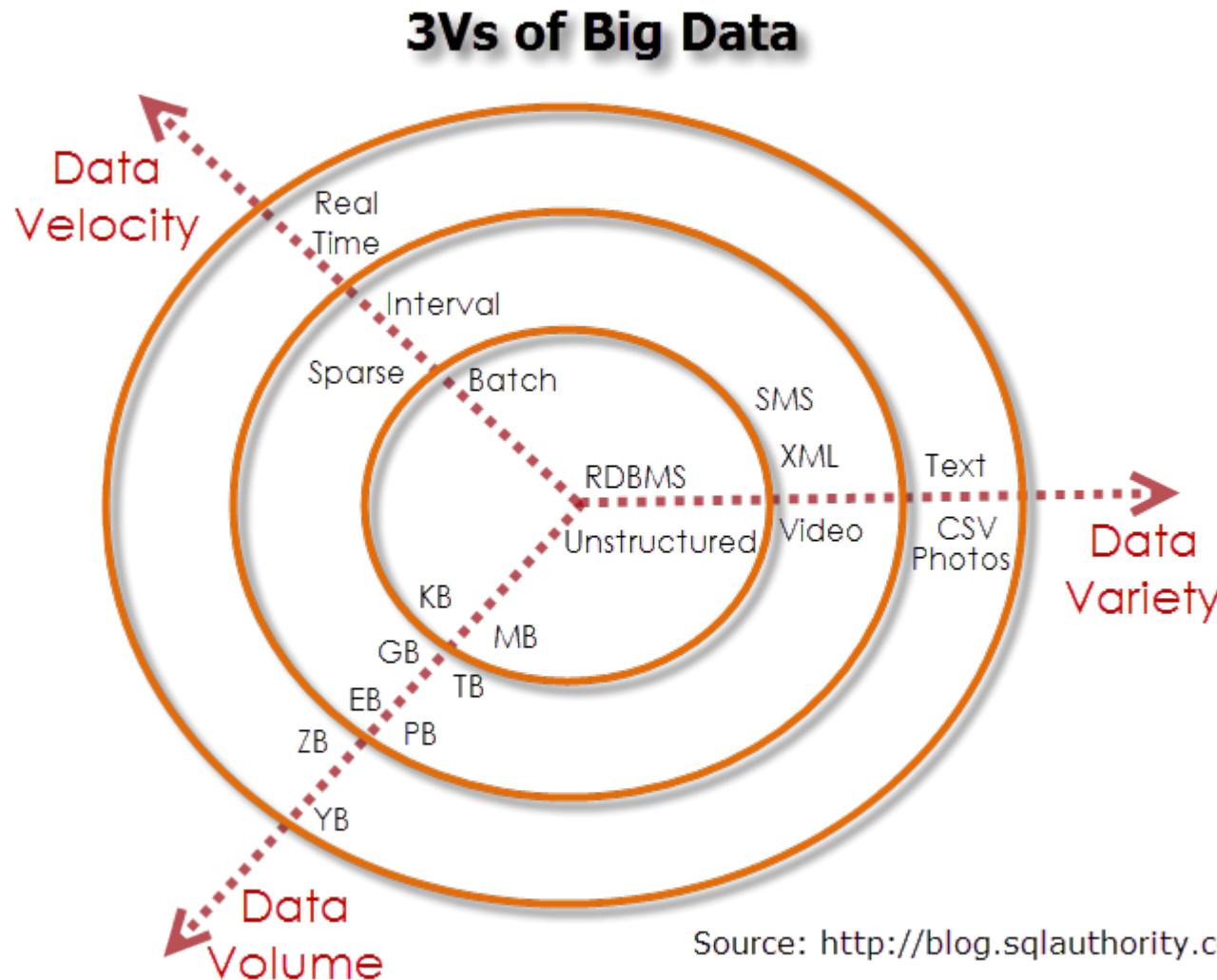


**Qmee**

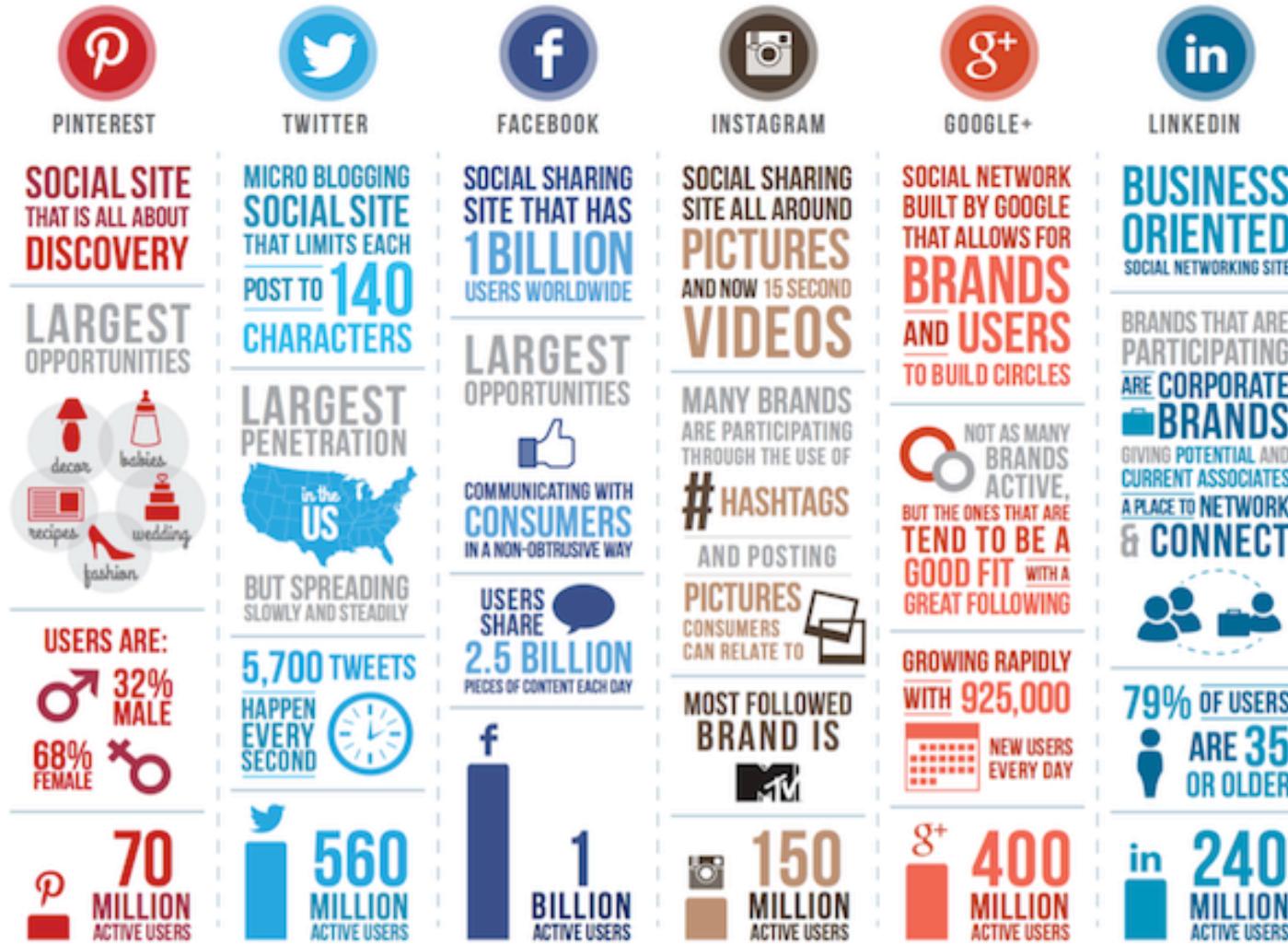
DATA  
[www.internetlivestats.com](http://www.internetlivestats.com)  
[www.thesocialbakers.com](http://www.thesocialbakers.com)  
[www.cnn.com](http://www.cnn.com)  
[www.washington.org](http://www.washington.org)  
[www.weiderl.com](http://www.weiderl.com)  
[www.linkedin.com](http://www.linkedin.com)  
[www.tumblr.com](http://www.tumblr.com)  
[www.amazon.com](http://www.amazon.com)  
[www.acipscamanda.org](http://www.acipscamanda.org)  
[www.mashable.com](http://www.mashable.com)

DESIGN BY **NoLimitAgency**

# Big Data: Volume Velocity Variety



# Variety (Social Media)

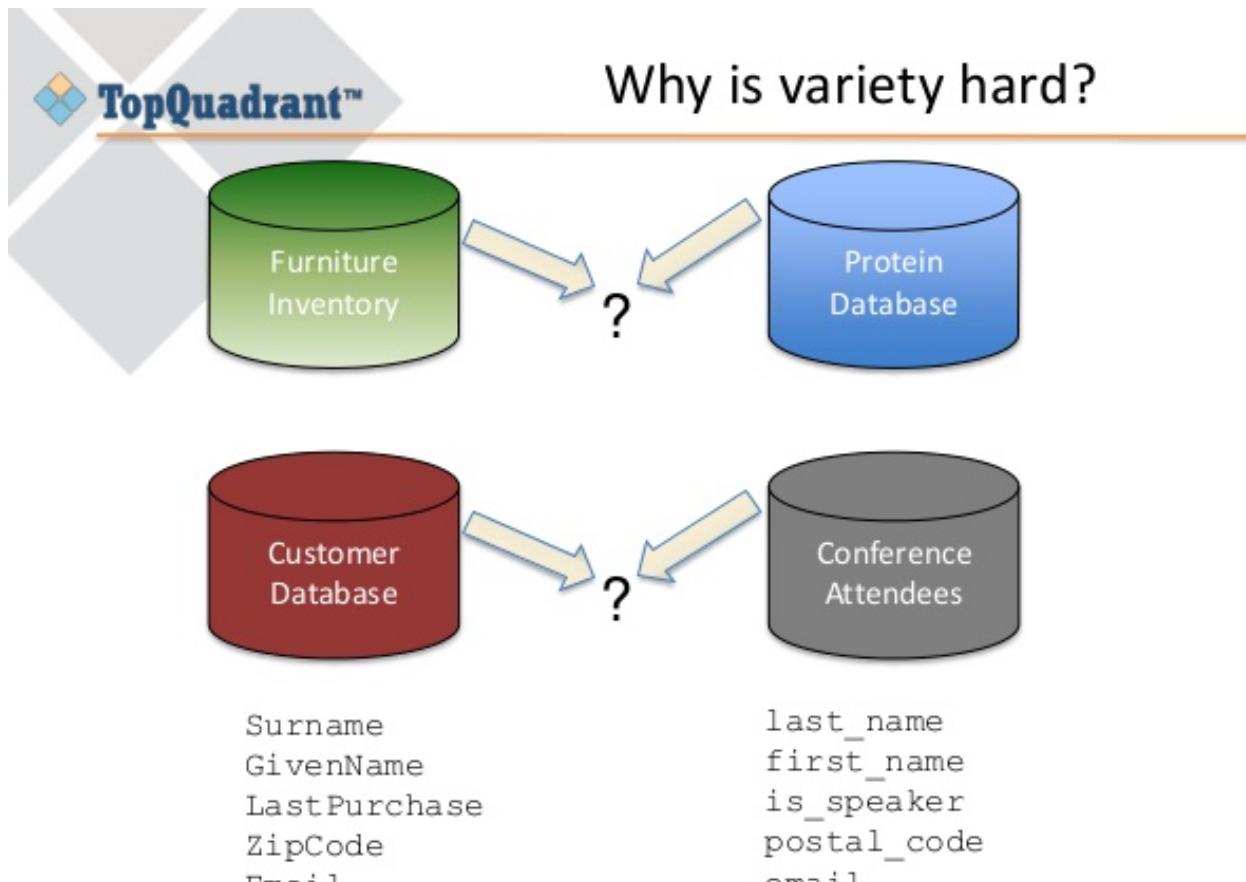


# Variety



According to Kapow Software, a comprehensive list of typical Big Data sources comprises these nine categories: archives, docs, media, business apps, social media, public Web, data storage, machine log data, and sensor data. Yellow means external data (public Web), red is internal (archives and data storage), and orange designates both (docs, media, business apps, social media, sensor data).

# Variety & Semantics (First Dive)



# Web Data

Stadium	City
Eden Gardens	Kolkata
Feroz Shah Kotla Ground	Delhi
M. A. Chidambaram Stadium	Chepauk, Chennai
Wankhede Stadium	Mumbai
Green Park	Kanpur

## 2. Semi-Structured Data

- Hierarchical organization
- May have no schema  
e.g., HTML Tables and spreadsheets with nested headers

### IBM's Watson Computer Made A BBQ Sauce, And It's Delicious

Watson, a cognitive computing system that can learn and process natural human language, has been one of IBM's most exciting projects of the last decade. Over the past few years, Watson has learned a variety of

## 1. Structured Data

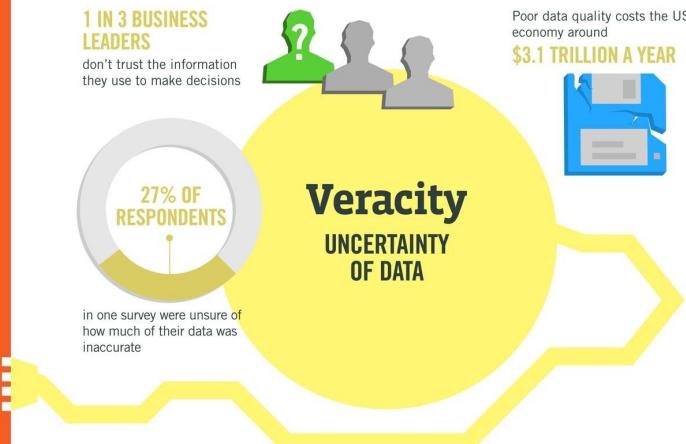
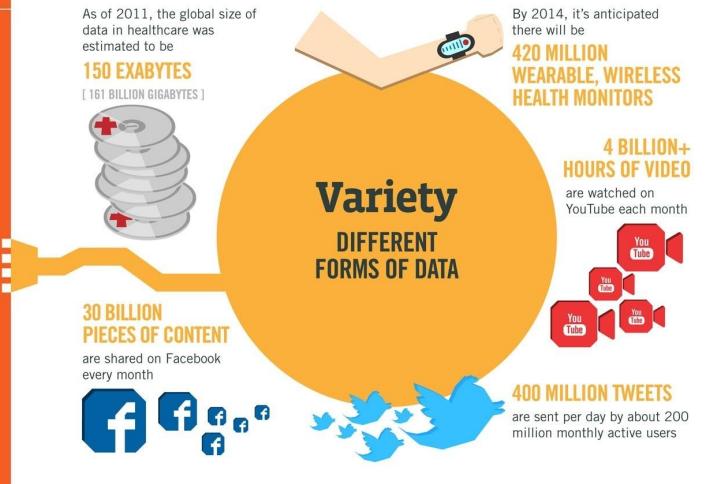
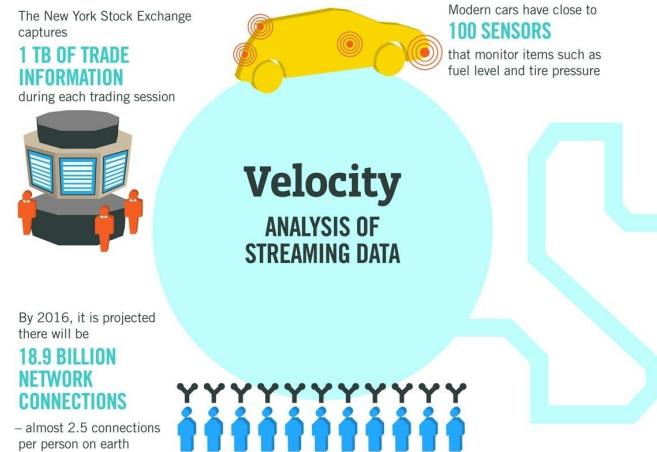
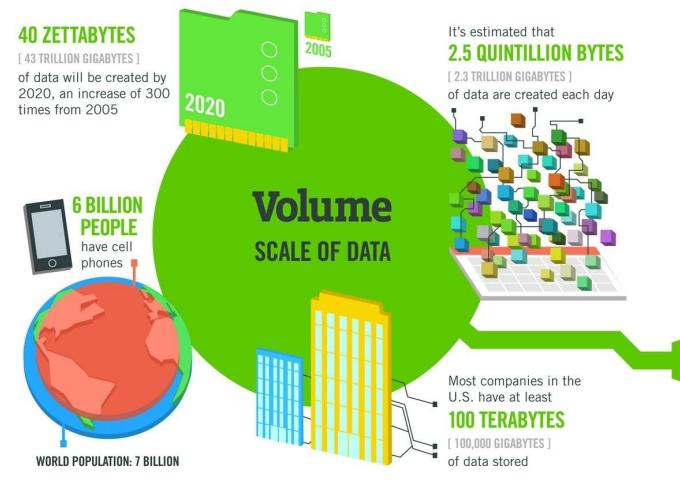
- Organized in attributes
- Well defined schema, follow some order  
e.g., Relational tables, tables in Wikipedia

	Temperate climate		Semi-arid climate		Arid climate	
	%	mm	%	mm	%	mm
Total precipitation	100	500–1,500	100	200–500	100	0–200
Evaporation /Evapotranspiration	~ 33	160–500	~ 50	100–250	~ 70	0–140
Groundwater recharge	~ 33	160–500	~ 20	40–100	~ 1	0–2
Surface runoff	~ 33	160–500	~ 30	60–150	~ 29	0–60

## 3. Unstructured Data

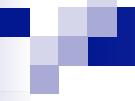
- Natural language text  
e.g., News articles, Wikipedia content

# VVV + Veracity



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

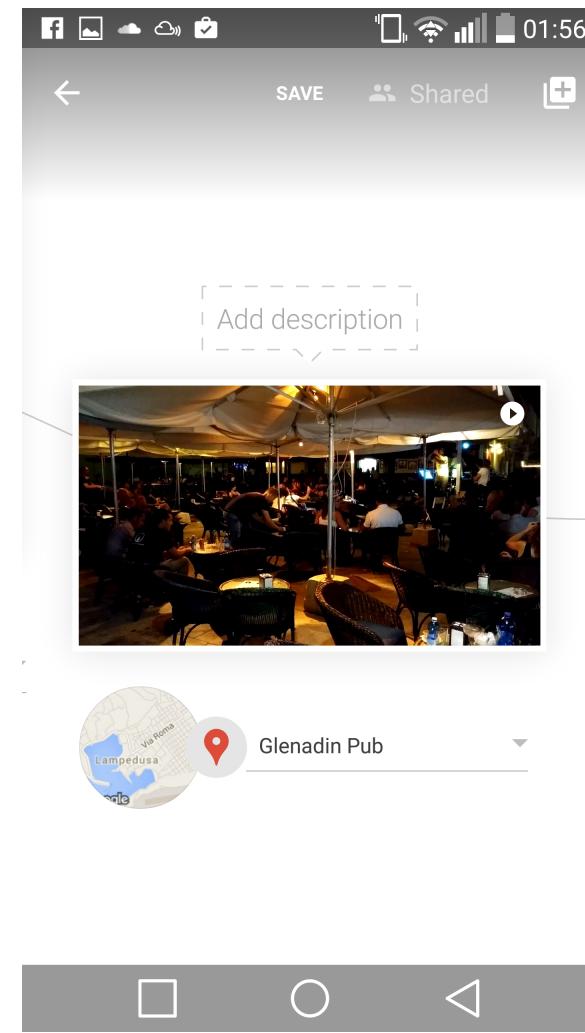
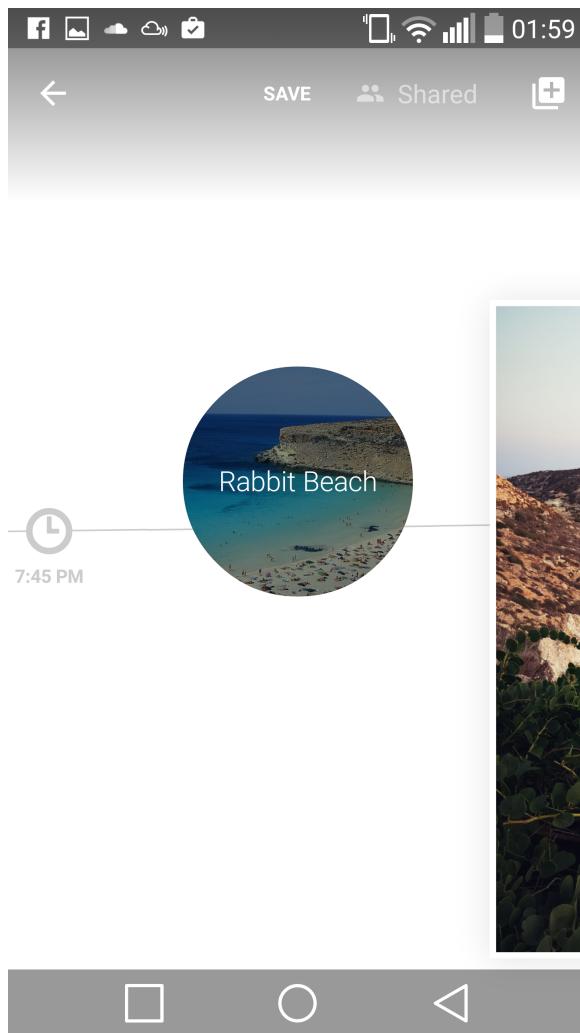




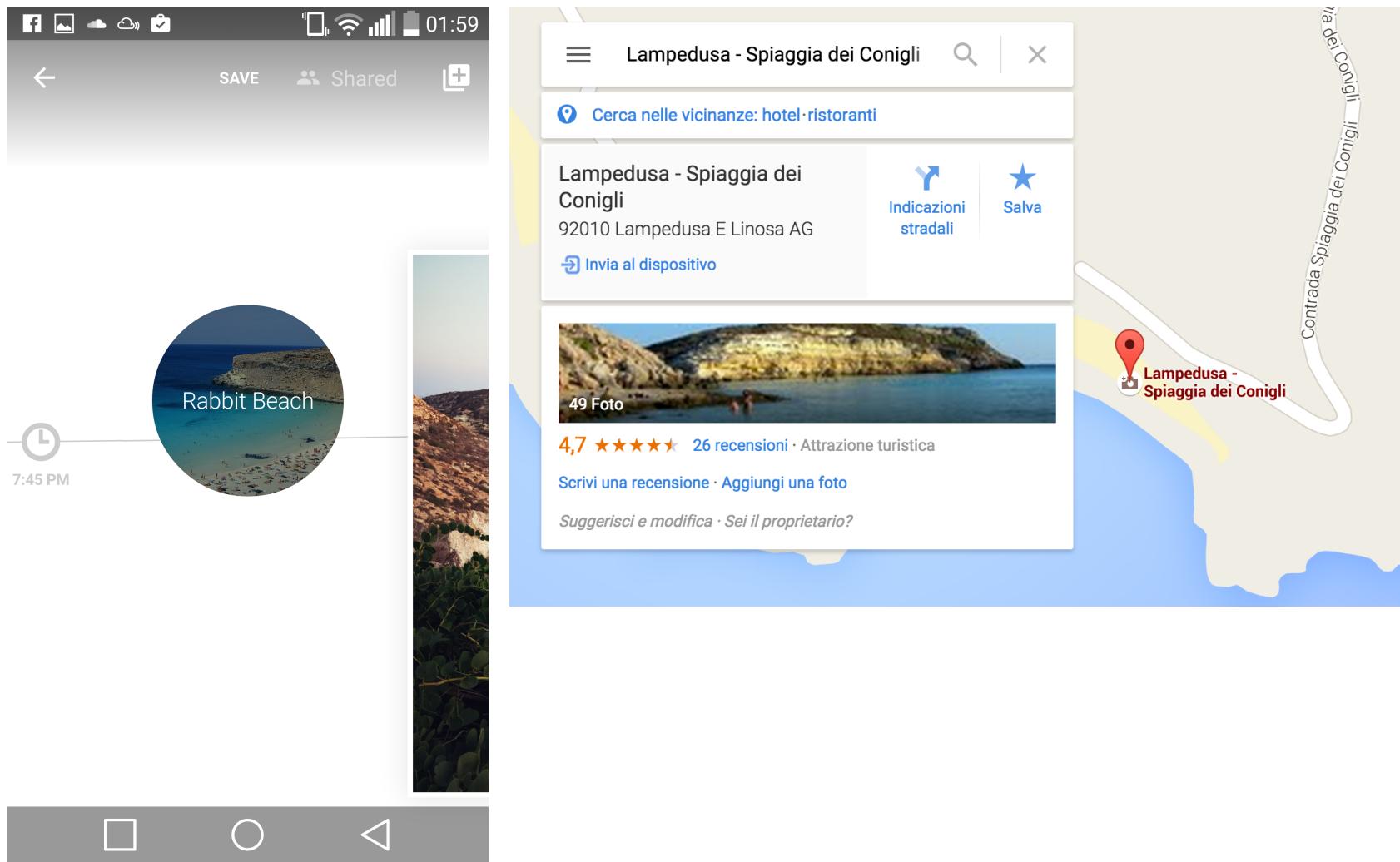
# Linking Web Data

Ovvero perché le connessioni sono importanti per passare da un web di documenti a un web di dati

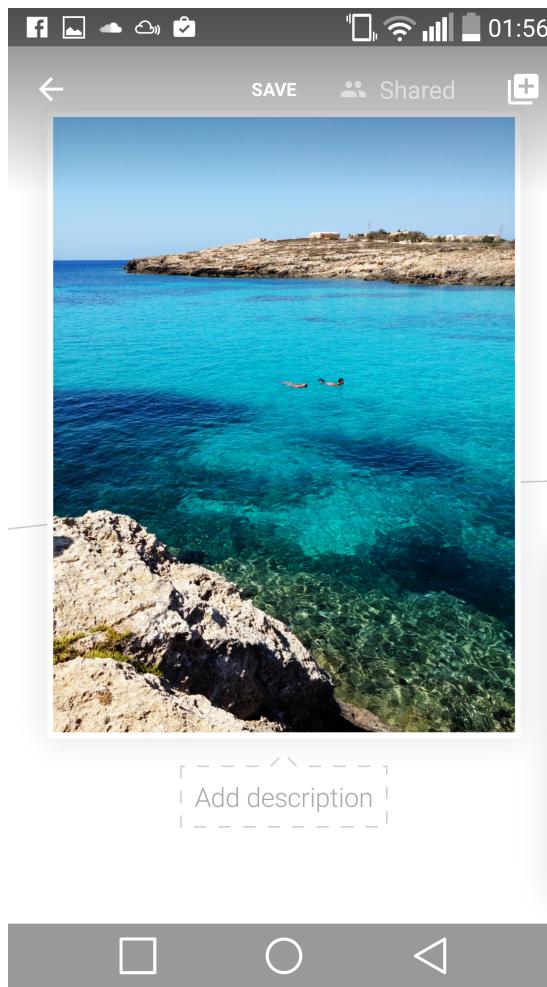
# Connettere esperienze personali a luoghi



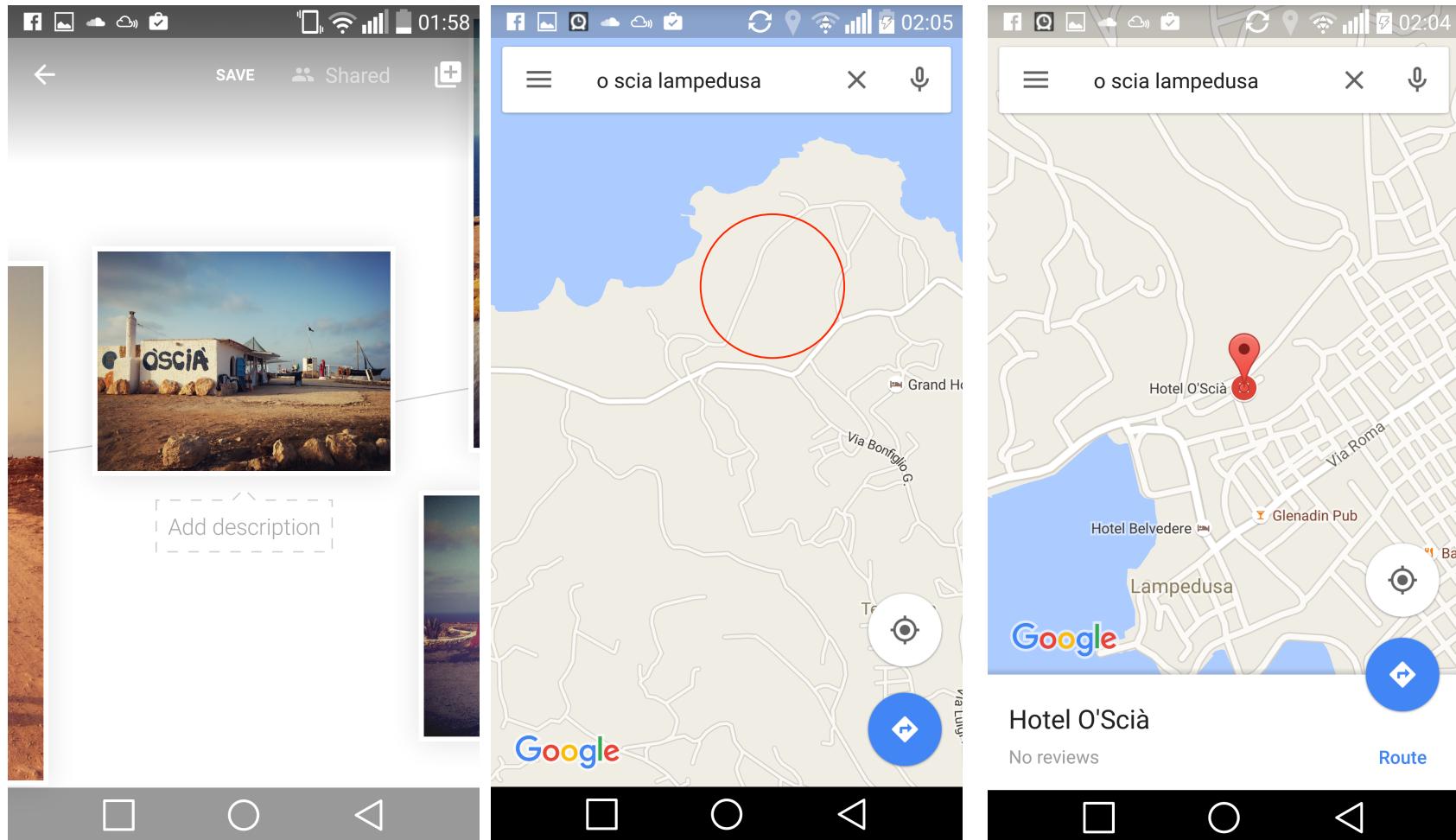
# Connettere esperienze personali a luoghi



# Il valore del dato per le applicazioni



# Il valore del dato per le applicazioni



# Perugia Film: Documenti vs Dati

- Un appassionato di cinema vuole fare un tour in giro per l'Italia visitando luoghi di film a lui cari...
- **Quali film sono stati girati in Umbria?**



Perugia nei film, video e curiosità



In questi giorni gira in rete un video caricato su Youtube contenente gli spezzi di film girati a Perugia, come *Fumo di Londra* (1966), con Alberto Sordi, o *I corpi presentano tracce di violenza carnale* (1973), con Luc Merenda, antesignano del sottogenere horror "Slasher", poi reso celebre dal serial cult Venerdì 13.

# Perugia Film: Documenti vs Dati

- Uno studioso di cinema vuole studiare il rapporto tra cinema e territorio umbro
- **Quali film sono stati girati in Umbria fuori da Perugia?**
- **Quali film sono stati girati a Perugia tra il 1970 e il 1979?**
- **Quali film recitati da Alberto Sordi sono stati girati a Perugia tra il 1960 e il 1979?**
- ...

# Documenti vs dati

Slide by

A. Bordes (Facebook)  
E. Gabrilovich (Google)

The screenshot shows a Google search results page for the query "san francisco". A red circle highlights the search bar. A large yellow arrow points from the search bar down to the right-hand sidebar, which displays detailed information about San Francisco. The sidebar includes a map of the city, its area (231.9 sq miles / 600.6 km²), founding date (June 29, 1776), weather (59°F / 15°C), local time (Sunday 3:55 PM PT), and population (812,826). The main search results list includes links to the SFGOV website, SFO International Airport, Wikipedia, and a travel guide.

san francisco

Web Images Maps Shopping News More Search tools

50 personal results | 1,510,000,000 other results.

Welcome to SFGOV City and County of San Francisco Official site  
[www.ci.sf.ca.us/](http://www.ci.sf.ca.us/)

SFGOV is the official website of the government of the City and County of San Francisco, providing information about departments, meetings, legislation, ...

SFO - San Francisco International Airport - Home Page  
[www.flysfo.com/](http://www.flysfo.com/)

Guides on the airlines, concessions, general services, ground transportation and shopping can be found, including flight information, statistics, future ...

San Francisco - Wikipedia, the free encyclopedia  
[en.wikipedia.org/wiki/San\\_Francisco](http://en.wikipedia.org/wiki/San_Francisco)

San Francisco officially the City and County of San Francisco, is the leading financial and cultural center of Northern California and the San Francisco Bay Area. San Francisco Bay Area - History of San Francisco - 1906 San Francisco earthquake

San Francisco Travel Guide: Things to Do, Hotels, Events ...  
[www.sanfrancisco.travel/](http://www.sanfrancisco.travel/)

The official travel and visitors guide for San Francisco. Only In San Francisco can you find San Francisco hotel reservations, tours, flights, maps, popular ...

10 Things Not to Miss in San - Visitor Information Center - San Francisco Events

**San Francisco**

©2012 Google

San Francisco, officially the City and County of San Francisco, is the leading financial and cultural center of Northern California and the San Francisco Bay Area. Wikipedia

Area: 231.9 sq miles (600.6 km²)  
Founded: June 29, 1776  
Weather: 59°F (15°C), Wind NE at 4 mph (6 km/h), 46% Humidity  
Local time: Sunday 3:55 PM PT  
Population: 812,826 (2011)

# Informazioni fattuali nei risultati di ricerca

Slide by  
**A. Bordes** (Facebook)  
**E. Gabrilovich** (Google)

**san francisco population**

About 79,700,000 results (0.38 seconds)

[Population, San Francisco, CA](#)

1M www.google.com/publicdata  
**812,826** - Jul 2011  
 Source: U.S. Census Bureau

[San Francisco](#)

San Francisco

**san francisco**

50 personal results | 1,510,000,000 other results.

[Welcome to SFGOV City and County of San Francisco Official site](#)  
 www.ci.sf.ca.us/  
 SFGOV is the official website of the government of the City and County of San Francisco, providing information about departments, meetings, legislation, ...

[SFO - San Francisco International Airport - Home Page](#)  
 www.flysfo.com/  
 Guides on the airlines, concessions, general services, ground transportation and shopping can be found, including flight information, statistics, future ...

[San Francisco - Wikipedia, the free encyclopedia](#)  
 en.wikipedia.org/wiki/San\_Francisco  
 San Francisco officially the City and County of San Francisco, is the leading financial and cultural center of Northern California and the San Francisco Bay Area. San Francisco Bay Area - History of San Francisco - 1906 San Francisco earthquake

[San Francisco Travel Guide: Things to Do, Hotels, Events ...](#)  
 www.sanfrancisco.travel/  
 The official travel and visitors guide for San Francisco. Only In San Francisco can you find San Francisco hotel reservations, tours, flights, maps, popular ...  
 10 Things Not to Miss in San - Visitor Information Center - San Francisco Events

**San Francisco**

San Francisco

Area: 231.9 sq miles (600.6 km²)  
 Founded: June 29, 1776  
 Weather: 59°F (15°C), Wind NE at 4 mph (6 km/h), 46% Humidity  
 Local time: Sunday 3:55 PM PT  
 Population: 812,826 (2011)

# Rappresentare informazioni (e connessioni) mediante *knowledge graph*

Slide by

A. Bordes (Facebook)  
E. Gabrilovich (Google)

## Why (knowledge) graphs?

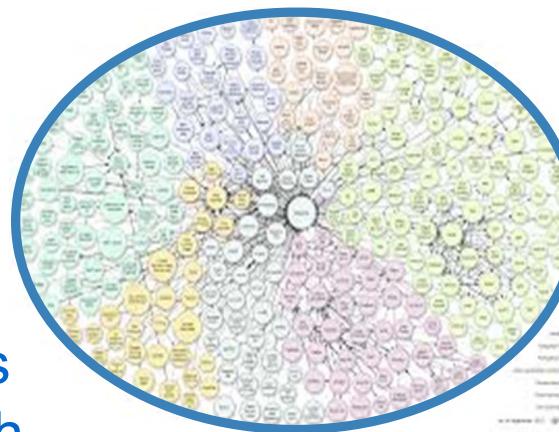
- We're surrounded by **entities**, which are connected by **relations**
- We need to store them somehow, e.g., using a **DB** or a **graph**
- **Graphs** can be processed **efficiently** and offer a convenient abstraction

# Rappresentare informazioni (e connessioni) mediante *knowledge graph*

Knowledge graphs



Facebook's  
Entity Graph



Microsoft's  
Satori



Slide by  
**A. Bordes** (Facebook)  
**E. Gabrilovich** (Google)



OpenIE  
(Reverb, OLLIE)

Google's  
Knowledge Graph

# Documenti vs dati

Google New York

Web Images News Maps Videos More Search tools

About 716,000,000 results (0.95 seconds)

**New York - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/New\\_York](http://en.wikipedia.org/wiki/New_York) ▾ Wikipedia  
 New York is a state in the Northeastern and Mid-Atlantic regions of the United States. New York is the 27th-most extensive, the third-most populous, and the ...  
 New York City - Albany - List of cities in New York - New York metropolitan area

**New York City - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/New\\_York\\_City](http://en.wikipedia.org/wiki/New_York_City) ▾ Wikipedia  
 For other uses, see NYC (disambiguation) and New York, New ...  
 Neighborhoods - History of New York City - Nicknames - Borough

**The Official New York City Guide to NYC Attractions, Dining ...**  
[www.nycgo.com/](http://www.nycgo.com/) ▾ New York City ▾  
 Visit NYCgo for official NYC information on travel, hotels, deals and offers like Restaurant Week, and the best restaurants, shops, clubs and cultural events.  
 Must-See NYC - Broadway Shows & Tickets - Events - Tours and Attractions

**The New York Times - Breaking News, World News ...**  
[www.nytimes.com/](http://www.nytimes.com/) ▾ The New York Times ▾  
 There were no reports of survivors on a Malaysia Airlines flight that crashed on Thursday in eastern Ukraine near the Russian border, the scene of fighting ...  
 Natalie Glance and one other person +1'd this

**New York Magazine -- NYC Guide to Restaurants, Fashion ...**  
[nymag.com/](http://nymag.com/) ▾ New York Magazine ▾  
 Daily coverage of New York's restaurants, nightlife, shopping, fashion, politics, and culture. NYMag.com is the online counterpart to New York Magazine.

**News for new york**  
  
**New York, Responding to Surge of Child Migrants, Forms ...**  
 New York Times - by Kirk Semple - 1 hour ago  
 Opposition to sheltering a wave of young migrants has mounted in many communities across the country, but in New York City, the reaction has ...

More news for new york

**NewYork.com - Your Official Site for Travelling To and Living ...**  
[www.newyork.com/](http://www.newyork.com/) ▾

Augmenting the presentation  
with relevant facts



## New York

US State

New York is a state in the Northeastern and Mid-Atlantic regions of the United States. New York is the 27th-most extensive, the third-most populous, and the seventh-most densely populated of the 50 United States. Wikipedia

Capital: Albany

Secretary of State: Cesar A. Perales

Minimum wage: 8.00 USD per hour (December 31, 2013)

Governor: Andrew Cuomo

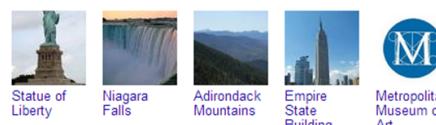
Colleges and Universities: Cornell University, More

## Destinations



View 45+ more

## Points of interest



View 40+ more

Feedback

**Slide by**  
**A. Bordes (Facebook)**  
**E. Gabrilovich (Google)**

# Rich Snippets in Search Results

The screenshot displays three search results with rich snippets:

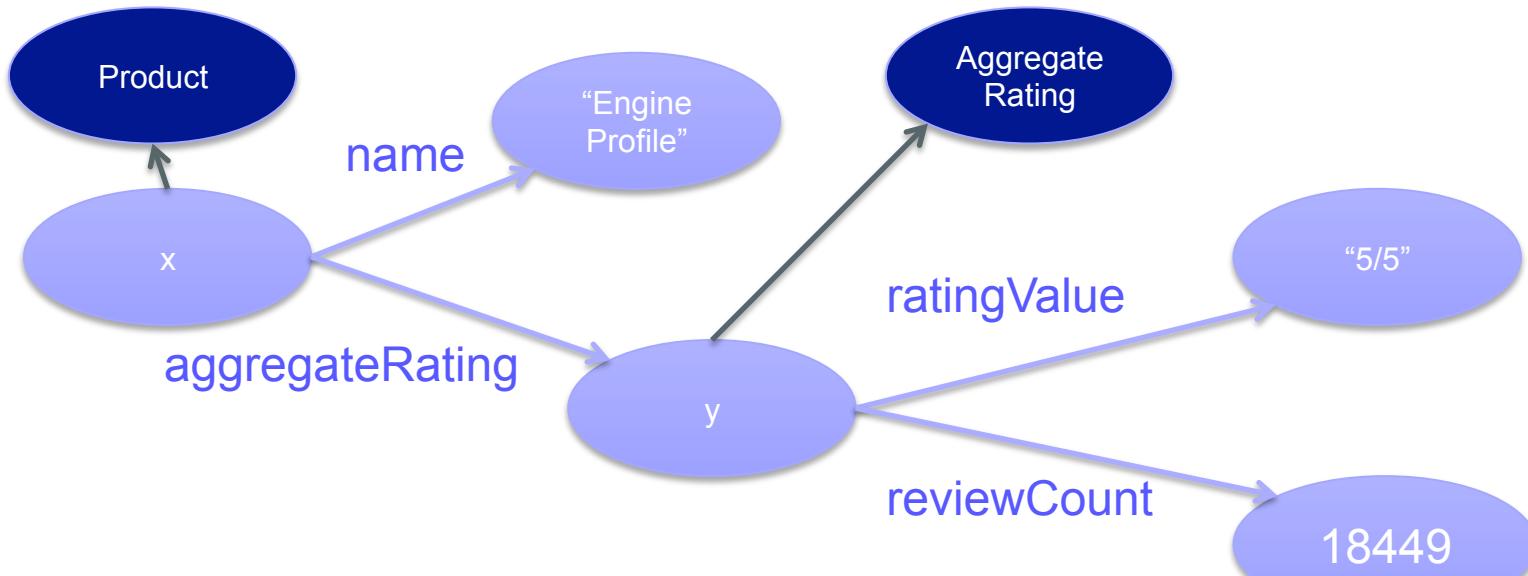
- Little Water Cantina - Eastlake - Seattle, WA**  
www.yelp.com › Restaurants › Mexican  
★★★☆☆ 90 reviews - Price range: \$\$  
90 Reviews of Little Water Cantina. "Three things are on my list when I eat out: great food, atmosphere, and
- Vegetarian Vegan Pizza No Cheese) Recipe - Food.com - 248865**  
www.food.com/recipe/vegetarian-vegan-pizza-no-c...  
★★★★★ 2 reviews - 1 hr 32 mins - 242.9 cal  
Aug 26, 2007 - This is from my dad, who developed some **vegan recipes**
- Leonard Cohen – Free listening, videos, concerts, stats, & pictures at ...**  
www.last.fm/music/Leonard+Cohen  
Watch videos & listen to Leonard Cohen: Suzanne, Hallelujah & more, plus 132 pictures. Leonard Cohen, (born September 21, 1934 in Montréal, Québec, ...  
Track Duration  
Suzanne 3:48  
The Darkness 4:29  
Going Home 3:51  
Hallelujah 6:12

Also useful for  
Search Engine Optimization

```
<span itemscope itemtype="http://schema.org/Product">
<span itemprop="name"><strong>Engine Profiles</strong></span>
<span itemprop="aggregateRating" itemscope
      itemtype="http://schema.org/AggregateRating">
  rated <span itemprop="ratingValue">5</span> / 5<br />
  based on <span itemprop="reviewCount">18449</span></span>
```

**Engine Profiles**  
rated **5 / 5**  
based on **18449**

# Rich Snippets in Search Results



```
<span itemscope itemtype="http://schema.org/Product">  
<span itemprop="name"><strong>Engine Profiles</strong></span>  
<span itemprop="aggregateRating" itemscope  
      itemtype="http://schema.org/AggregateRating">  
    rated <span itemprop="ratingValue">5</span> / 5<br />  
    based on <span itemprop="reviewCount">18449</span></span>
```

**Engine Profiles**  
rated 5 / 5  
based on 18449

# Ricerche Esplorative

**Google** new york sightseeing

Web Maps Images Videos Shopping More Search tools

Sightseeing near New York, NY

	Gray Line New York	City Sightseeing New York	Circle Line Sightseeing Cruises	CitySights NY	NYC & Co	The New York Pass	Rockefeller Center	The Metropolitan Museum of Art	Empire State Building	Statue of Liberty						
2.6 ★★★★☆	64 reviews	1 review	4.0 ★★★★★	57 reviews	2.2 ★★★★☆	60 reviews	3 reviews	3.6 ★★★★★	15 reviews	4.4 ★★★★★	341 reviews	4.7 ★★★★★	725 reviews	4.5 ★★★★★	1,157 reviews	2.08
8th Ave	455 12th Avenue a...	W 42nd St	W 42nd St	7th Ave	W 44th St	5th Ave	Rockefeller Plaza	5th Ave	5th Ave	5th Ave	5th Ave	5th Ave	5th Ave	5th Ave	5th Ave	

**Gray Line New York Sightseeing Tours, Cruises & Attractions**  
[www.newyorksightseeing.com/](http://www.newyorksightseeing.com/) ▾ Gray Line New York ▾  
 New York's famous Empire State Building, a New York City and a National Historic ... to Top of the Rock AND 1-hour Statue of Liberty New York Harbor Cruise! NYC double decker tours - Loops Tour Map - All Loops Tour Plus - Contact Us

**New York attractions: The 50 best sights and attractions in ...**  
[www.timeout.com/newyork/attractions.../new-york-attractions](http://www.timeout.com/newyork/attractions.../new-york-attractions) ▾ Time Out ▾  
 by Amy Pitt - Apr 25, 2013 - Sights like the Empire State Building and the Statue of Liberty are perennial favorites, but we've also highlighted newcomers and lesser-known ...  
 BLDG 92 - Free attractions in New York - Brooklyn Flea - Chrysler Building

**New York City Tours and Attractions - NYC Sightseeing ...**  
[www.nycgo.com/toursandattractions/](http://www.nycgo.com/toursandattractions/) ▾ New York City ▾  
 ITINERARIES. Novel New York. by Jessica Allen. NYC's novelistic life comes alive with self-guided tours to landmarks from books like The Catcher in the Rye, ...

**New York: Sightseeing in NYC - TripAdvisor**  
[www.tripadvisor.com.../New%20York%20\(NY\)%20Before%20You%20Go](http://www.tripadvisor.com.../New%20York%20(NY)%20Before%20You%20Go) ▾ TripAdvisor ▾  
 Inside New York: Sightseeing in NYC - Before you visit New York, visit ... the main sights of the New York City harbor, including the Statue of Liberty, Ellis Island, ...

**NYC Sightseeing Tour | New York City Double Decker Tou...**  
[skylinesightseeing.com/](http://skylinesightseeing.com/) ▾  
 See NYC's landmarks your way with our Hop-on, ... The most historic neighborhood in New York City, downtown Manhattan features the Empire State Building, ...

**City Sightseeing New York, Hop On - Hop Off Bus Tours**  
[www.city-sightseeing.com/tours/united-states-of-.../new-york.htm](http://www.city-sightseeing.com/tours/united-states-of-.../new-york.htm) ▾  
 Choose your own way in which you use your ticket according to your own itinerary. There are three tour routes to choose from, allowing you to explore.

**Top 25 New York City Tours - New York Magazine**  
[nymag.com/visitorsguide/sightseeing/citytours.htm](http://nymag.com/visitorsguide/sightseeing/citytours.htm) ▾ New York Magazine ▾  
 Aug 16, 2013 - Views From the Top The Big Apple Tour See the city streets from a

Slide by  
**A. Bordes** (Facebook)  
**E. Gabrilovich** (Google)

# Connettere Persone e Luoghi

The image displays three Facebook screenshots arranged vertically, illustrating the connection between people and places:

- Top Screenshot:** The Harvard University page. It features a large photo of the university's campus with red brick buildings and green lawns. Below the photo is the Harvard University shield logo. The page has 2,703,624 likes and 47,280 talking about it.
- Middle Screenshot:** The 'People who like Harvard University' page. It lists several users with their profile pictures and basic information:
  - Paul McDonald:** Engineer at Facebook. Likes Harvard University, iRunFar.com and 182 others. Studied Computer Science at Harvard University '03. 3 mutual friends including Clodagh Chloe Takeuchi and Serkan ...
  - Ekaterina Skorobogatova:** Works at Facebook. Likes Harvard University, Loves Company and 3,455 others. Studied Interactive Multimedia at New York University. 2 mutual friends: Amina Belighti and Alexey Spiridonov
  - Gary Johnson:** Corporate Development at Face... Likes Harvard University and 278 others. Studied at Wharton School, University of Pennsylvania '08. 5 mutual friends including Jen Holmstrom and Clodagh Chloe T...
  - Greg Marra (马格雷):** Product Manager at Facebook. Likes Harvard University, Emmy's Spaghetti Shack and 693 ot... Studied Electrical and Computer Engineering at Franklin W. Olin ... 1 mutual friend: Ledell Wu
- Bottom Screenshot:** The 'People who visited Harvard University' page. It lists several users with their profile pictures and basic information:
  - Florence Trouche:** Global Client Partner at Facebook. Visited Harvard University, Marché Poncelet and 317 other pla... Studied at Rouen Business School '90. 35 mutual friends including Michelle Gilbert and Lisa Carucci
  - Andrew Tulloch:** Machine Learning at Facebook. Visited Harvard University, City Beer Store and 90 other places. Studied Machine Learning at University of Cambridge. 21 mutual friends including Jason Weston and Nicolas Vasilache
  - Joseph Barillari (joeb):** Software Engineer at Facebook. Visited Harvard University, Philz Coffee At Facebook and 990 o... Studied Computer Science at Harvard University '07. 10 mutual friends including Tudor Bosman and Jessica Traynor
  - Sheryl Sandberg:** Chief Operating Officer at Facebook. Visited Harvard University and 423 other places. Studied at Harvard Business School. 14 mutual friends including Laurent Solly and Catalina Fries Sa...

Slide by  
A. Bordes (Facebook)  
E. Gabrilovich (Google)

# Question Answering: risposte dirette a domande espresse in linguaggio naturale

Slide by A. Bordes (Facebook) & E. Gabrilovich (Google)

Google search results for "Barack Obama place of birth". The results show a map of Honolulu, HI, and a snippet: "Honolulu, HI" followed by "Barack Obama, Place of birth". Below the snippet is a link to "Barack Obama citizenship conspiracy theories - Wikipedia".

Google

EVI (Amazon)

EVI  
(Amazon)

Siri  
(Apple)

# Online Marketing: Digital Data Trends

Search, Social, & Content Fusion (by Andy Betts, April 10, 2014)

- “One of the biggest challenges marketers face is understanding what to do with mass data at hand. This was a key theme from the recent [ClickZ Live New York](#).”
- “Identifying what content engages with your digital audience is the key to driving efficient, and scalable, integrated marketing campaigns. The panel was in agreement that content is the catalyst that should be placed at the center of all digital campaigns.”
- key trends that focus on breaking down big data into small pieces to help marketers understand how content, search, and social (and subsequent measurement) are morphing together.

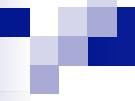


# Online Marketing: Digital Data Trends

Search, Social, & Content Fusion (by Andy Betts, April 10, 2014)

- “top factors that impact conversion and how best to optimize campaigns around data.”





# Complex Questions & Decision Making

# Complex Questions over Social Networks

Structured search within the graph

The screenshot shows a Facebook search results page with the query "People who like Harvard University and Basketball and work at Facebook". The results list five profiles:

- Mike Vernal**: VP Engineering at Facebook. Likes Harvard University, Harvard Crimson and F@ceb00k Su... Studied Computer Science at Harvard University '02. 10 mutual friends including Keith Adams and Philip Bohannon.
- Jared Morgenstern**: Product Manager / Ninja - Games ... Likes Harvard University, F@ceb00k Summer Basketball Leag... Studied Computer Science at Harvard University. 5 mutual friends including Clodagh Chloe Takeuchi and Pierre ...
- Florin Rățiu**: Software Engineer at Facebook. Likes Harvard School of Public Health, Stanford 6th Man and ba... Studied Management Science and Engineering at Stanford Univ... 3 mutual friends including Alexey Spiridonov and Serkan Plantino
- Ning Zhang (张宁)**: Software Engineer at Facebook. Likes Harvard University, Basketball and 314 others. Studied Computer Science at University of Waterloo '06. 5 mutual friends including Tudor Bosman and Ves Stoyanov
- Zhongyuan Xu (徐重远)**: Software Engineer at Facebook. Likes Harvard University, Basketball and 364 others. Studied at Stony Brook University. 1 mutual friend: Ledell Wu

Each profile card includes an "Add Friend" button and a "..." menu. The bottom right corner of the page shows a performance metric: 2147 ms.

# Questions over Web information



Qual è il cantante rock di maggior successo in questo momento in UK?

# Questions over Web information



Quali sono gli ultimi 3 eventi che hanno determinato l'aumento di ascolto di canali televisivi stranieri?

# Questions over Web information



Qual è il cantante rock di maggior successo in questo momento in UK?

# Data Analysis Scenario: School Closings in Chicago

- Utilization Crisis (May 2013)
  - 403,000 students enrolled in Chicago Public Schools (CPS)
  - 511,000 available seats
- Proposed solution
  - 47 schools closed (affecting 12,000 students)
  - Displaced students assigned to a *welcoming school* that is rated higher than the closed school
  - Safe passage routes implemented
- Problems
  - Potential to destabilize fragile neighborhoods (South and West sides) already grappling with high poverty, crime, and unemployment
  - Disproportionately affect the most vulnerable students

# Welcoming School Choice

FamilyConnections

ClosedSchoolStudentsTransferred

School PerformanceMeasures

ClosedSchoolStaffTransferred

Bullying Fighting TransportationCosts ClassesCurriculum

WelcomingEnvironment

CloseToHome

GoodCommunication

AfterSchoolPrograms

Ratings

SafeCommute

IndividualizedAcademicAttention

SpecialEducationSupports

# Current Observations

- 93% of displaced students attended schools with higher performance ratings than the closed schools
- One-quarter of students attended schools that were lower performance than their designated welcoming schools
- Only a small portion of displaced students from these past closings attended substantially higher-performing schools

# Evaluation of Educational Outcomes

- Displaced students have (on average) better academic outcomes
- Displaced students who attend substantially higher-performing schools have better academic outcomes
- Displaced students have (on average) the same academic outcomes
- Displaced students have (on average) worse academic outcomes

# Analytic Method

(2009)

- Has access to individual student data
- Uses two methods
  - Longitudinal survey (temporal), by matching expected student performance trajectory with actual trajectory after displacement
  - Performance comparison, by matching performances of displaced students with that of students from “similar” schools.
- Links to poverty data from Census
- Ignores other Census variables, and economical, societal, and transportation data
- Concludes: Displaced students who attended substantially higher-performing schools improve their academic outcomes. For others no change after being displaced (some change prior).

# Conclusions on Scenarios

- Use of many datasets
- Use of specific domain analytic methods
- However, domain scientists lack
  - Automatic methods for data extraction/cleaning/integration
  - Support for interacting with data and analytics
- In general, complex real-world problems are difficult to tackle

# School Closing Datasets

- Academic: CPS schools (S), School Quality Rating (SQR), demographic variables of students (D), e.g., gender, race, age, attendance/preference boundaries
- Community Profile: Chicago census data (U), e.g., population, income, poverty, infrastructure (I), e.g., roads, transportation.
- Safety: crime statistics (C), police districts (P), sex offenders and their blocks (SO), safety passages (SP).
- Background information: geospatial ontologies (location ontology, transportation ontology), time ontology, academic institution ontology

# Information Processing Methods

## ■ Visualization (V)

- The answer is not given in terms of a quantitative analysis but information is displayed so as to provide insight into the addressed question

## • Query Answering (QA)

- The answer can be computed by queries over the integrated dataset (e.g., group by + count), including geospatial queries (points, regions, trajectories)

## • Descriptive Statistical Analysis (DSA)

- The answer is given by descriptive statistical analysis, which may include methods such as correlation analysis, multivariate analysis, spatial correlation, and so on.

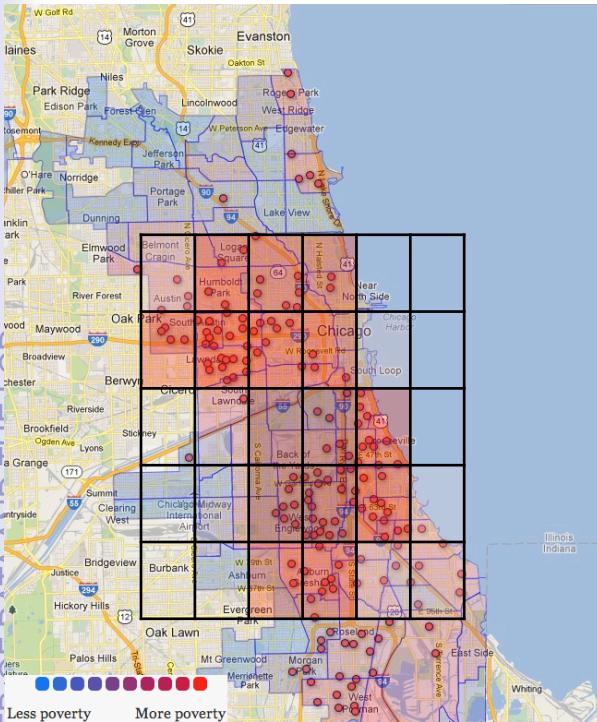
## • Predictive Statistical Analysis (PSA)

- The answer is given by making predictions about unknown variables of a problem, e.g., using machine learning methods

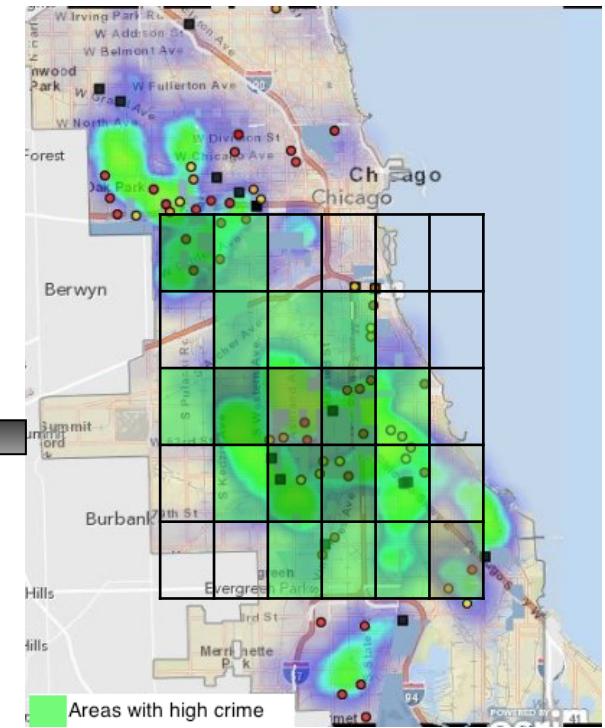
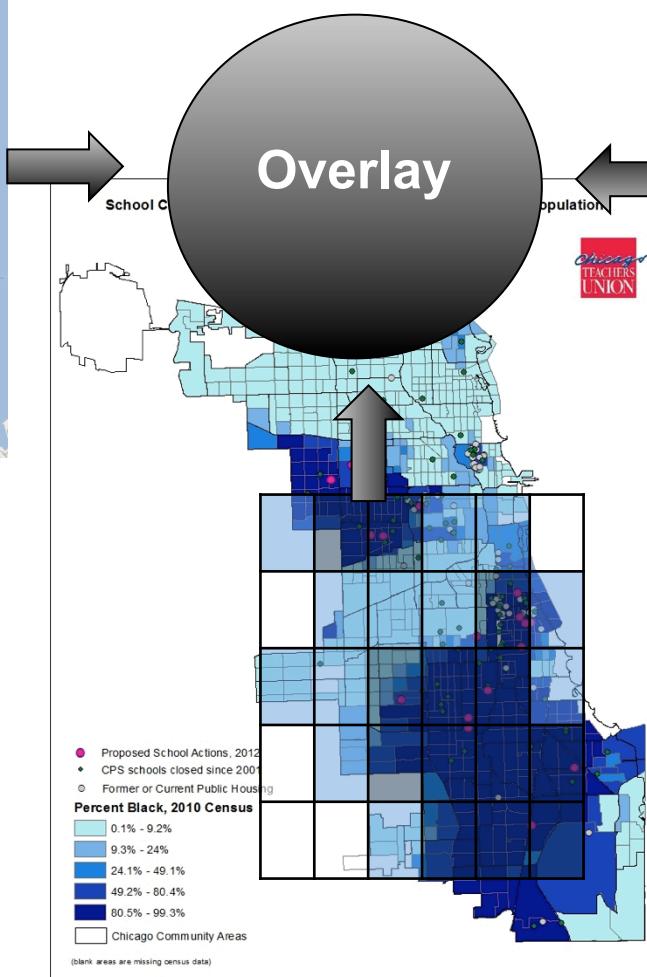
# Process Complexity

Question	Data	Geospatial Integration	Information Processing	Difficulty
Show school performance, poverty and crime rate on one map	S,SQR,C	Map Overlaying	V	
What are the characteristics of displaced students and of underperforming schools?	S,SQR	-	QA	Green
Is there a correlation between imbalanced school demography and school performance?	S,D	-	DSA	Yellow
Is there a correlation between poverty and school performance?	S,SQR, C	Transformation + Join	DSA	Yellow
Is there a correlation between violent crime rate and performance of school quality indicators?	S,SQR,C	Transformation + Join	DSA	Yellow
Do charter schools perform better than district schools under similar vulnerability conditions?	S,SQR, D, U	Transformation + Join	PSA	Red
Are safe passages close to crime areas and sex offenders?	S,SO,SP	Transformation + Join	QA, PSA	Red
Is there optimal student relocation in terms of safe passage coverage, performance, and demographic balance of the hosting school?	S,SQR, D, I	Transformation + Join	PSA	Red

# Visualization



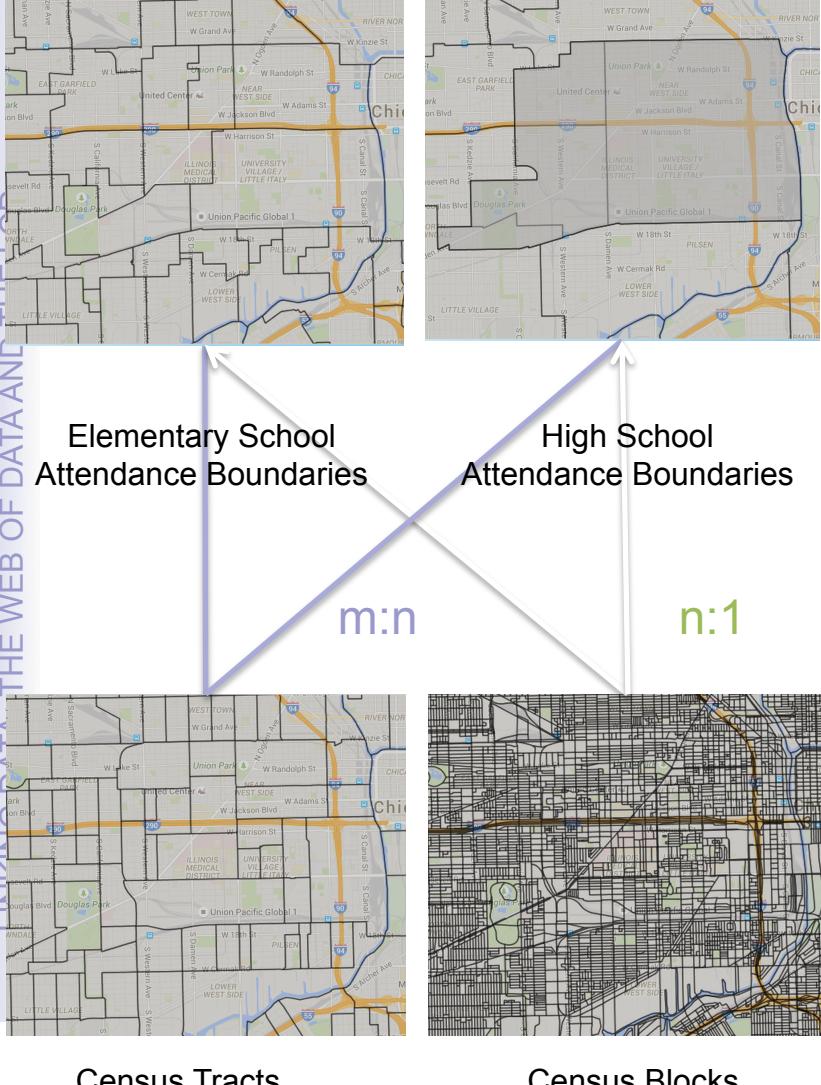
Underutilized schools in the CPS system based on poverty in the neighborhood



School closings in neighborhoods with high crime rate

# Geospatial Integration with Transformation

“Is there a correlation between poverty and school performance?”



1. Transform format (to obtain harmonized Coordinate Reference System (CRS) or geometry serialization)

E.g., Convert every point in the geometries of School Boundaries to WGS 84 CRS

2. Compute geometric similarity for 1:1 mappings (e.g., with Hausdorff Distance) or discover spatial relations for n:1/1:n/m:n mappings using RCC8 or DE-9IM (e.g., contains, within, intersects)

E.g., compute containment between blocks and attendance boundaries

# Open Data

## Il valore del dato integrato e integrabile

# Open Data: Tabelle e Spreadshit

Agenzie Turistiche Umbria

Query?

Integrazione?

	A	B	C	E	F	G	H	I
1	tipo	lingua	id contenuto	denominazione	indirizzi	telefoni	fax	email
2	Agenzia viaggio	it_IT	35362	YELLOW RABBIT TRAVEL	[ 06012   CITTA' DI CASTELLO   PG   Via E. Kant 29/G   ]	075/8511766	075/8511486	turismo@autonoleggivaltiberina.it
3	Agenzia viaggio	it_IT	35372	HERMITAGE TRAVEL FILIALE	[ 05015   FABRO   55011   TR   Via del Casermone,54   ]   [ 80763/832149   081/8706966	0763/832149   08	contabilita@hermitagetravelsrl.it   booking@h	
4	Agenzia viaggio	it_IT	35382	ADMIRE	[ 06122   PERUGIA   54039   PG   Via del Sole, 6   ]	075/5730808   3356247240	075/5730303	admire.italy@libero.it
5	Agenzia viaggio	it_IT	35392	SPAZIO E TEMPO LIBERO VIAGGI	[ 05100   TERNI   55032   TR   Via I Maggio, 13   ]	0744/59679	0744/406943	info@spaziotempolibero.it
6	Agenzia viaggio	it_IT	35402	STRAVAGARIO VIAGGI E TURISMO	[ 05018   ORVIETO   55023   TR   C. Cavour, 97   ]	0763/340890	0763/343422	stravagario@tiscali.it
7	Agenzia viaggio	it_IT	35412	INVIAGGI	[ 05100   TERNI   55032   TR   V.le della Stazione 27   ]	0744/803711	0744/817404	contabilita@inviaggi.it   info@inviaggi.it
8	Agenzia viaggio	it_IT	35422	AMBASSADOR TRAVEL SERVICE FILIALE	[ 05021   ACQUASPARTA   55001   TR   Via Roma, 9   ]	0744/930831   0541/605509	0744/930831   05	ambassador@ambassadortravel.it
9	Agenzia viaggio	it_IT	35432	MELPIDEA VIAGGI	[ 05015   FABRO   55011   TR   P.le C.Levi, 17   ]	0761/556566   3493818827	0761/1760298	melpidea@sensidiViaggio.it
10	Agenzia viaggio	it_IT	34522	IMPAEKT	[ 06062   CITTA' DELLA PIEVE   PG   Via G. Marconi, 1   ]	0578/298446   3461380007		haret@impaekt.com   info@impaekt.com
11	Agenzia viaggio	it_IT	34532	LOCCHI VIAGGI	[ 06049   SPOLETO   54051   PG   Piazza Garibaldi, 2   ]	0743/54415	0743/202677	fabio@locchiviaggi.it
12	Agenzia viaggio	it_IT	34542	ITALSPRING	[ 06121   PERUGIA   54039   PG   Via Baglioni, 12   ]	075/5731732	075/5090929	info@italspring.it
13	Agenzia viaggio	it_IT	34552	VENTO DI VACANZE	[ 05022   AMELIA   55004   TR   Via Europa, 48   ]	0744/978673	0744/060064	info@ventodivacanze.it
14	Agenzia viaggio	it_IT	34562	ANTHOS VIAGGI	[ 06126   PERUGIA   54039   PG   Via della Pallotta, 18/c   ]	075/30465	075/36194	anthos@anthosviaggi.it
15	Agenzia viaggio	it_IT	34572	I VIAGGI DI MACOS	[ 06121   PERUGIA   54039   PG   Via dei Filosofi, 25/C   ]	075/36396	075/34015	macosviaggi@tiscali.it
16	Agenzia viaggio	it_IT	34582	MENIGATTI	[ 06132   PERUGIA   54039   PG   Via del Pettiroso, 4   ]	075/5149707	075/774611	amministrazione@menigattiviaggi.it

Significato dei dati?

Eventi

	A	B	C	D	E	F	G	H	I	J	K	L
1	tipo	lingua	id contenuto	id contenuto	url risorsa	titolo	descrizione					
2	Evento	it_IT	2916883	90470	http://www.umbriatourism.it/-/it-fits	IT FITS	Il 25 settembre, ad Assisi, la terza edizione del Forum Italiano sul Turismo e la Sostenibilità.	25/09/15		26/09/15		
3	Evento	it_IT	2032108	90422	http://www.umbriatourism.it/-/il-lusso-del-s	Il Lusso del Sonno	Fino al 31 agosto all'Antiquarium di Corciano la mostra dedicata ai letti etruschi in bronzo rinvenuti	21/08/15		31/08/16		
4	Evento	it_IT	2032245		http://www.umbriatourism.it/-/ecobike	Ecobike	Fino a ottobre escursioni e visite guidate attraverso l'Umbria in sella a una bicicletta elettrica.	21/08/15		31/10/15		

	M	P	Q	T	U	V	W	X	Y	Z	AA	AB
1	titolo testo	caregorie evento	immagini evento	immagine copertina evento	testo alternativo immagi	latitudine	longitudine	comune	codice IS	via o piazza	localitv†	cap
2	Forum Italiano sul Turismo e la Sostenibilità	Mostre e rassegne   Assisi	http://www.umbriatourism.it	http://www.umbriatourism.it	IT FITS	430.707.017	12.619.596.600.000.000	Assisi	54001			
3	I letti in bronzo di Strozziacapponi	Mostre e rassegne   Corciano	http://www.umbriatourism.it	http://www.umbriatourism.it	Il lusso del Sonno	431.237.872	12.289.259.600.000.000	Corciano	54015			
4	Alla scoperta dell'Umbria con la bici elettrica	Eventi sportivi	http://www.umbriatourism.it	http://www.umbriatourism.it	Ecobike							

# Open Data: Tabelle e Spreadsheet

## ■ Query & Integrazione

- CSV non interrogabili così come sono
- Esplorazione e manipolazione richiede applicazioni
  - Locali / Non web friendly
- Integrazione manuale (vedi Open Refine)

## ■ Semantica

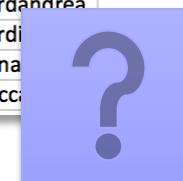
- Ambiguità: tipo (evento) vs tipo (agenzia)
- Tipizzazione debole: e.g., codici
- Vocabolari locali e non interpretabili da terze parti

### \*REMARK\*

La finalità degli Open Data è proprio quella di poter essere usati da *terze parti*, tipicamente con conoscenza approfondita dell'organizzazione che ha prodotto i dati

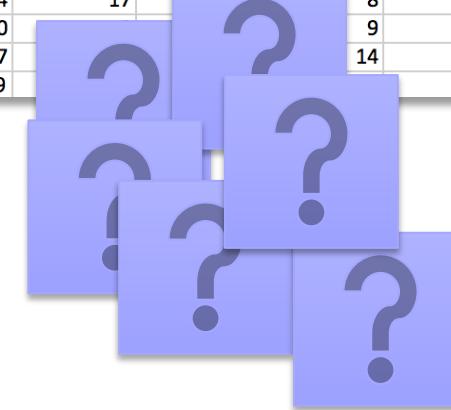
# ISTAT Open Data: CSV file

COMUNE	PROCOM	LOC2011	CODLOC	LOCALITA	TIPOLOC
Assisi	54001	5400110001	10001	Armenzano	1
Assisi	54001	5400110002	10002	Assisi	1
Assisi	54001	5400110003	10003	Capodacqua	1
Assisi	54001	5400110004	10004	Case Nuove	1
Assisi	54001	5400110005	10005	Castelnuovo	1
Assisi	54001	5400110006	10006	Palazzo	1
Assisi	54001	5400110007	10007	Petrignano	1
Assisi	54001	5400110008	10008	Pianello	1
Assisi	54001	5400110009	10009	Rivotorto	1
Assisi	54001	5400110010	10010	Santa Maria	1
Assisi	54001	5400110011	10011	San Vitale	1
Assisi	54001	5400110012	10012	Sterpeto	1
Assisi	54001	5400110013	10013	Torchiagina	1
Assisi	54001	5400110014	10014	Tordandrea	1
Assisi	54001	5400110015	10015	Tordigliano	1
Assisi	54001	5400120007	20007	Rena	2
Assisi	54001	5400120008	20008	Rocca	2



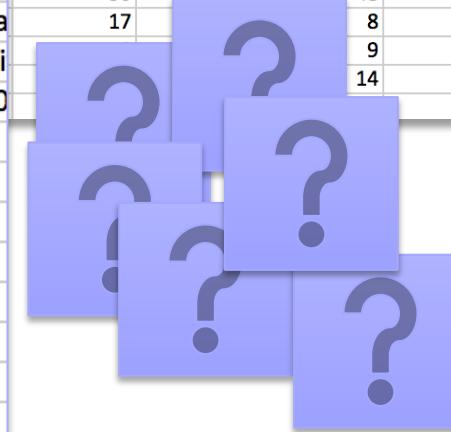
...

P1	P2	P3	P4	P5	P6	P7
33	21	12	8	16	1	5
3732	1690	2042	1664	1526	77	378
142	68	74	49	69	2	17
634	316	318	237	306	17	62
791	385	406	307	394	11	68
1817	866	951	721	890	41	137
3184	1538	1646	1200	1624	68	229
107	52	55	40	55	1	11
1425	687	738	513	734	27	122
7719	3710	4009	3336	3442	164	597
603	287	316	254	285	9	46
13	8	5	6	3	2	1
655	319	336	247	332	12	60
939	480	459	357	483	17	64
293	137	156	102	151	10	23
108	50	58	48	48	0	10
34	17	8	15	0	0	10
20		9	7	0	0	1
27		14	8	1	1	4
119				69	2	6



# ISTAT Open Data: CSV file

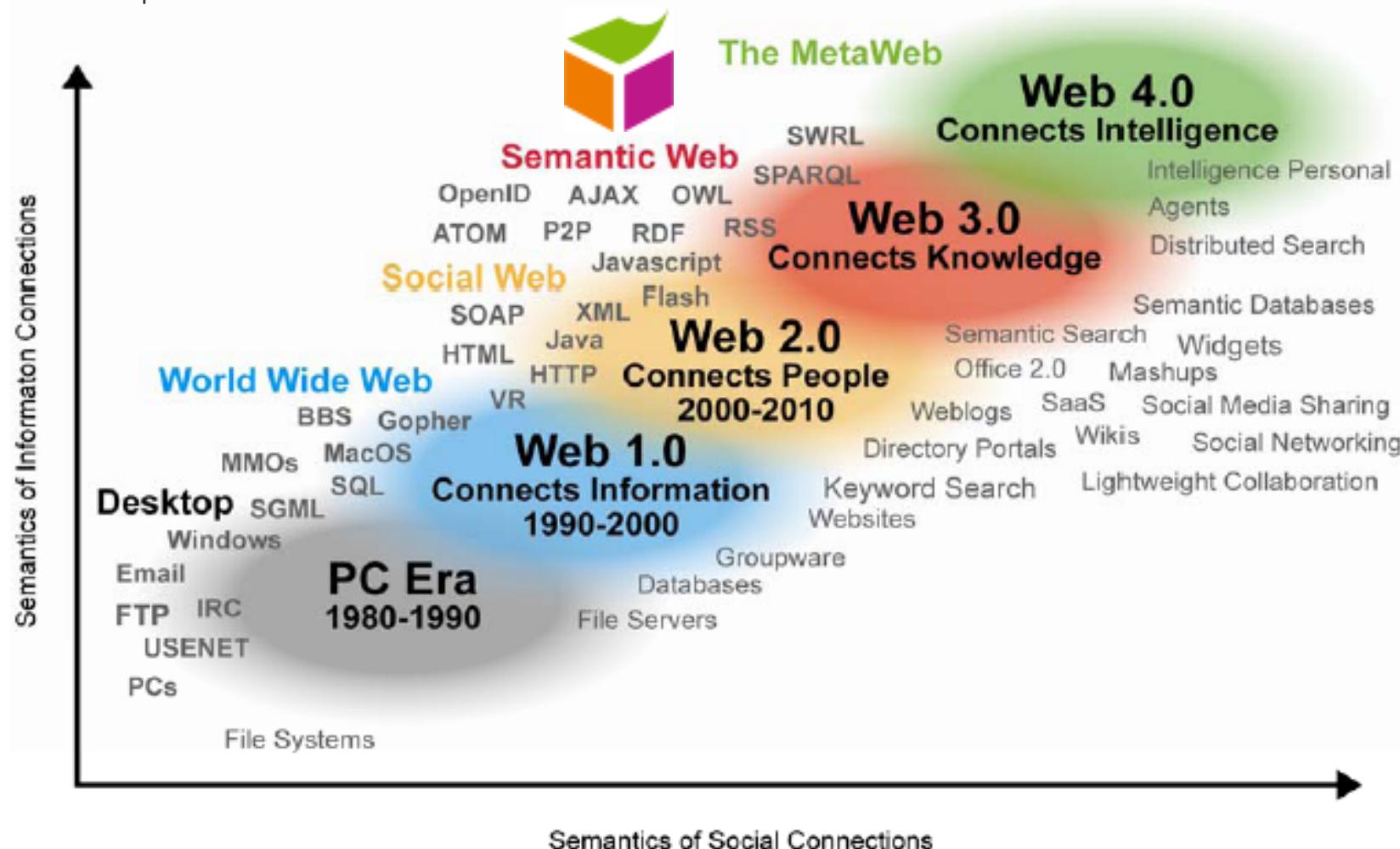
COMUNE	PROCOM	LOC2011	CODLOC	LOCALITA	TIPOLOC	P1	P2	P3	P4	P5	P6	P7
Assisi	54001	5400110001	10001	Armenzano	1		21	12	8	16	1	5
Assisi	54001	5400110002	10002	Assisi	1		1690	2042	1664	1526	77	378
<b>Nome Campo Definizione</b>												
CODREG	Codice numerico che identifica univocamente la regione nell'ambito del territorio nazionale.											
REGIONE	Denominazione della regione											
CODPRO	Codice numerico che identifica univocamente la provincia nell'ambito della regione.											
PROVINCIA	Denominazione della provincia											
CODCOM	Codice numerico che identifica univocamente il comune nell'ambito della provincia.											
COMUNE	Denominazione del comune											
PROCOM	Codice numerico che identifica univocamente il comune nell'ambito del territorio nazionale.											
LOC2011	Codice numerico che identifica univocamente la località 2011 nell'ambito del territorio nazionale.											
CODLOC	Codice numerico che identifica la località 2011 nell'ambito del territorio nazionale.											
LOCALITA	Denominazione della località 2011											
TIPOLOC	Codice numerico che identifica la tipologia della località 2011. Il codice è composto da due cifre.											
AMPLOC	Codice numerico che identifica l'ampiezza demografica della località 2011.											
CAPOLUOGO	Codice numerico valorizzato a 1 nel caso di centro capoluogo e a 0 nei casi di frazioni.											
ALITUDINE	Altitudine della località 2011											
P1	Popolazione residente - Totale											
P2	Popolazione residente - Maschi											
P3	Popolazione residente - Femmine											
P4	Popolazione residente - Celibi/nubili											
P5	Popolazione residente - Coniugati/e (+ separati/e di fatto)											
P6	Popolazione residente - Separati/e legalmente											
P7	Popolazione residente - Vedovi/e											
P8	Popolazione residente - Divorziati/e											
P9	Popolazione residente - Maschi celibi											
P10	Popolazione residente - Maschi coniugati o separati di fatto											



Spiegazione in file separato

# The Web: Evolution

Source: <http://www.radarnetworks.com>

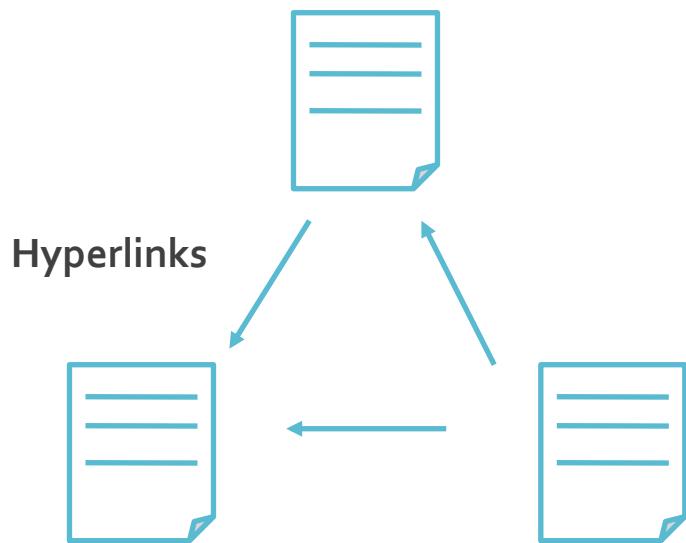


# The Web: Evolution

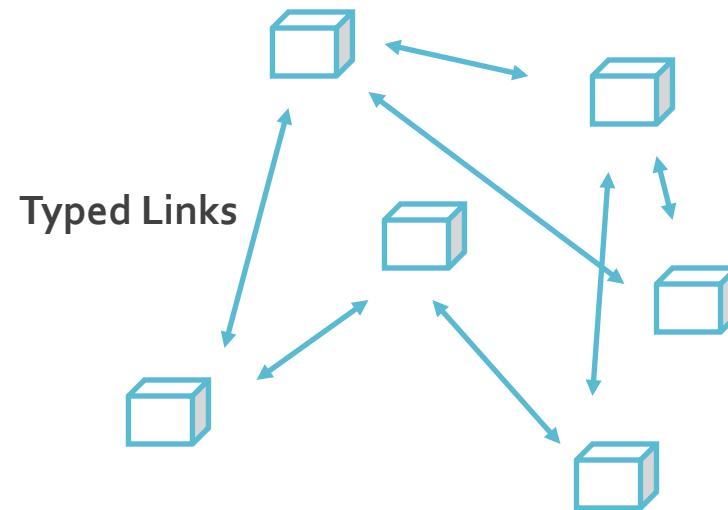
Web of Documents



Web of Data



"Documents"



"Things"

# Dai documenti ai dati

- Entità e relazioni
- Entità e attributi
- Processabili da una macchina, supporto a interrogazioni (complesse)
- Integrazione di testi, immagini, video con dati strutturati
- Valore del dato di qualità
- Politiche sui dati di ampio respiro (costruzione di knowledge graph): grandi investimenti per alti ricavi



# Linked Data

# Linked Data

- Set of best practices for **publishing data on the Web**
- Data from different knowledge domains, self-described, linked and accessible
- Follows 4 simple principles...

