**International Conference on Information Quality (ICIQ)**
Paris, France – 16 November 2012

# Big Data Must Overcome Big Data Quality Challenges

*by*

***Prof Stuart Madnick***
***Sloan School of Management***
***& Systems Engineering Division***
***Massachusetts Institute of Technology***

# Agenda

## 4 Parts

1. Motivation for “Big Data”
2. Evolution of Data Quality research
3. Importance of Data Quality to Big Data
4. Ending: (Slight) Connection between Data Quality and History of France

**For start: What is Big Data?**

# Part 1: McKinsey strategic consulting firm: "Big Data is the next frontier for innovation, competition, and productivity" [1]

- "The amount of data in our world has been exploding
- So called, 'Big Data,' will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus …
- Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers…
- The rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future. …"

[1] http://www.mckinsey.com/mgi/publications/big_data/

# It must be important !

# New Tools Beget Revolutions

# What is Big Data?

## The V's of Big Data

- **Volume** – Large quantities of data
- **Velocity** – Speed to digest and generate results
- **Variety** – Diverse sources and types of data
  - Types: Structured, Unstructured, Semi-structured
- **Veracity** – The quality and life-cycle of data } Later
- **Value** – Does the data have any value?
  - Design of experiments
  - Evaluate results

# New Sources of Data (some examples)

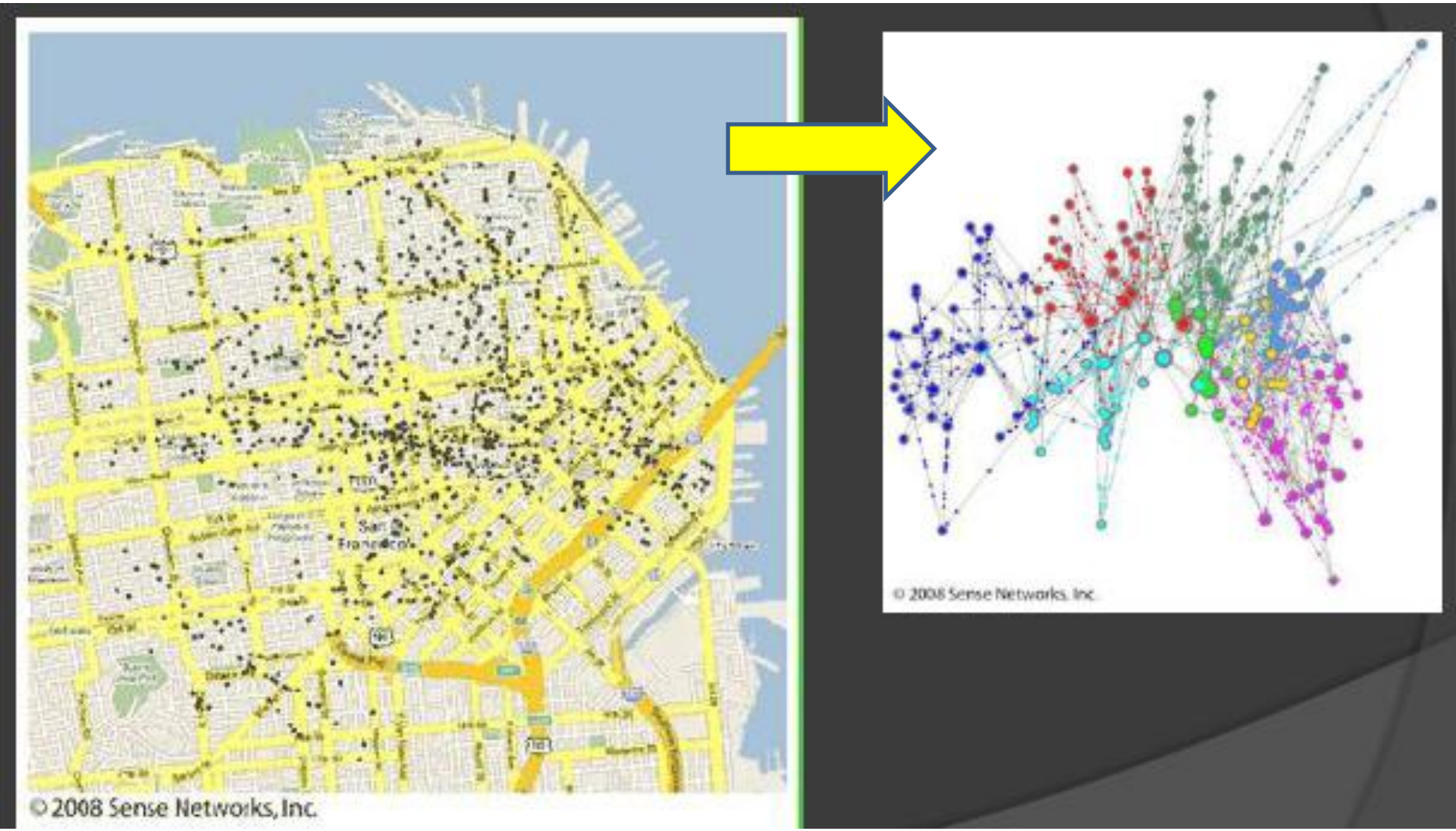
- Web traffic: Clickstream/ Page views/ Web activities
- Web links/ Blog references
- Search engines: Google/ Bing/ Yahoo
- Social media: Facebook / Twitter feeds
- Location and Activity: Mobile phone/ GPS
- Email messages
- Transactions: ERP/ CRM/ SCM
- RFID (Radio Frequency Identification), Bar Code Scanner
- Real-time: Machinery diagnostics/ engines/ equipment
- Automated scientific equipment: DNA sequencers
- Financial transactions: Stock markets / foreign exchanges
- User generated content: Wikipedia updates
- Open Linked Data
- Online repositories
- Etc….

# Search Engines: The Future of Prediction

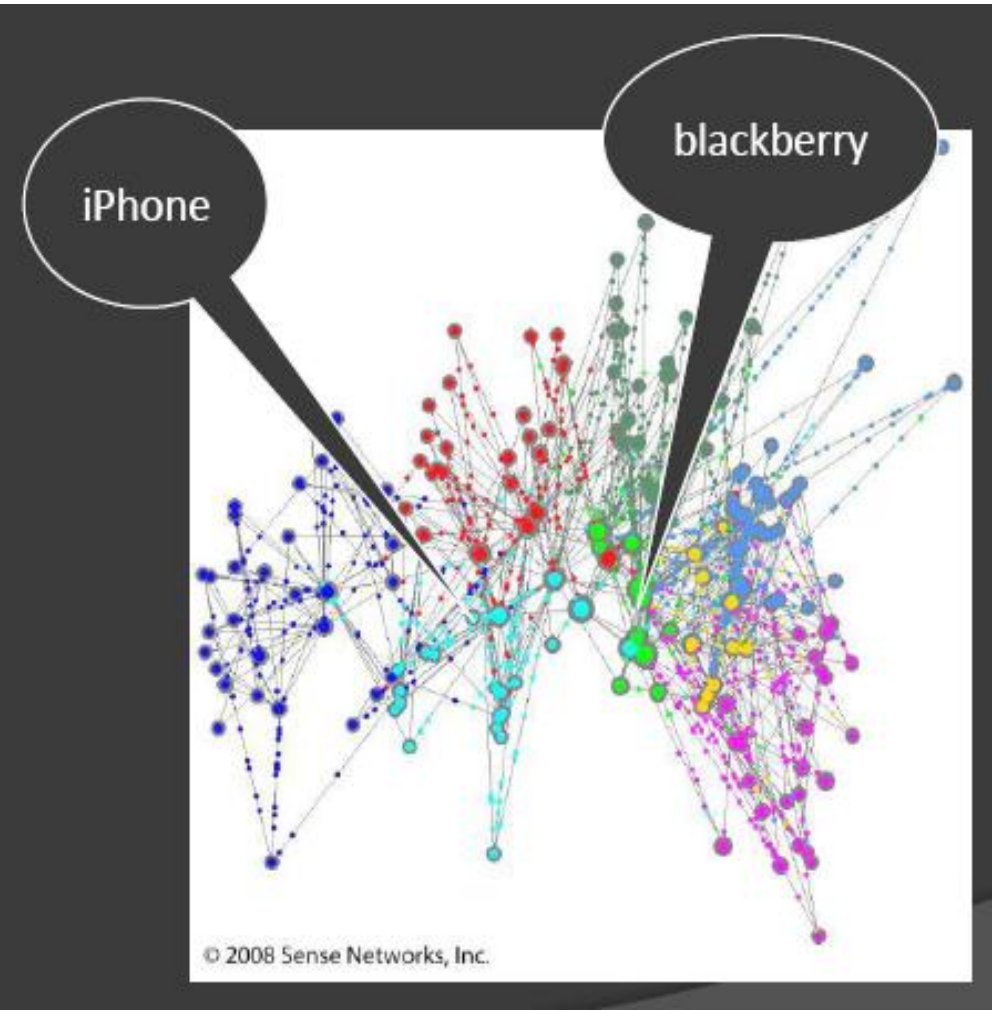
- **Insight: *We know what you are thinking!***
- Google Search Foreshadow Housing Prices and Sales
  - Work by Lynne Wu and Erik Brynjolfsson, see http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2022293
- Data sources used in economics have substantial lag and high level of aggregation -> difficult to use for real-time predictions
- Data from search engines like Google provide highly accurate way to predict future business activities
- *Example*: Predict housing market trends, each % increase in housing search index is correlated with added sales of 67,220 houses in next quarter
  - Produced much better results than conventional models
- This is a form of ***implicit*** Collective Intelligence
- Also used to identify emerging "hot" research – before known (from analyzing 100,000's of research reports)

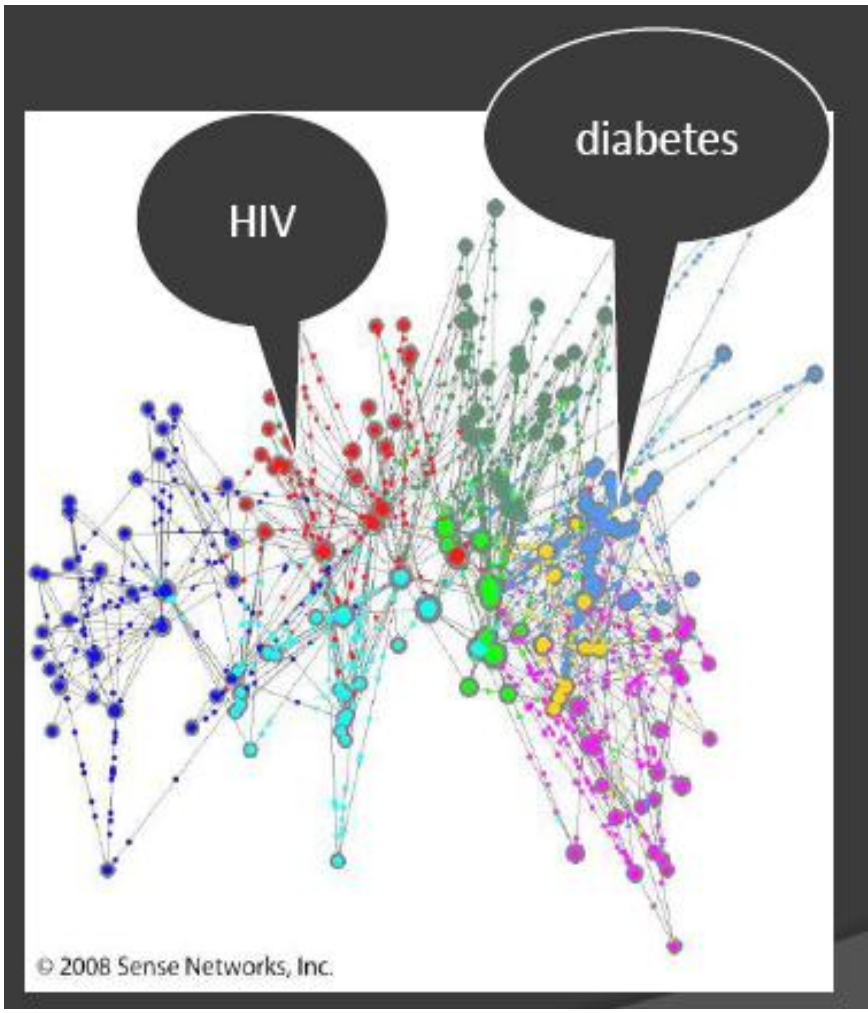
# Location Data: Tracked over Time / Commonalities

- **Insight:** ***We know where you are and what you doing!***

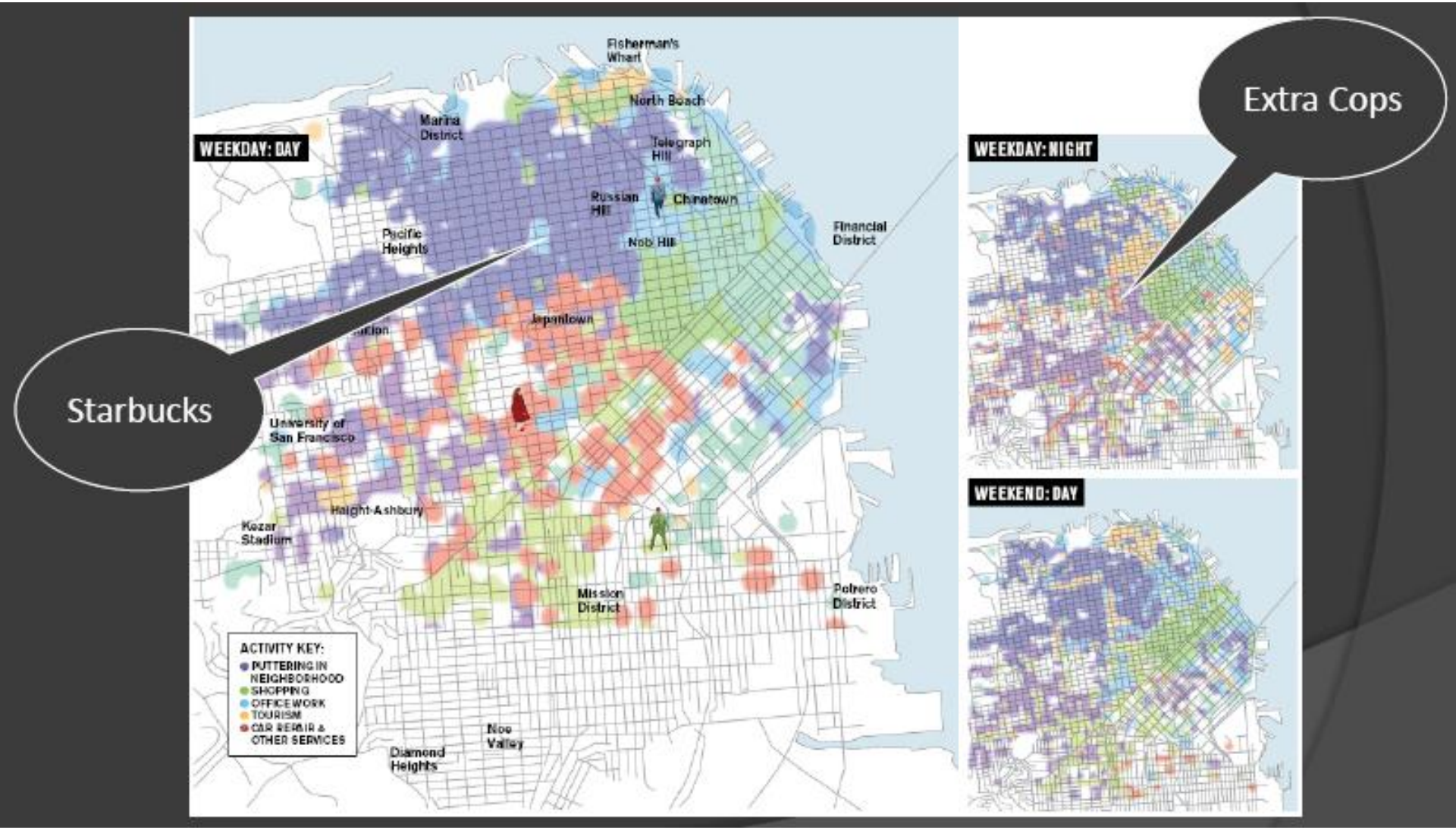


# Commonalities Discovered



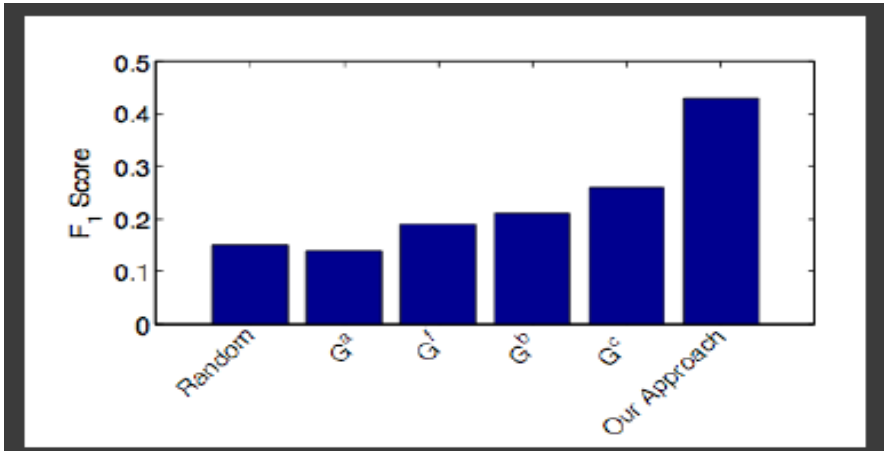



# Use of Patterns for Planning



# Detailed Sensor & Social Data

- **Insight: We may be able to know things about you that even you don't (yet) know …**
- Using accelerometer and other smart phone sensors:
  - Anticipate medical developments, e.g., depressions, post trauma distress syndrome, some mental illness
- For certain types of products (e.g., smartphone apps), using prior purchase behavior and friends' behavior:
  - Anticipate what you will buy next …



# Some potentially controversial uses …

- **Insight: There are many things that can be learned about you by studying your social network …**

HOME / GLOBE / IDEAS

The Boston Globe

## Project 'Gaydar'

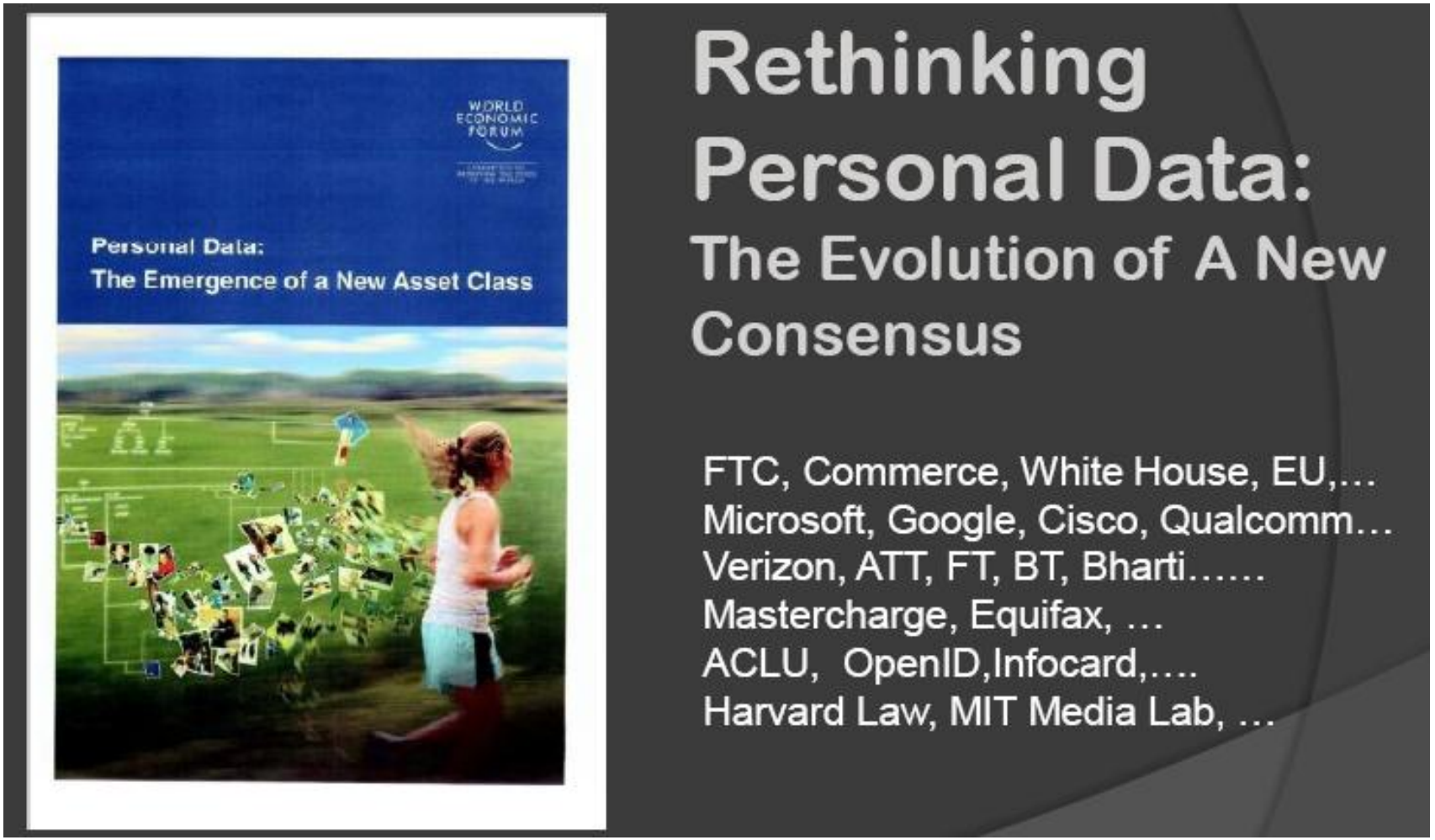
At MIT, an experiment identifies which students are gay, raising new questions about online privacy

By Carolyn Y. Johnson
Globe Staff / September 20, 2009

E-mail | Reprints | Text size

It started as a simple term project for an MIT class on ethics and law on the electronic frontier. (Full article: 1768 words)

# Social Media Data

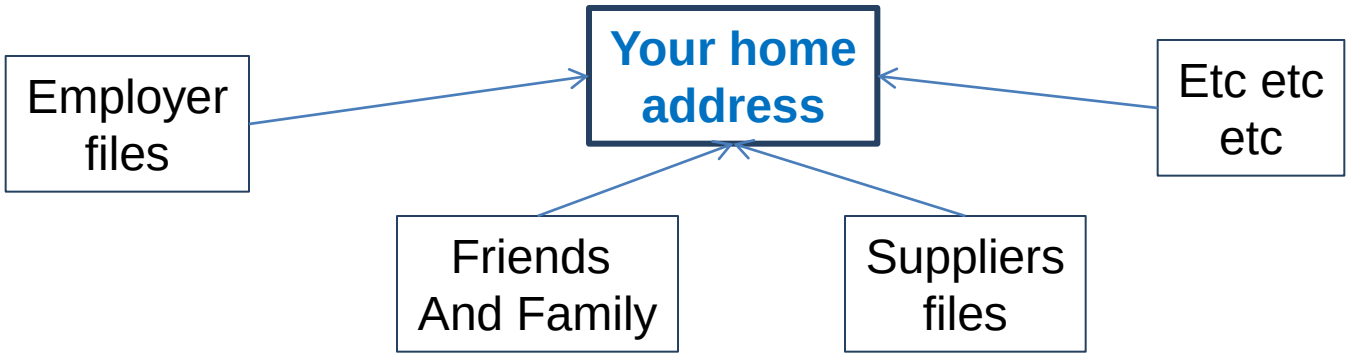



**"Personal data is the new oil of the Internet and the new currency of the digital world",**

Meglena Kuneva, European Consumer Commissioner

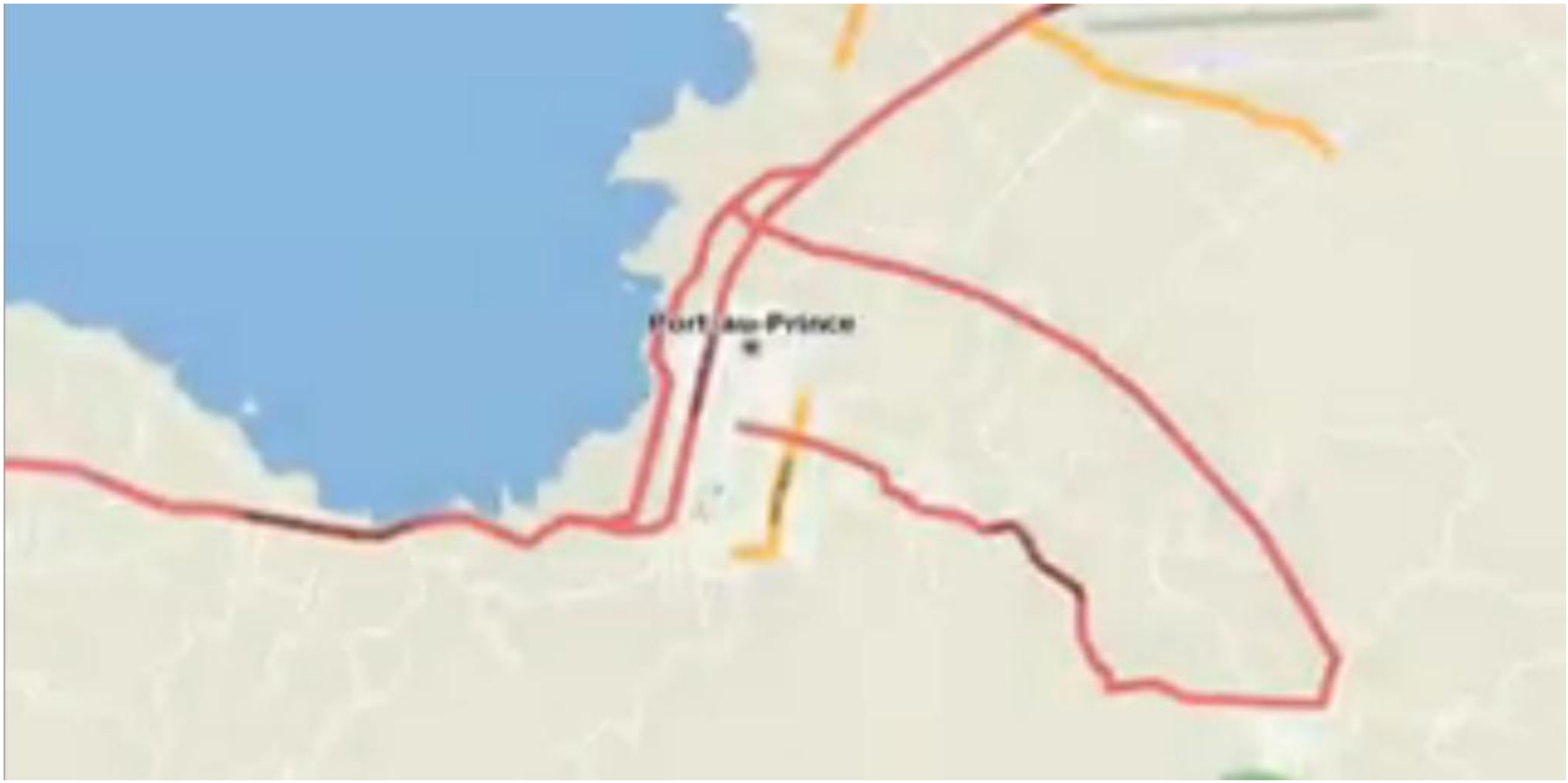
# "Open Linked Data" – What's the big deal?

- How many places does your home address appear?
  - Employer's files (multiple places)
  - Friends and family (many  many)
  - Suppliers (telephone, electric, cable TV, etc., etc., etc.)
- What if you move?
  - How long before all are up-to-date? (If ever …)
- Linked Data approach




- **Insight: "Open Linked Data" – You can contribute**
  - ➢ Like Wikipedia, but for Data

# "Open Linked Data" – What's the big deal?

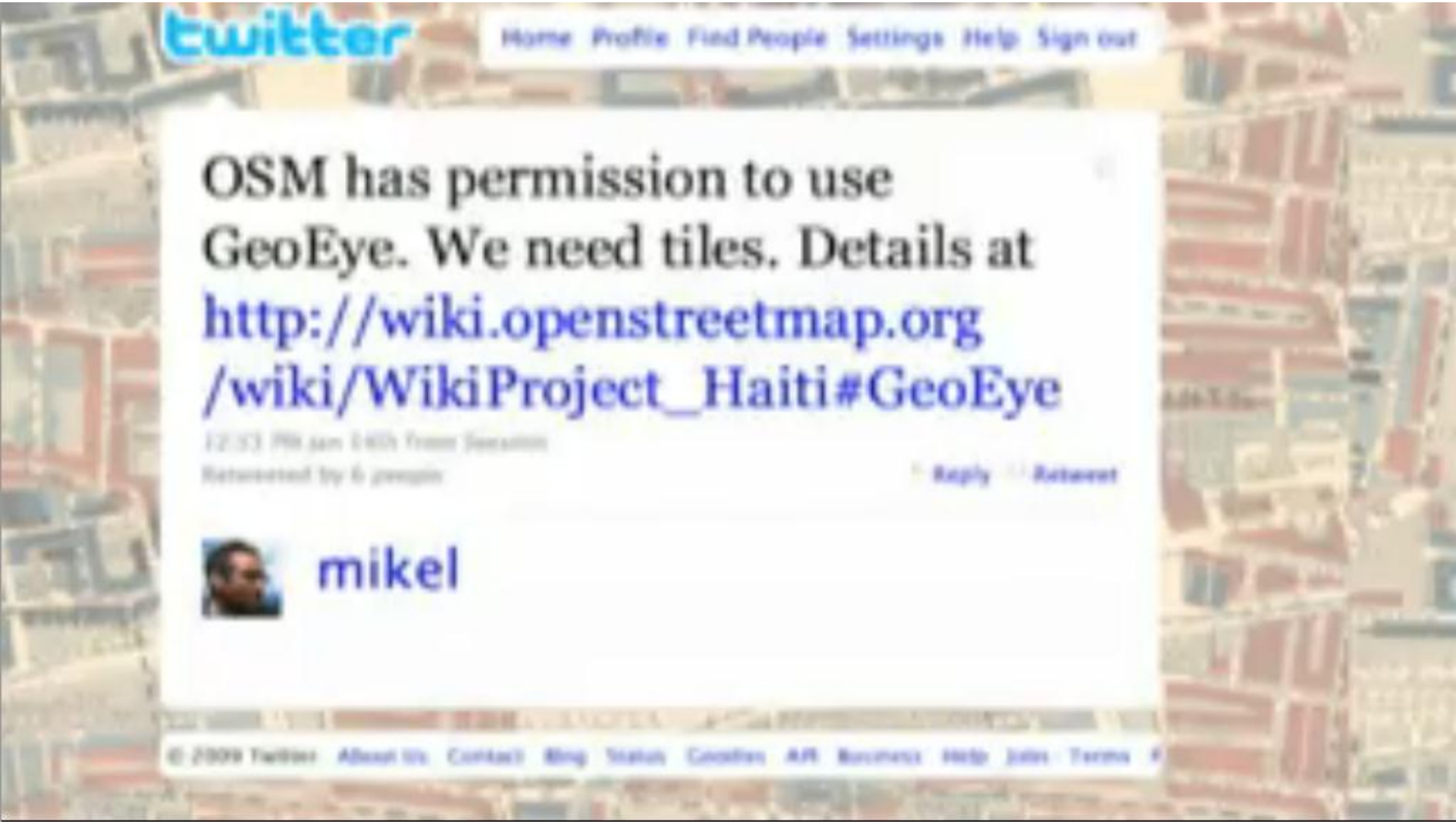



- **Terrible Quality of the map of Port au Prince, Haiti at time of 2010 earthquake**
  - **Many roads missing and unnamed, buildings not identified (hospitals, hotels), out of date information (refugee camps)?**



See http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide.html from 4:10 to 4:51)

# Open Street Map (OSM) Project

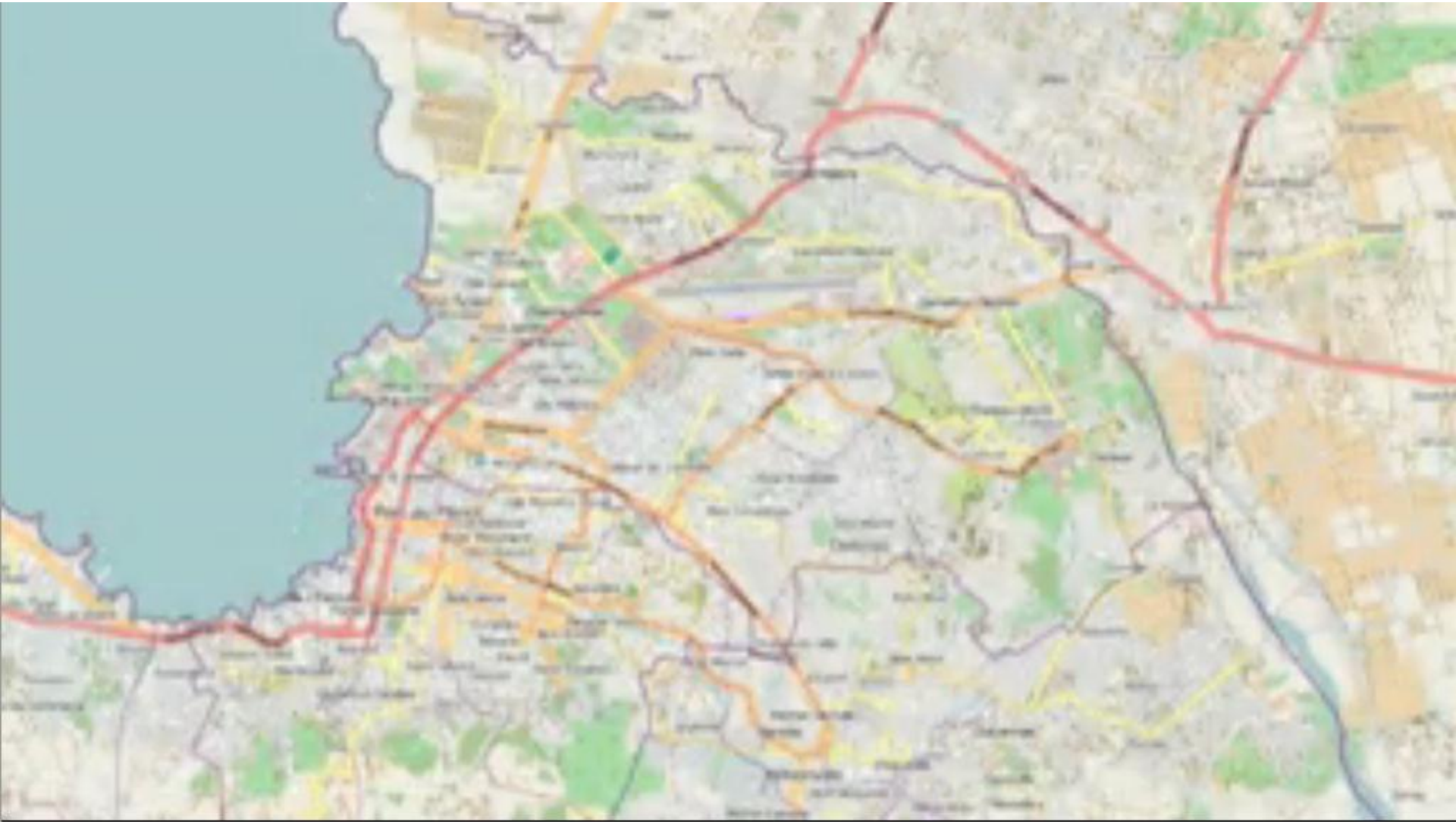


# Each “Dot of Light”

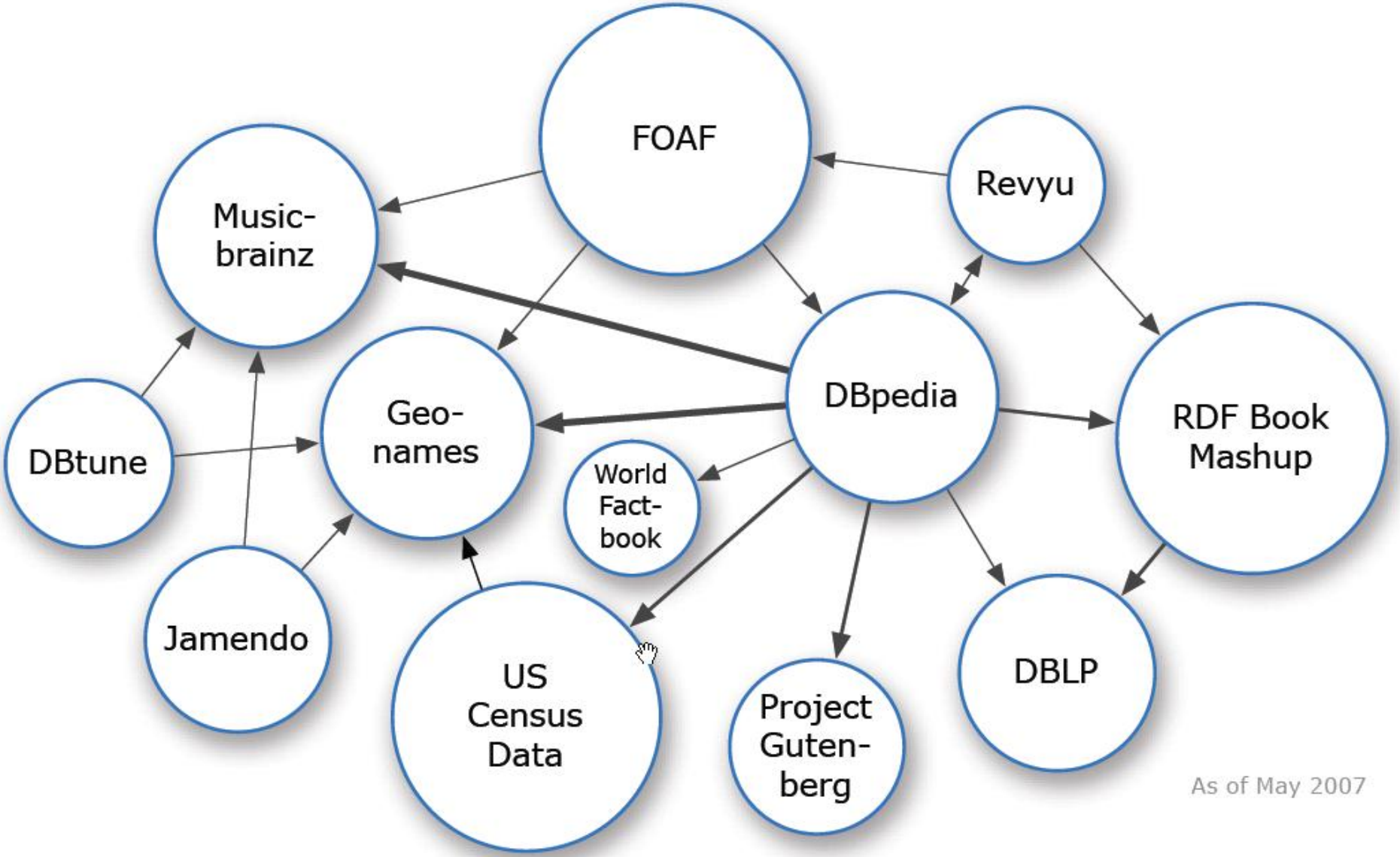
## Someone in the World Adding Detail to the Map

# Resulting Map of Port au Prince, Haiti

## Roads added and named, buildings identified, up-to-date

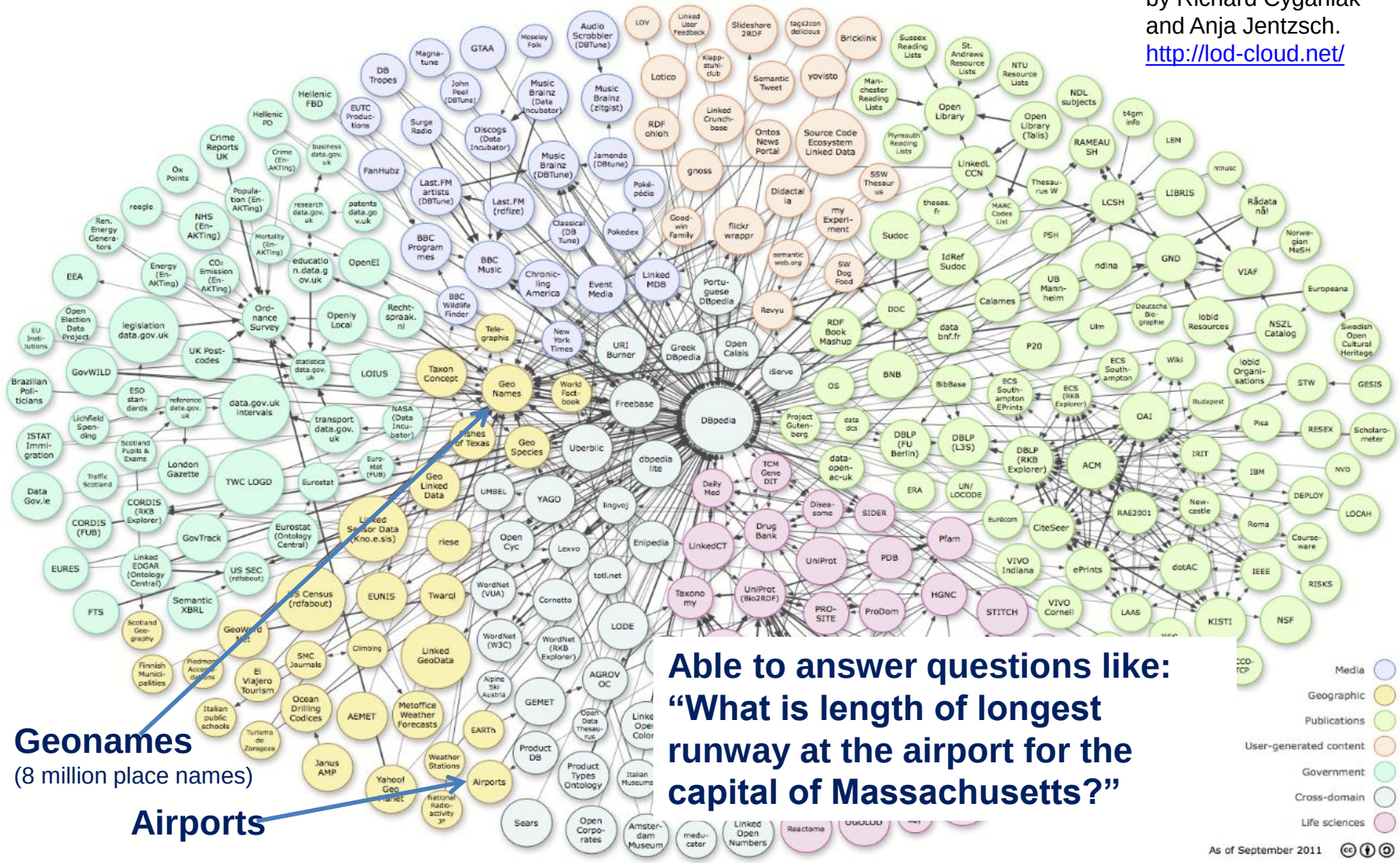


# Open Linked Data Movement:
# The Linked Data Web, May 2007



# Open Linked Data Movement:
# The Linked Data Web, 2011

by Richard Cyganiak and Anja Jentzsch.
http://lod-cloud.net/



# Big Data < == > New Sources of Data

- Web traffic: Clickstream/ Page views/ Web activities
- Web links/ Blog references
- Search engines: Google/ Bing/ Yahoo
- Social media: Facebook / Twitter feeds
- Location and Activity: Mobile phone/ GPS
- Email messages
- Transactions: ERP/ CRM/ SCM
- RFID (Radio Frequency Identification), Bar Code Scanner
- Real-time: Machinery diagnostics/ engines/ equipment
- Automated scientific equipment: DNA sequencers
- Financial transactions: Stock markets foreign exchanges
- User generated content: Wikipedia updates
- Open Linked Data
- Online repositories
- Etc….

# Part 2: Evolution of Data & Information Quality (DQ/IQ)


**Rich Wang**
**(our Harry Potter)**

*Lots of time & energy*

## Journals *

**- 2007 ACM Journal on Data and Information Quality (JDIQ)**

## Conferences and Certification Programs *

- 1996 International Conference on Information Quality (ICIQ)
- 2002 MIT-IQ program for Executives
- 2003 IQ-1: Principles and Foundations
- 2007 IQ Industry Symposium
- **2012 ICIQ #17 Paris …**

## Education

MS IQ and IQ PhD Degree Programs

## Books *

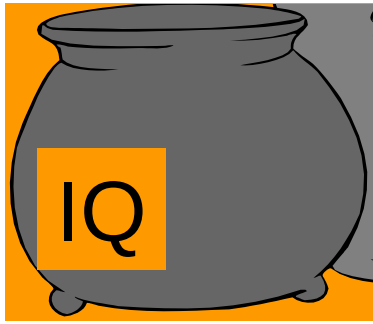
- *Information Quality & Knowledge* (1999)
- *Data Quality* (2000)

## Articles *

- 1990 Polygen Data Quality Model (VLDB + ICIS)
- 1996 Beyond Accuracy
- 1998 Managing Information as a Product, etc …

## Research Projects *

- 1988 Total Data Quality Management Program (TDQM)
- 2002 MIT Information Quality (MITIQ) Program






* Not complete list

# Some Data Quality Research Areas

- **Data Quality is multi-dimensional**
- **Organizational Data Quality assessment**
- **Interplay of Data Quality and Data Semantics**

- Manage information as a product
- Data integrity analysis
- Data Quality root cause analysis
- Data Source/Provenance – mathematics of DQ

# What is Data Quality?

- **Naïve / Conventional view:**

  Data Quality = Accuracy

- **Research finding:**

  Data Quality *Goes Beyond* Accuracy

  Initial survey of data users resulted in over 100 different data quality dimensions!

- **What are some other dimensions ?**

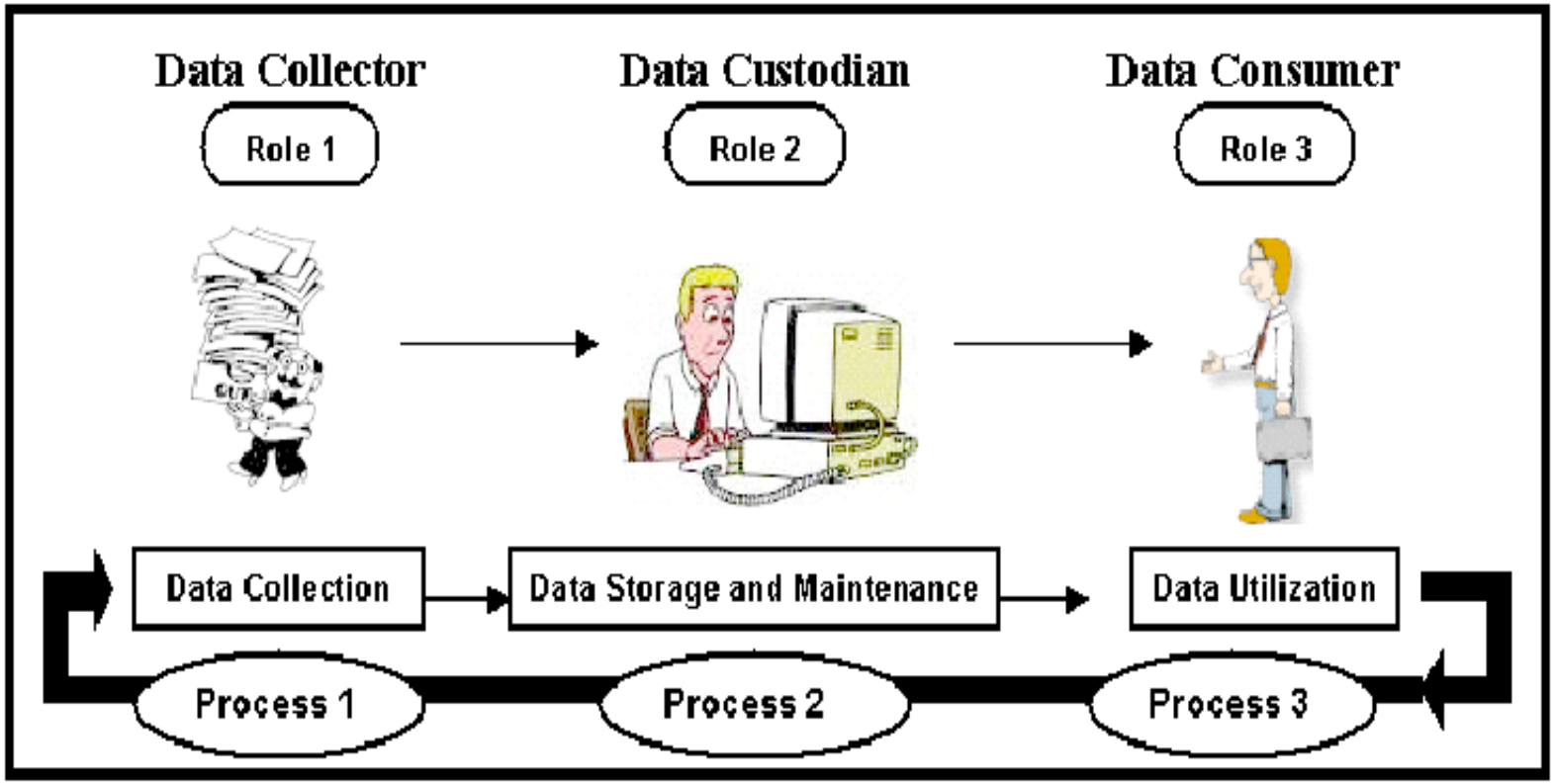
# Data Quality Dimensions:

## 16 Key dimensions, organized into 4 categories

| DQ Category | DQ Dimensions |
|---|---|
| Intrinsic DQ | Accuracy, Objectivity, Believability, Reputation |
| Accessibility DQ | Access, Security |
| Contextual DQ | Relevancy, Value-Added, Timeliness, Completeness, Amount of data, Ease of manipulation |
| Representational DQ | Interpretability, Ease of understanding, Concise representation, Consistent representation |

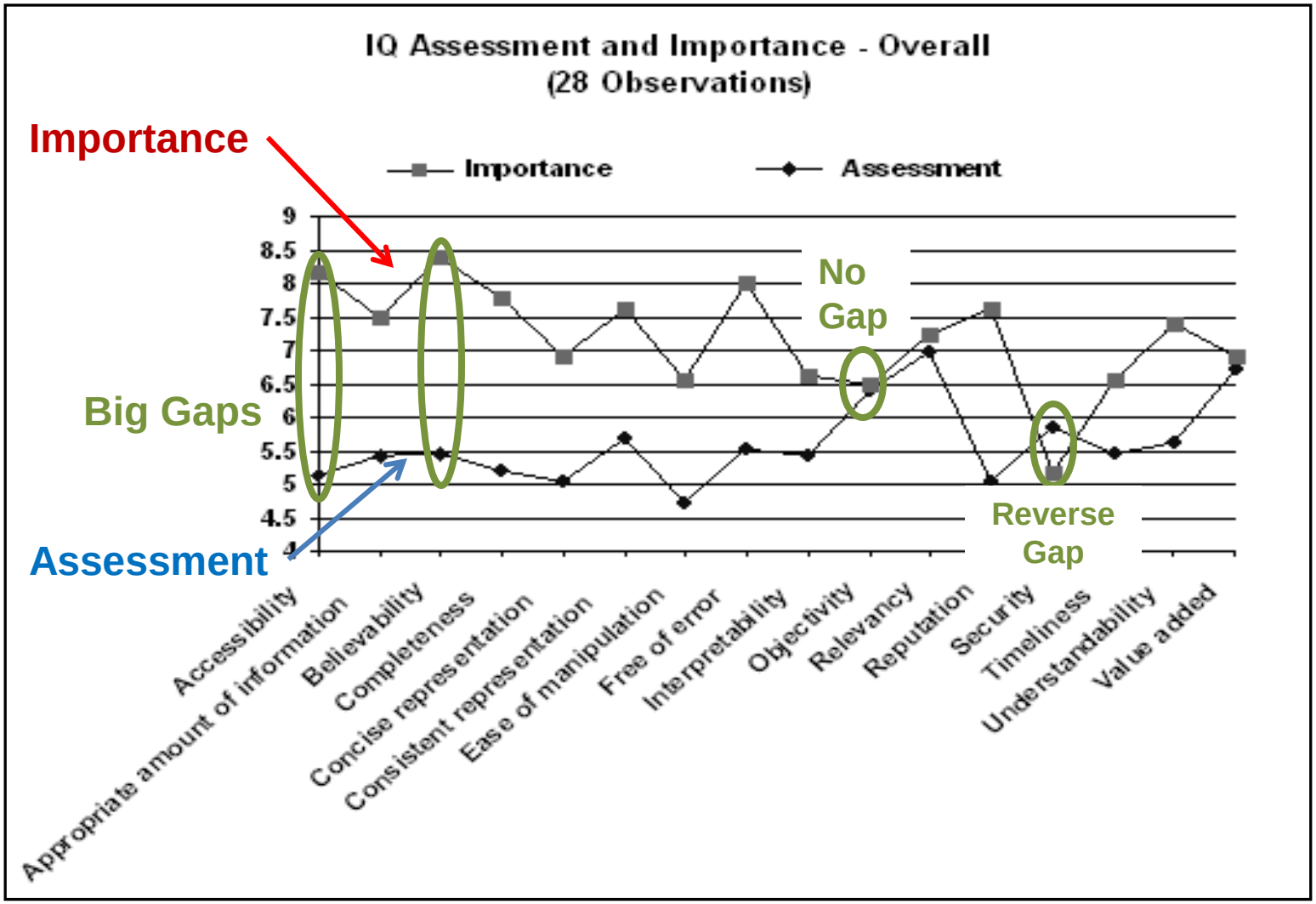
# Organizational DQ Assessment

## Many different roles in involved with data in an organization …

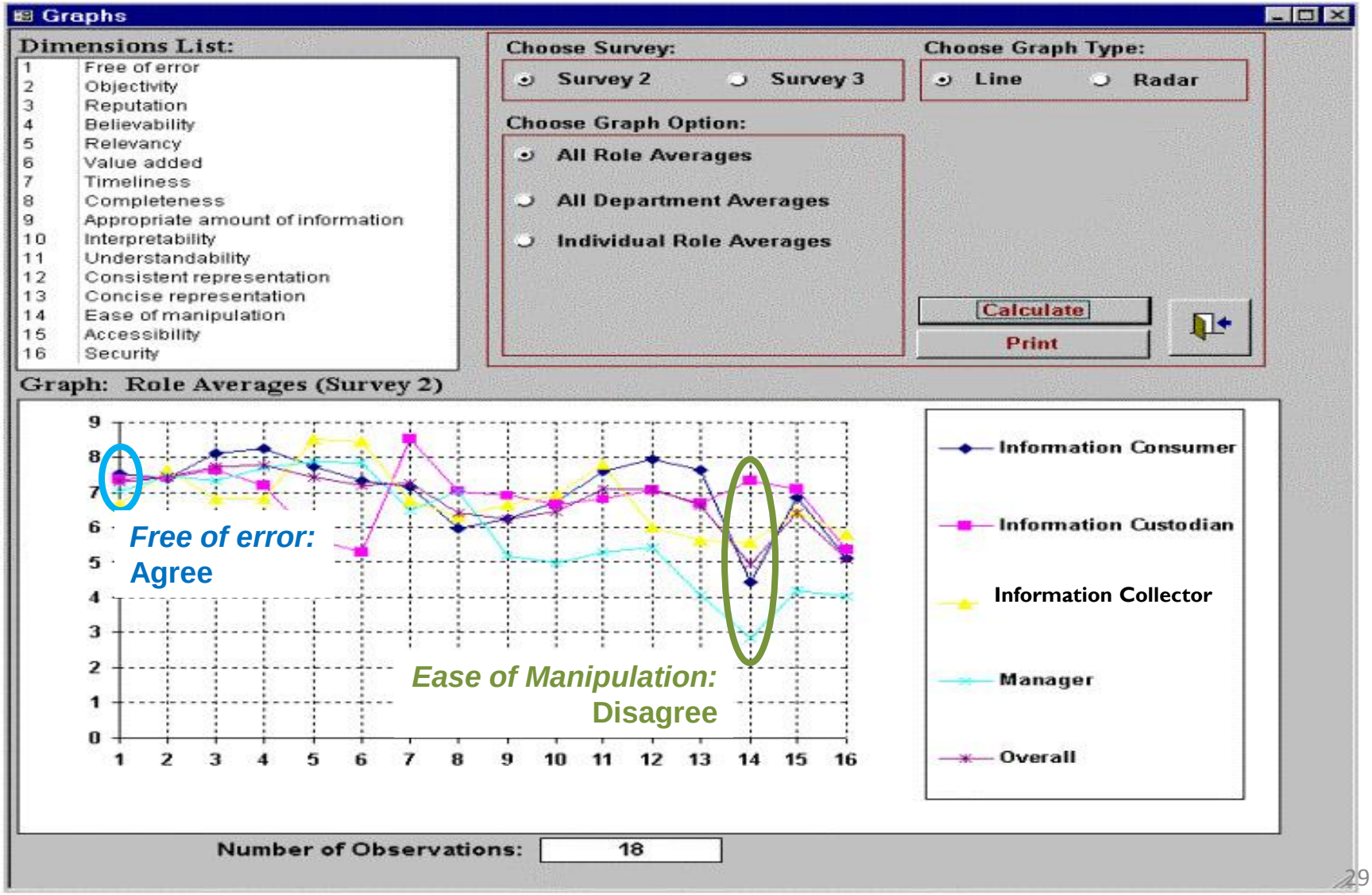



- **Method: Questionnaire to Assess Perceptions of Data Quality**
- **Analysis: Statistical Significance, Statistical Reliability and Statistical Validity (**Convergent Validity and Discriminant Validity)

# Organizational DQ Assessment: Some sample results



# Organizational DQ Analysis: Some sample results



# Interplay of Data Quality and Data Semantics

## Daimler Benz（DCX）Financial Data

| Source | P/E Ratio |
|---|---|
| ABC | 11.6 |
| Bloomberg | 5.57 |
| DBC | 19.19 |
| MarketGuide | 7.46 |

**Which one is correct? Why?**

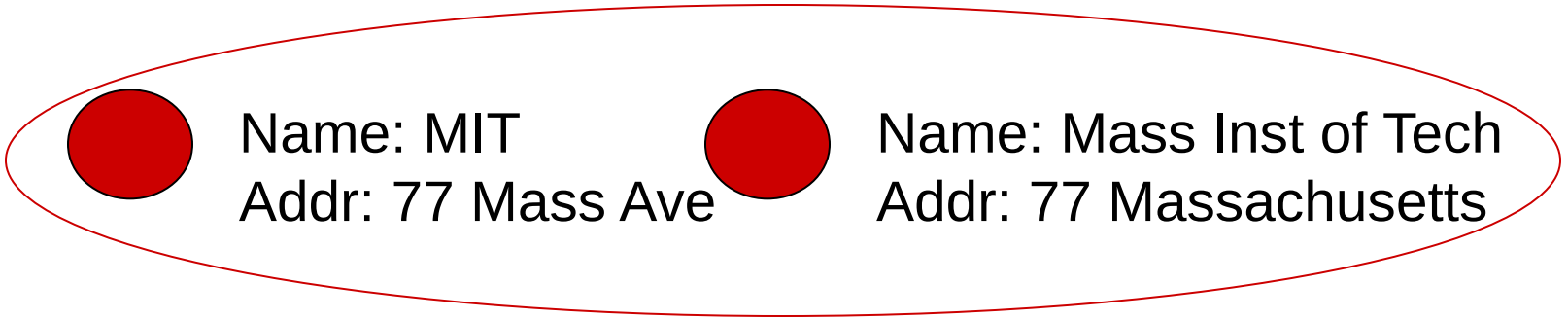
# More complex "simple" Example Questions

- Simple questions:
  - "How much did Merrill Lynch loan to IBM last year ?"
  - "How many employees does IBM have ?"
  - **"How many faculty does MIT have?"**

  - "How much did MIT buy from IBM last year ?"
  - "How much did IBM sell to MIT last year ?"

    *[ Do you expect the answers to be the same ?]*

# a. Identical entity instance identification (Record Linkage)



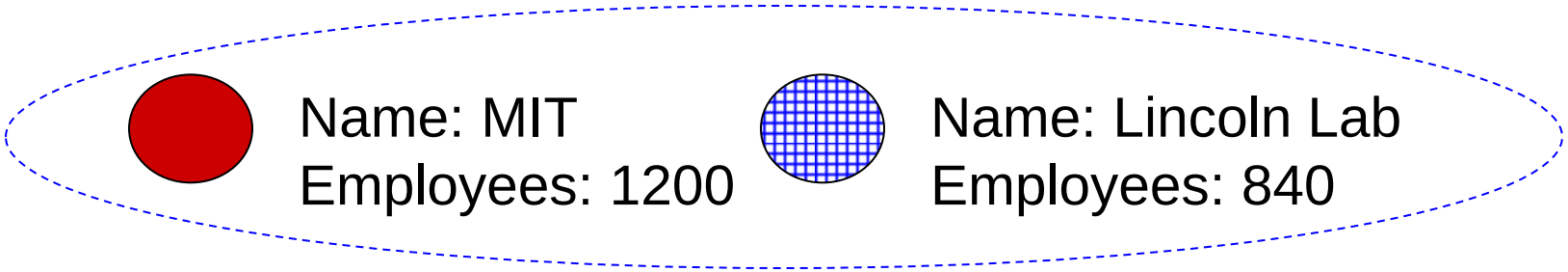

- Unambiguous universal identifiers rare (or rarely used)
- Examples:

  Massaschusetts Institute of Technology

  Mass Inst of Tech

  MIT, M.I.T., M I T
- In practice a frequent problem for mailing lists
  - "Record Linkage" research

# b. Entity aggregation




- What should be included as part of an entity ?
- Example:

  "Lincoln Lab" is "Federally Funded R&D Center of MIT"
- Is Lincoln Lab included in answer to questions, such as:

  How many employees does MIT have ?

  What was MIT's total budget last year ?

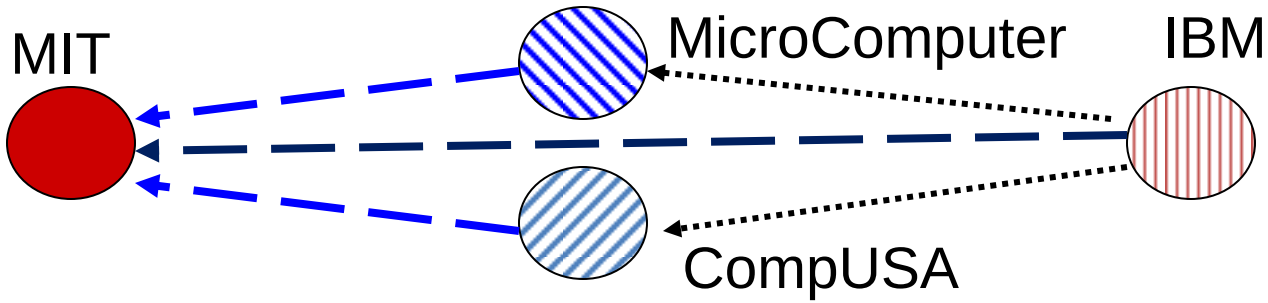
  How much have we sold to MIT ?
- The different circumstances are called "contexts"

# Example: What is "IBM" ?

What is the relationship among these entities
(and the changes over time – "temporal context"):

- International Business Machines Corporation
- IBM
- IBM Global Services
- IBM Global Network (1999-)
- IBM de Colombia, S.A (90%)
- Lotus Development Corporation (100%)
- Software Artistry, Inc. (1998+, 2000-)
- Dominion Semiconductor Company (50/50 jv)
- MiCRUS (majority jv)
- Computing-Tabulating-Recording Co.

# c. Transparency of inter-entity relationships




- Relationships might be direct or indirect
- Understand what circumstances (i.e., contexts)
  should they be collapsed ?
- This can be multi-leveled, especially in
  - financial transactions
  - supply chain management

# Part 3: Connection between Big Data & Data Quality

- **Remark for many Executives:**

  "I now have more and more information, that I know less and less about …"

- **Big Data provides:**
  - Even more data, including personal data
  - From even more diverse sources
- **To get true and effective value from Big Data**
  - It must be high quality Big Data
    - You need to know the quality of the data
    - You need to know the origin (provenance) of the data

# Ending: (Slight) Connection of Data Quality and France

- **In 1805, the Austrian and Russian Emperors agreed to join forces against Napoleon.**
  - The Russians promised that their forces would be in the field in Bavaria by **Oct. 20**.
  - The Austrian staff planned its campaign based on that date in the **Gregorian calendar**.
  - Russia, however, still used the ancient **Julian calendar**, which lagged 10 days behind.
- **That allowed Napoleon to surround Austrian General Mack's army at Ulm and force its surrender on Oct. 21, well before the Russian forces could reach him**.

- **How might history have changed if the Austrian and Russian Emperors had gotten their calendars right?**

Source: David Chandler, *The Campaigns of Napoleon*, New York: MacMillan 1966, pg. 390.

**Thank you for your attention.**

**Questions?**