BIG DATA

Big Data: A definition

- Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools.
- The challenges include:
  - capture
  - curation
  - Storage
  - Search
  - Sharing
  - analysis
  - visualization.

Big Data Trend

The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions. (Wikipedia)

Big Data

- Definition
  - Big data is defined by some as the realization of greater business intelligence by storing, processing, and analyzing data that was previously ignored due to the limitations of traditional data management technologies.
  - And by others as processing of a lot of raw data from many different sources in a way that would make it intelligible to users.

Source: *Harness the Power of Big Data: The IBM Big Data Platform*

Lots of data

- 2.5 quintillion bytes of data are generated every day!
  - A quintillion is 10^{18}
- Data come from many quarters.
  - Social media sites
  - Sensors
  - Digital photos
  - Business transactions
  - Location-based data

Source: *IBM* http://www-01.ibm.com/software/data/bigdata/

The four dimensions of Big Data

- **Volume**: Large volumes of data
- **Velocity**: Quickly moving data
- **Variety**: structured, unstructured, images, etc.
- **Veracity**: Trust and integrity is a challenge and a must and is important for big data just as for traditional relational DBs

Source: *IBM* http://www-01.ibm.com/software/data/bigdata/

The four dimensions of use

- Aspects of the way in which users want to interact with their data…
  - **Totality**: Users have an increased desire to process and analyze all available data
  - **Exploration**: Users apply analytic approaches where the schema is defined in response to the nature of the query
  - **Frequency**: Users have a desire to increase the rate of analysis in order to generate more accurate and timely business intelligence
  - **Dependency**: Users' need to balance investment in existing technologies and skills with the adoption of new techniques

Source: *IBM* http://www-01.ibm.com/software/data/bigdata/

So, in a nutshell

- **Big Data is about better analytics!**
  - Big data analytics is the process of examining large data sets containing a variety of data types to uncover:
    - hidden patterns,
    - unknown correlations,
    - market trends,
    - customer preferences and
    - other useful business information.
  - The analytical findings can lead to
    - more effective marketing,
    - new revenue opportunities,
    - better customer service,
    - improved operational efficiency,
    - competitive advantages over rival organizations and
    - other business benefits.

Type of Data

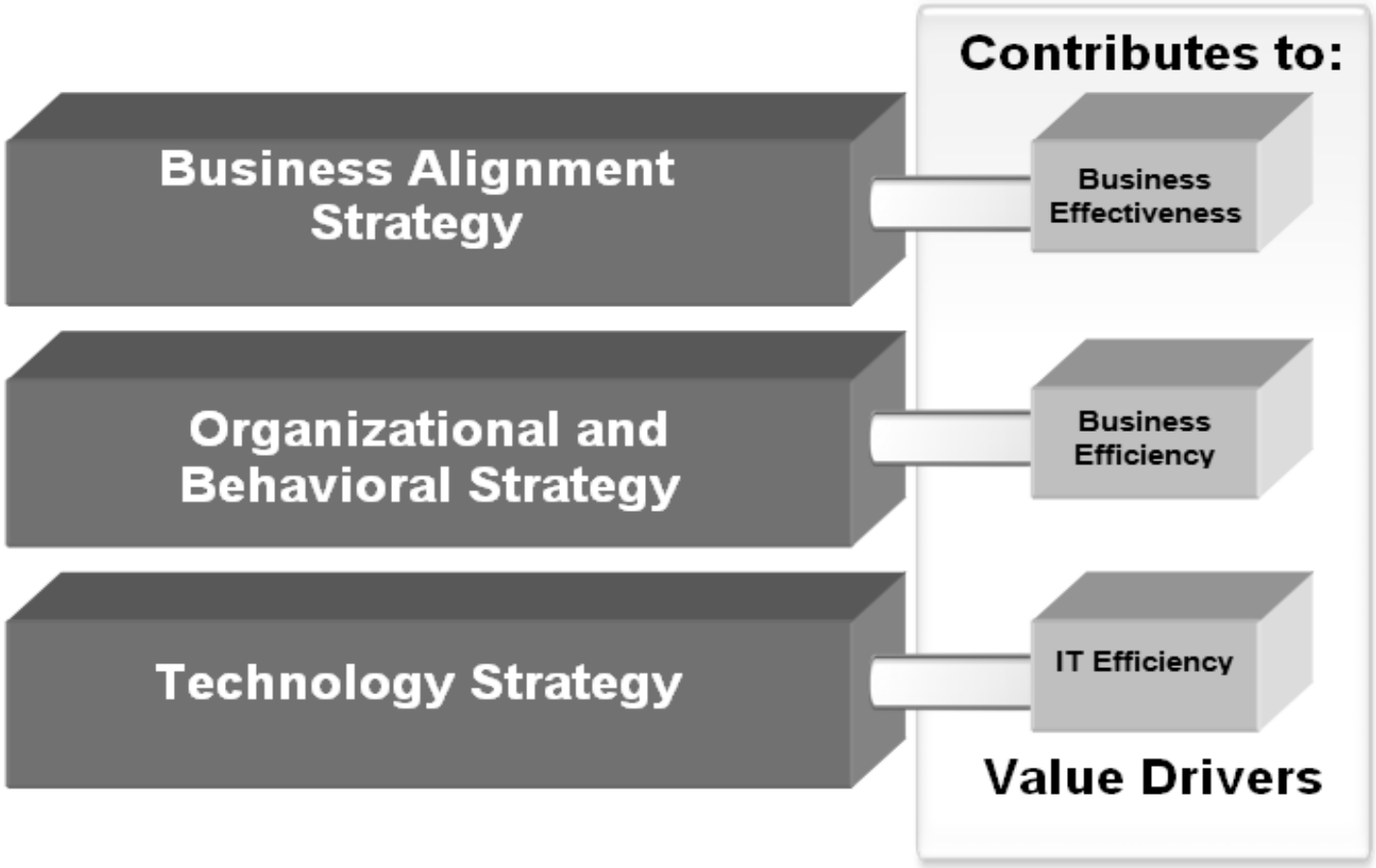
- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once

What to do with these data?

- Aggregation and Statistics
  - Data warehouse and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

larger the sampler is , more meaningful it is

Why Big Data and BI



- Business value outcomes tied to the business strategy resulting from business decisions
- Higher productivity, faster time to complete tasks
- Lower total cost of ownership and greater efficiencies in IT

Source: [Business Intelligence Strategy: A Framework for Achieving BI Excellence]

Source: [Business Intelligence Strategy: A Framework for Achieving BI Excellence]

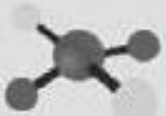
Big Data Conundrum

- Problems:
  - Although there is a massive spike available data, the percentage of the data that an enterprise can understand is on the decline
  - The data that the enterprise is trying to understand is saturated with both useful signals and lots of noise

Source: *IBM* http://www-01.ibm.com/software/data/bigdata/

The Big Data platform Manifesto

imperatives and underlying technologies

1	Discover, explore, and navigate Big Data sources		Federated Discovery, Search, and Navigation
2	Extreme performance–run analytics closer to data		Massively Parallel Processing Analytic appliances
3	Manage and analyze unstructured data		Hadoop File System/MapReduce Text Analytics
4	Analyze data in motion		Stream Computing
5	Rich library of analytical functions and tools		In-Database Analytics Libraries Big Data Visualization
6	Integrate and govern all data sources		Integration, Data Quality, Security, Lifecycle Management, MDM, etc

IBM's Big Data Platform

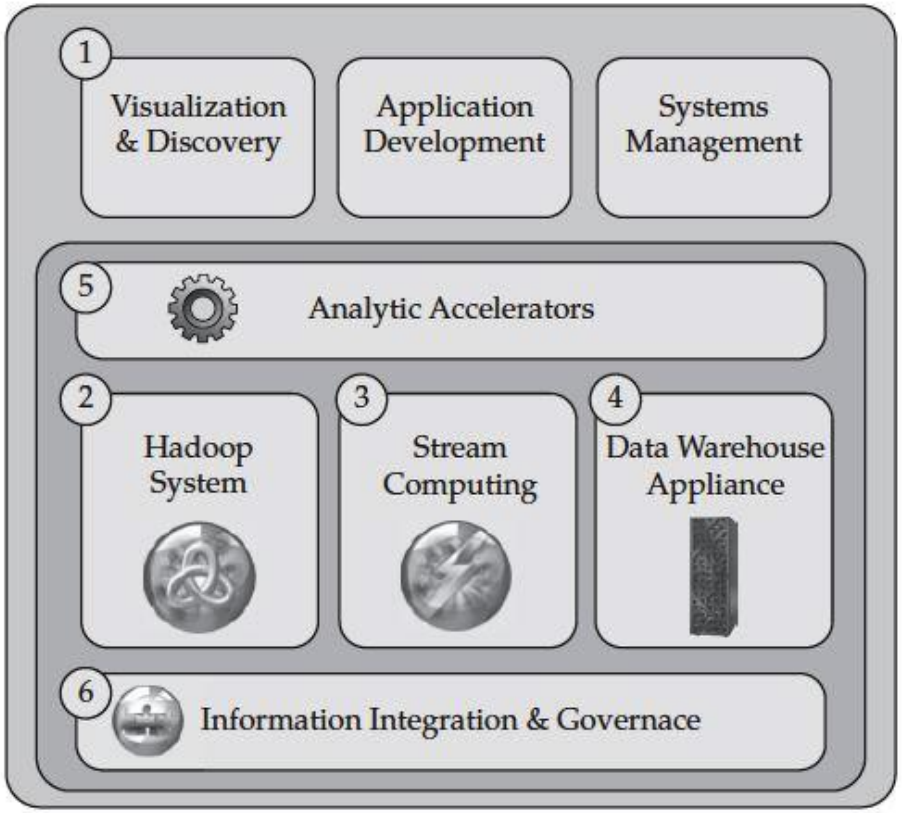


Figure 3-3 *The IBM Big Data platform*

Some concepts

- NoSQL (Not Only SQL): Databases that "move beyond" relational data models (i.e., no tables, limited or no use of SQL)
  - Focus on retrieval of data and appending new data (not necessarily tables)
  - Focus on key-value data stores that can be used to locate data objects
  - Focus on supporting storage of large quantities of unstructured data
  - SQL is not used for storage or retrieval of data
  - No ACID (atomicity, consistency, isolation, durability)

NoSQL

- NoSQL focuses on a schema-less architecture (i.e., the data structure is not predefined)
- In contrast, traditional relation DBs require the schema to be defined before the database is built and populated.
  - Data are structured
  - Limited in scope
  - Designed around ACID principles.

Hadoop

- The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.
- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop is a distributed file system and data processing engine that is designed to handle extremely high volumes of data in any structure.

Hadoop

- Hadoop has two components:
  - The Hadoop distributed file system (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between
  - The MapReduce programing paradigm for managing applications on multiple distributed servers
    - The focus is on supporting
      - redundancy,
      - distributed architectures,
      - and parallel processing

Some Hadoop Related Names to Know

- **Apache Avro**: designed for communication between Hadoop nodes through data serialization
- **Cassandra and Hbase**: a non-relational database designed for use with Hadoop
- **Hive**: a query language similar to SQL (HiveQL) but compatible with Hadoop
- **Mahout**: an AI tool designed for machine learning; that is, to assist with filtering data for analysis and exploration
- **Pig Latin**: A data-flow language and execution framework for parallel computing
- **ZooKeeper**: Keeps all the parts coordinated and working together

What to do with the data

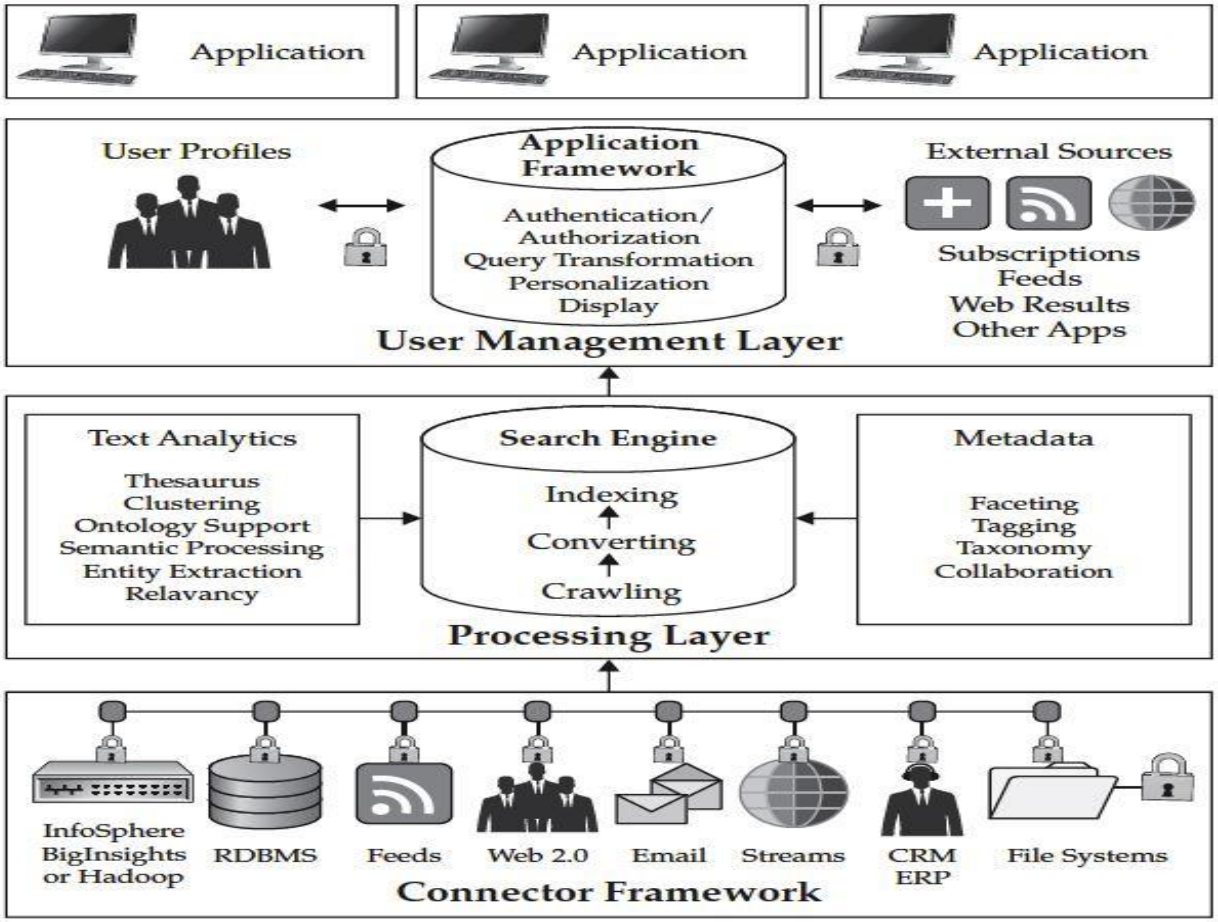
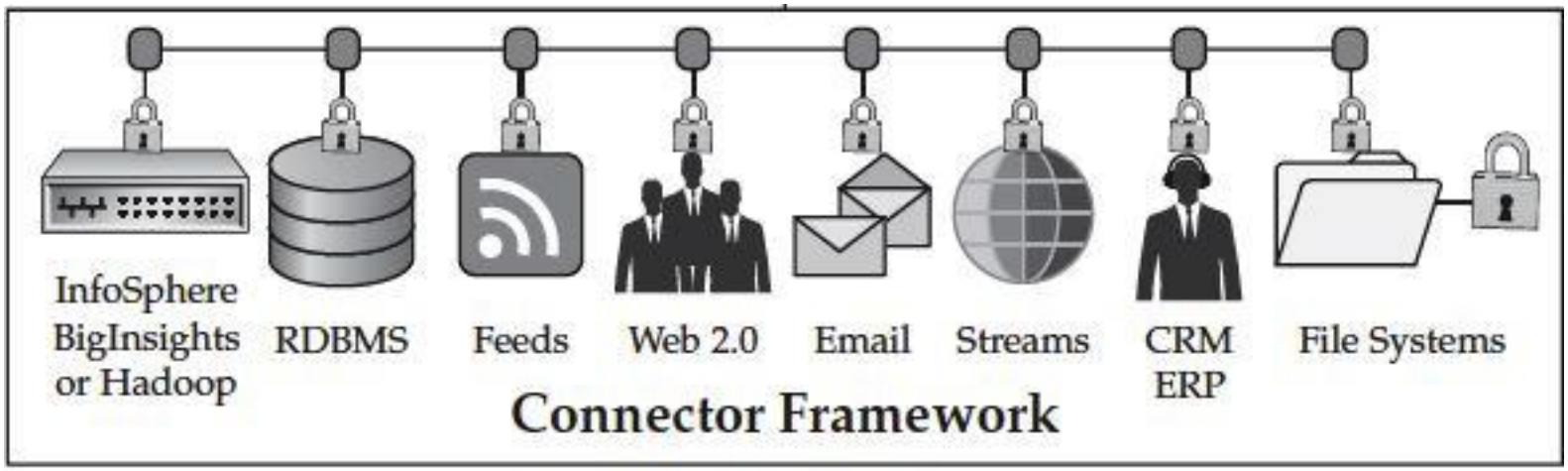


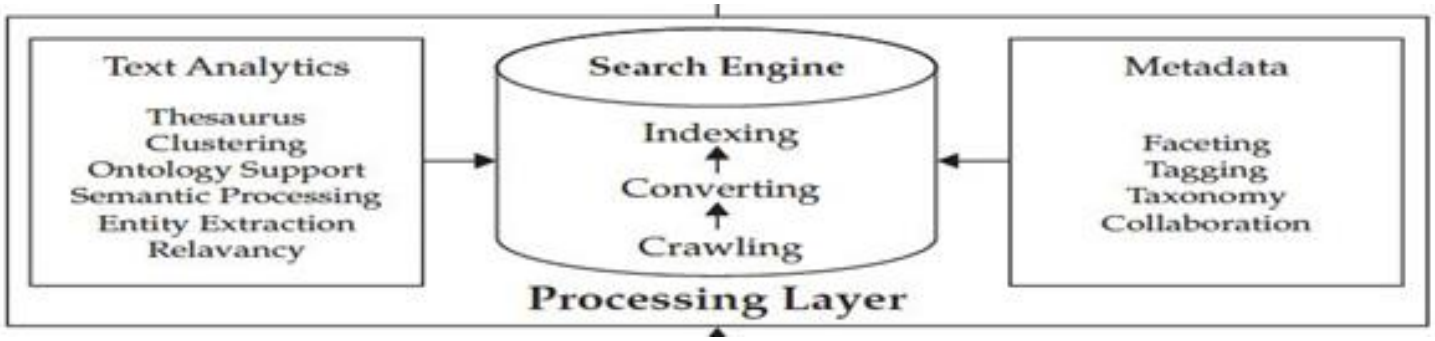
Figure 7-1 *Data Explorer architecture*

Connector Framework

- Supports access to data by creating indexes that can be used for access to the data in its native repository (i.e., it does not manage the data, it keeps track of where it is located)



Processing Layer

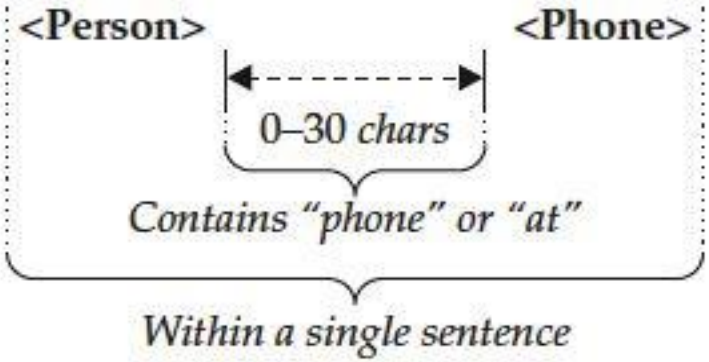


- Two primary functions:
  - Indexes content: data are crawled, parsed, and analyzed with the result that contents are indexed and located
- Processes queries
  - Manages access to various servers hosting the indexed and searchable content

Annotated Query Language

- AQL is an SQL-like declarative language for performing text analysis and extraction

```
create view PersonPhone as select P.name as person, N.number as phone
from Person P, Phone PN, Sentence S where Follows(P.name. PN.number, 0, 30)
    and Contains(S.sentence, P.name) and Contains(S.sentence, PN.number)
    and ContainsRegex(/\b(phonelat)\b/, SpanBetween(P.name, PN.number));
```



BigInsights Text Analytics Workflow - String-based file: \4Q2006.txt - Eclipse

File Edit Navigate Search Project Run Window Help

BigInsigh | Extractio | Package

Step 1 : Select Documents

Step 2 : Label Examples and Clues

a. Label Snippets of Interest

In the open documents, select text snippets that can serve as examples of what you want to extract. To label an example, select the text, right-click the text, and then click either 'Add Example with New Label' or 'Label Example As'.

See Example

Tip: Labeled examples are shown in the extraction plan.

b. Label extraction clues

From within the example snippets or nearby text, define other labels to use as clues for extraction.

See Example

Step 3 : Develop the Extractor

Step 4 : Test the Extractor

Step 5 : Profile the Extractor

Step 6 : Export the Extractor

InfoSphere BigInsights Task Launcher | main.aql | \4Q2006.txt

or 7 percent compared with the fourth-quarter 2005, excluding pension curtailment charge;

operations, an increase of 12 percent as reported, compared with diluted earnings of $2.01

in the fourth quarter of 2005, an increase of 8 percent. Income from continuing operations

terrific quarter and a good year with record cash performance, profit and EPS, as well as re

increase of 6 percent as reported (5 percent, adjusting for currency) from the 2005 period.

adjusting for currency) compared with the fourth quarter of 2005. Revenues from IBM's midd

a wide variety of business processes using open standards to interconnect applications, da

7 percent (4 percent, adjusting for [illegible], and segment revenues from G

quarter, up 3 percent (flat, adjustin[illegible]nues from System z server pro

fourth quarter to $620 million.

with 44.1 percent in the 2005 period.

-year period. SG&A expense increased 7 percent to $5.6 billion. RD&E expense increased 9 per

cent in the fourth quarter of 2005. The decrease in the tax rate was caused by the favorable

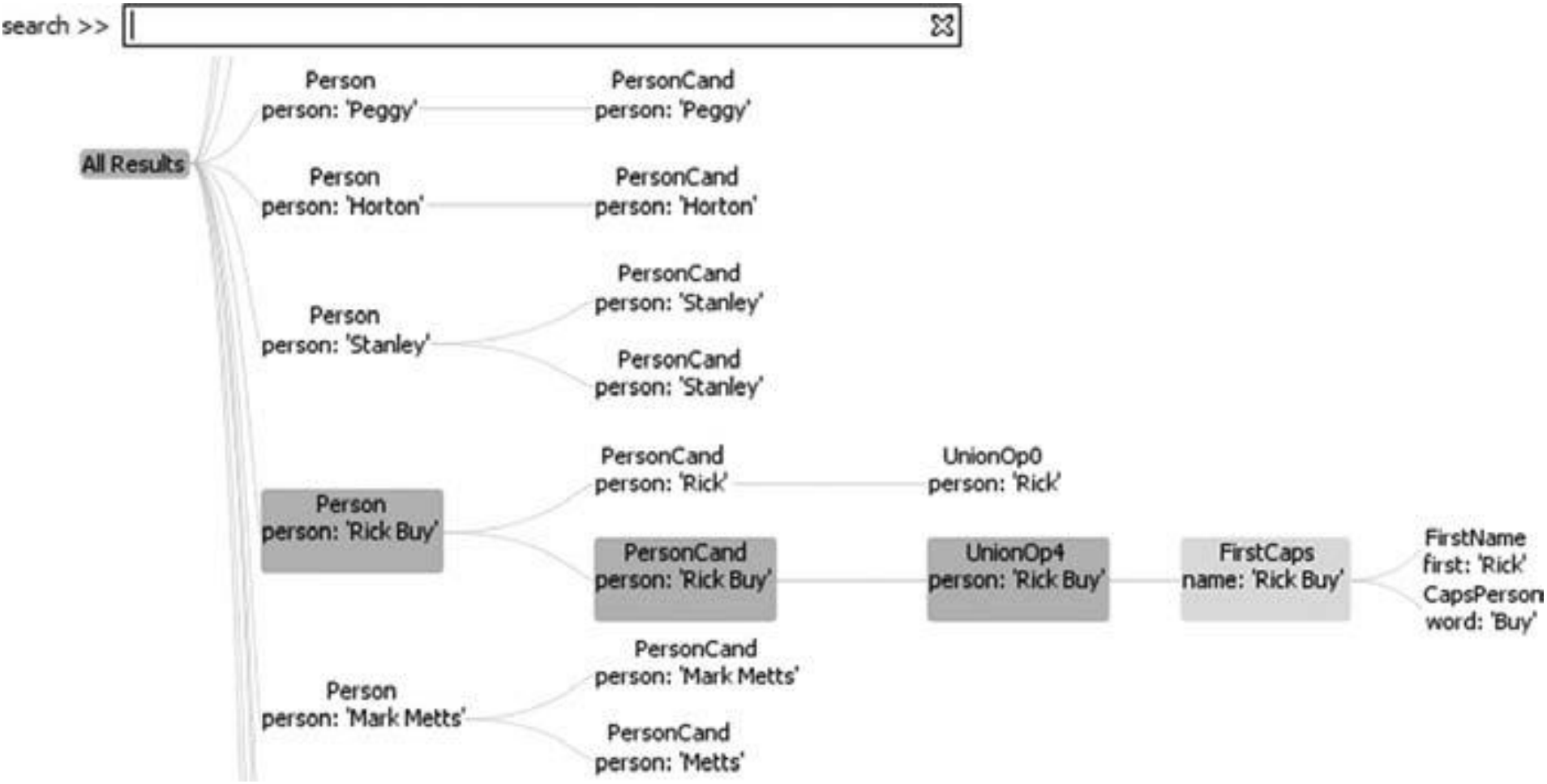
diluted share, which included a gain from discontinued operations related to country tax se

Add Example with New Label

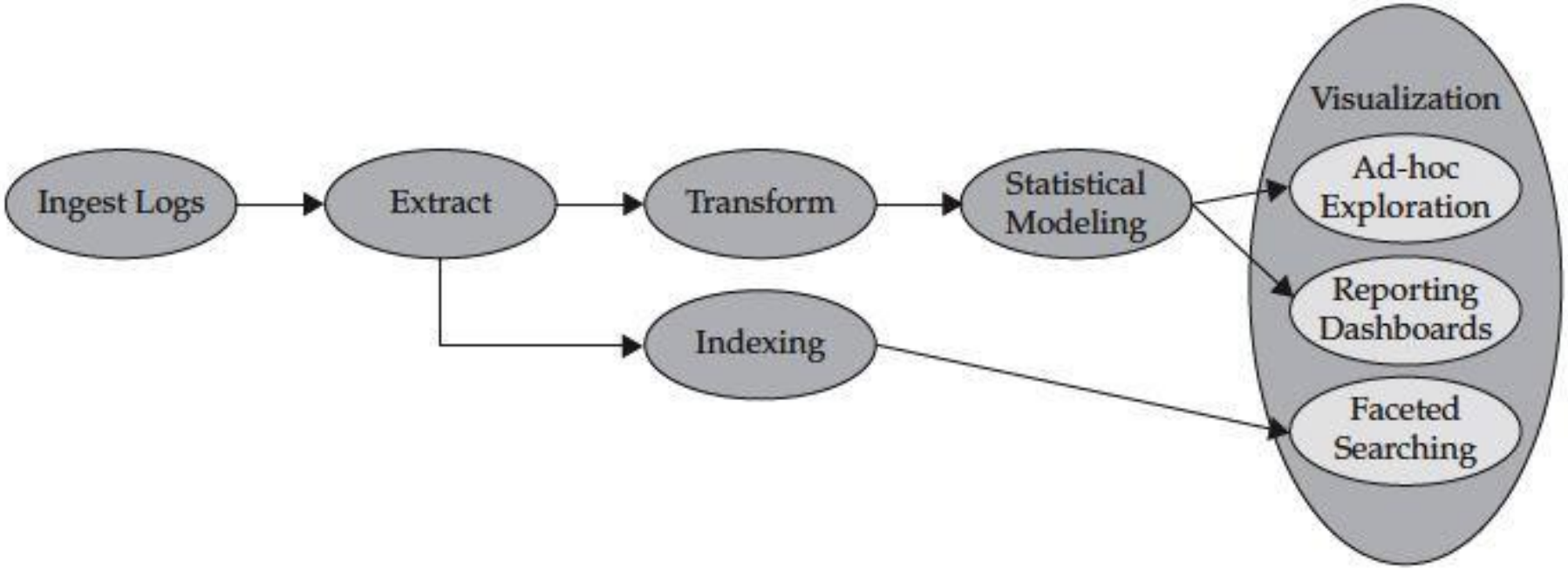
Add to Snippets...

Read-Only | Insert | 29 : 155

The background viewer



Machine data analysis



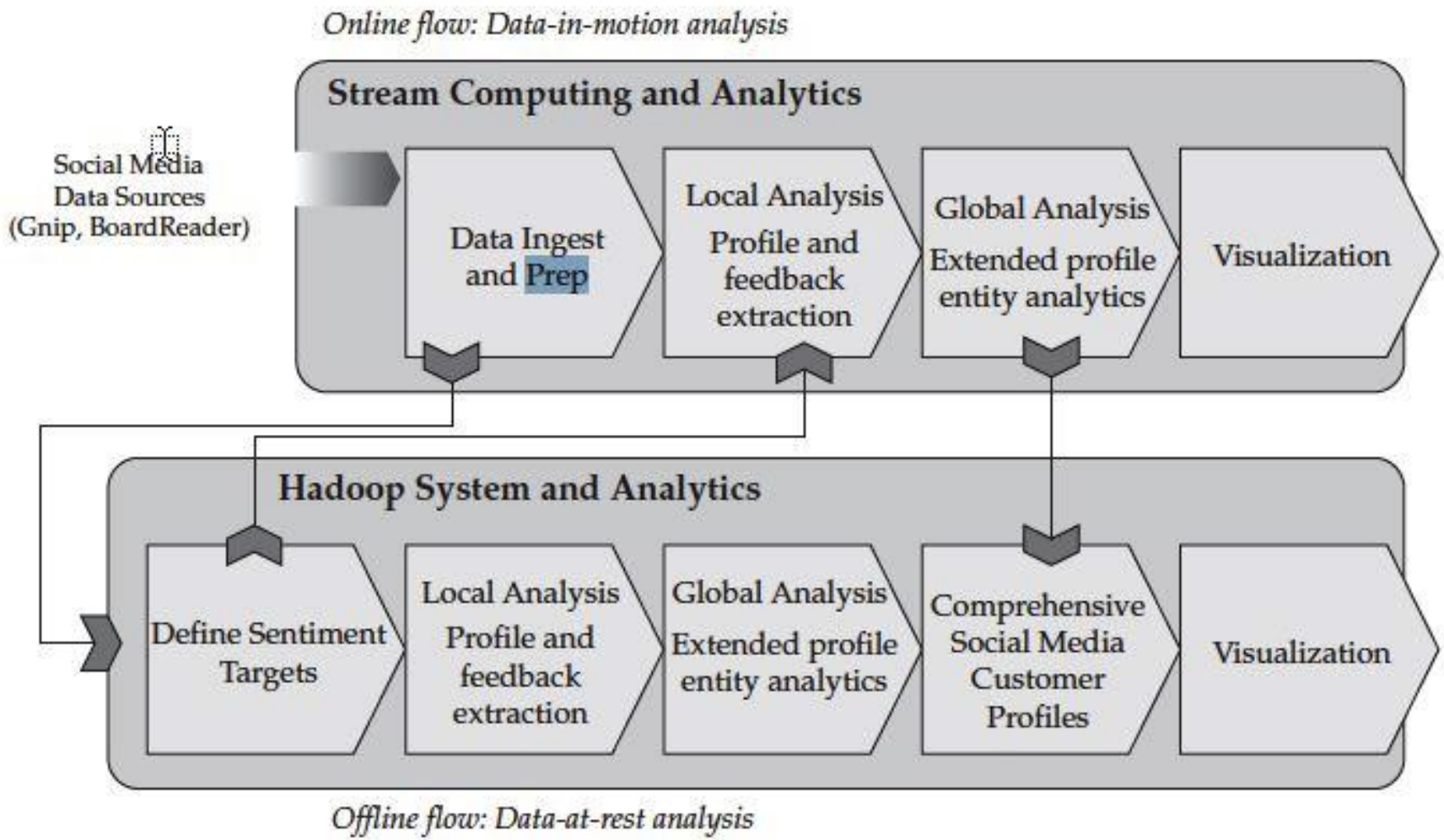


Figure 9-3 *The lifecycle of social data analysis*

OPTIMUM INVESTMENTS

Logged in as **Frank Gelato** | Help

Followed High Net Wealth Clients

- Isabella Jones
- Thomas Jackson
- Theresa Mayer
- Michael Kleinfelder

Tracked Products

- 529 Plan
- 401K
- Money Market IRA
- Fixed Income & Bonds

Financial Blogs

CNBC: Warren Buffett: 'Disruptive' Debt Limit Debates Are 'Waste of...
Everything Warren Buffet

CNBC: CNBC Transcript: Warren Buffett on Russian Roulette, Tax...
Everything Warren Buffet

BLOOMBERG: Munger Treats 'Hard-Core Addicts' as Wesco Stock Exits...
Everything Warren Buffet

Investment News

Texas Gains New Billion Dollar Bank
Business Wire - 35 minutes ago

Ally Financial Reports Preliminary Second Quarter 2011 Financial Results
PR Newswire - 54 minutes ago

Pinnacle West Reports Second-Quarter Results
TheStreet.com - 55 minutes ago

Action Needed

Sentiment	Customer	Type	Format	Time	Product
Negative	Isabella Jones	Support	Tweet	05.29.12 10:30 am EST	Mutual funds
Neutral	Thomas Jackson	Support	Support ticket - email	05.29.12 10:25 am EST	401K
Positive	Isabella Jones	Sales	Tweet	05.29.12 10:05 am EST	529
Negative	Theresa Mayer	Satisfaction	Blog	05.29.12 09:30 am EST	401K

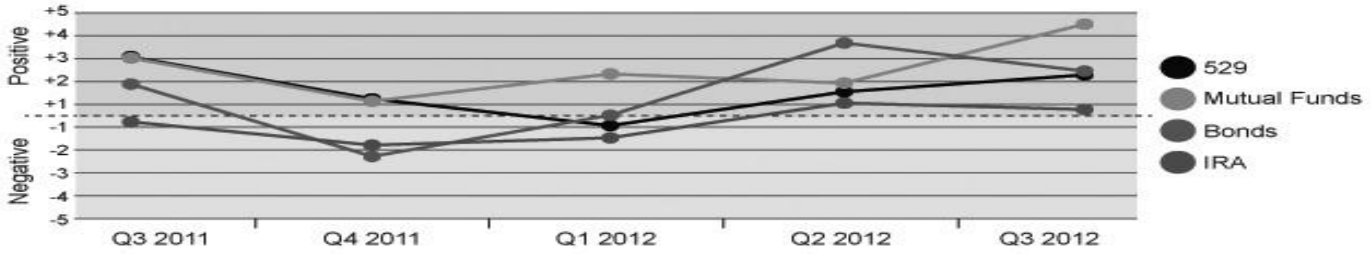
Activity Feed

Isabella Jones - @IzzyJones
Wow is the market really this bad, my monthly mutual fund statement looks terrible! Time to call my financial advisor.
Twitter - Minutes ago

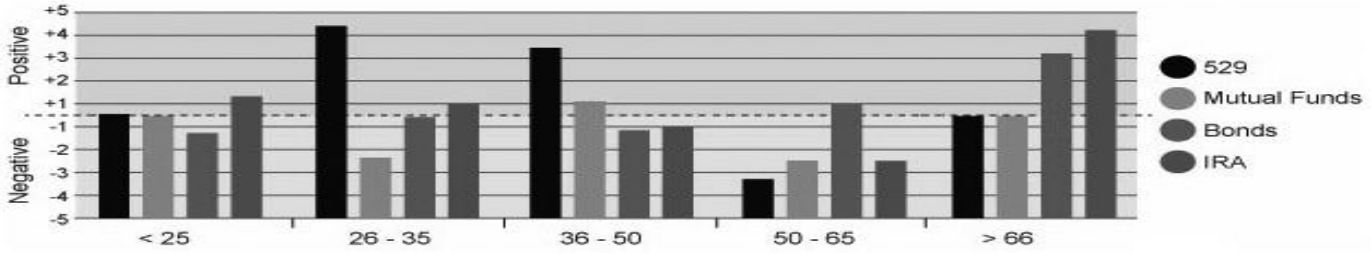
Theresa Mayer - t.mayer@gmail.com
Great advice on retirement planning! Thinking about increasing my 401K but are there other retirement plans I should be looking at given I hope to retire in 10-15 yrs?
http://nextavenue.org/blog/why-women-need-embrace-retirement-planning
Blog - Minutes ago

Thomas Jackson
How can I change my 401K contribution using your online system?
Remedy - Minutes ago

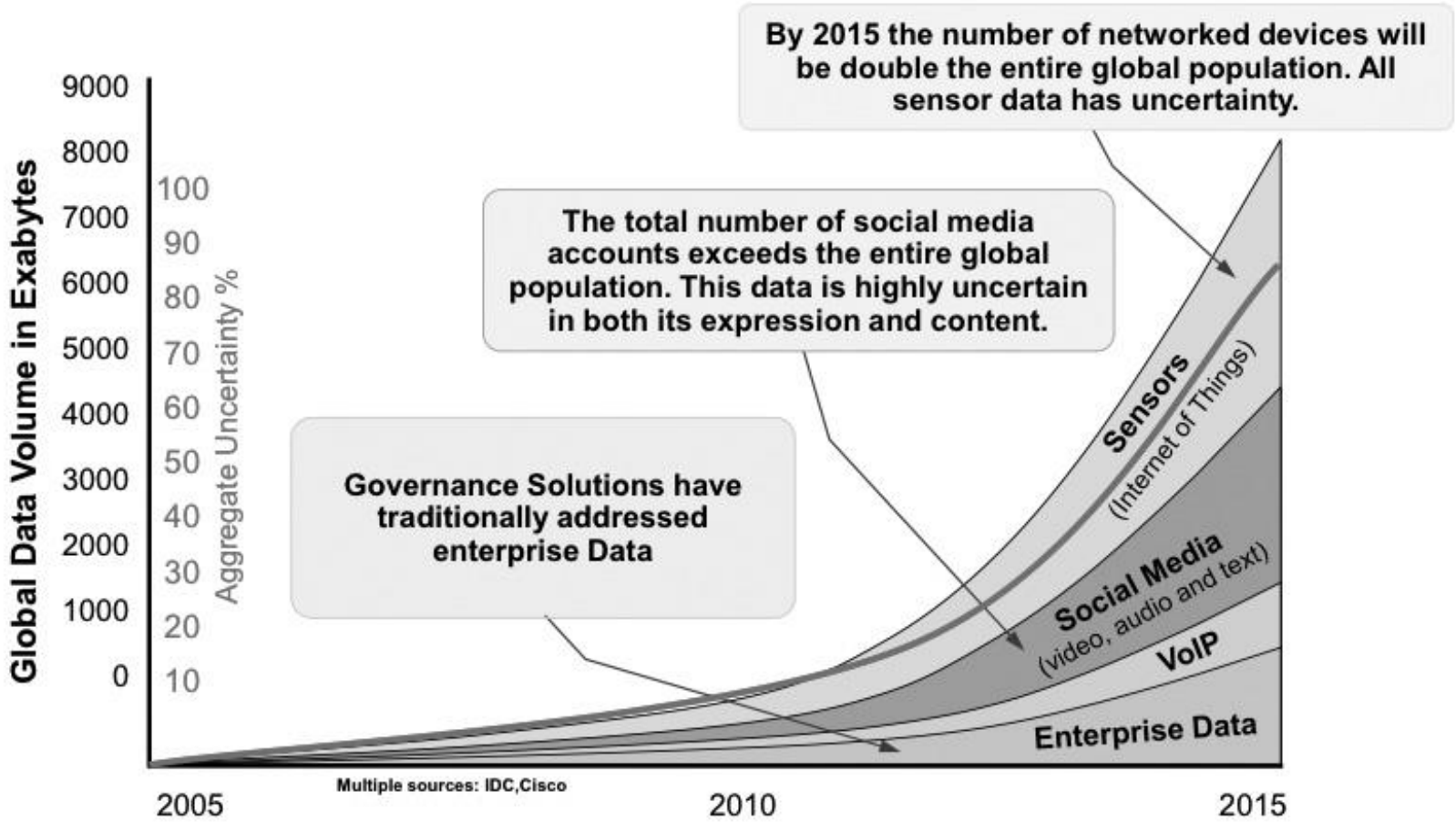
Sentiment



Sentiment By Age:



By 2015, 80% of all available data will be uncertain



1 in 3 Leaders make decisions on untrusted information

1 in 2 Leaders don't have the information they need

60% of CEOs have more data than they can use

Utilizing massive data to discover and explain

Is not as easy as you might think…

- Poor and sparse samples, surrogates, bias…
- As number of dimensions increases it becomes increasingly difficult to add in any data point without giving rise to some kind of statistically significant ‘pattern’ or ‘cluster’
- And parametric distributions become unreliable
- It is very difficult to discover **useful** things that are **unknown** by experts

We need to capture the meaning of data, not just the data itself

- aligning heterogeneous definitions in content, schema

ID : Table

ERA	PERIOD	SERIES	ROCK_TYPE	LITH1
Precambrian	Algonkian		igneous rocks - intrusive	dioritic sills
Cenozoic	Quaternary	Pleistocene	igneous extrusive rocks	basalt

Era

Eon

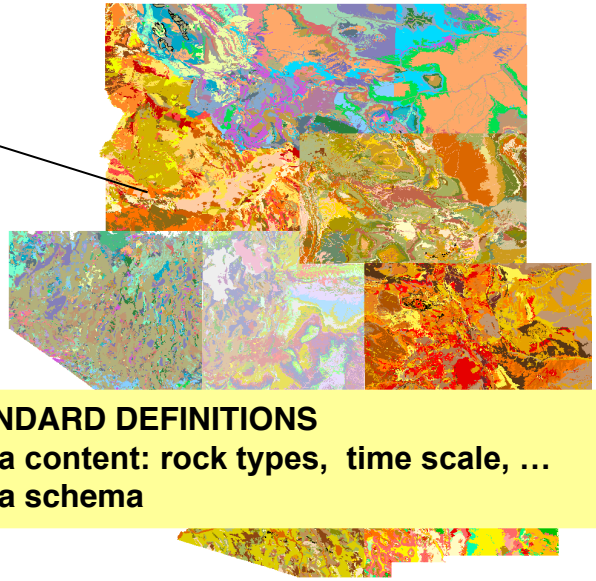
Period

Series

GEOLOGIC AGE	ROCK TYPE
	Volcanic

AZ : Table

PERIOD	ROCK_DESCR
Algonkian	GRANITOID ROCKS
Paleozoic	PALEOZOIC ROCKS
Cretaceous	VOLCANIC ROCKS



STANDARD DEFINITIONS
- **data content: rock types, time scale, …**
- **data schema**

from www.GEONgrid.org

Bid data Example:

Hotel Chain Uses Big Data to Increase Bookings
Bad weather reduces travel, which then reduces overnight lodging. That's not good news if you're in the hotel business.

However, Red Roof Inn turned this trend on its head. The hotel chain recognized that cancelled flights leave travelers in a bind and in need of a place to sleep overnight. The company sourced freely available weather and flight cancellation information, organized by combinations of hotel and airport locations, and built an algorithm which factored weather severity, travel conditions, time of the day and cancellation rates by airport and airline among other variables. With its big data insights, and recognition that travelers will be using mobile devices for this use case, the company used Search, PPC and SoLoMo mobile campaigns to deliver targeted mobile ads to stranded travelers and make it easy for them to book a nearby hotel.

This big data payback is compelling.

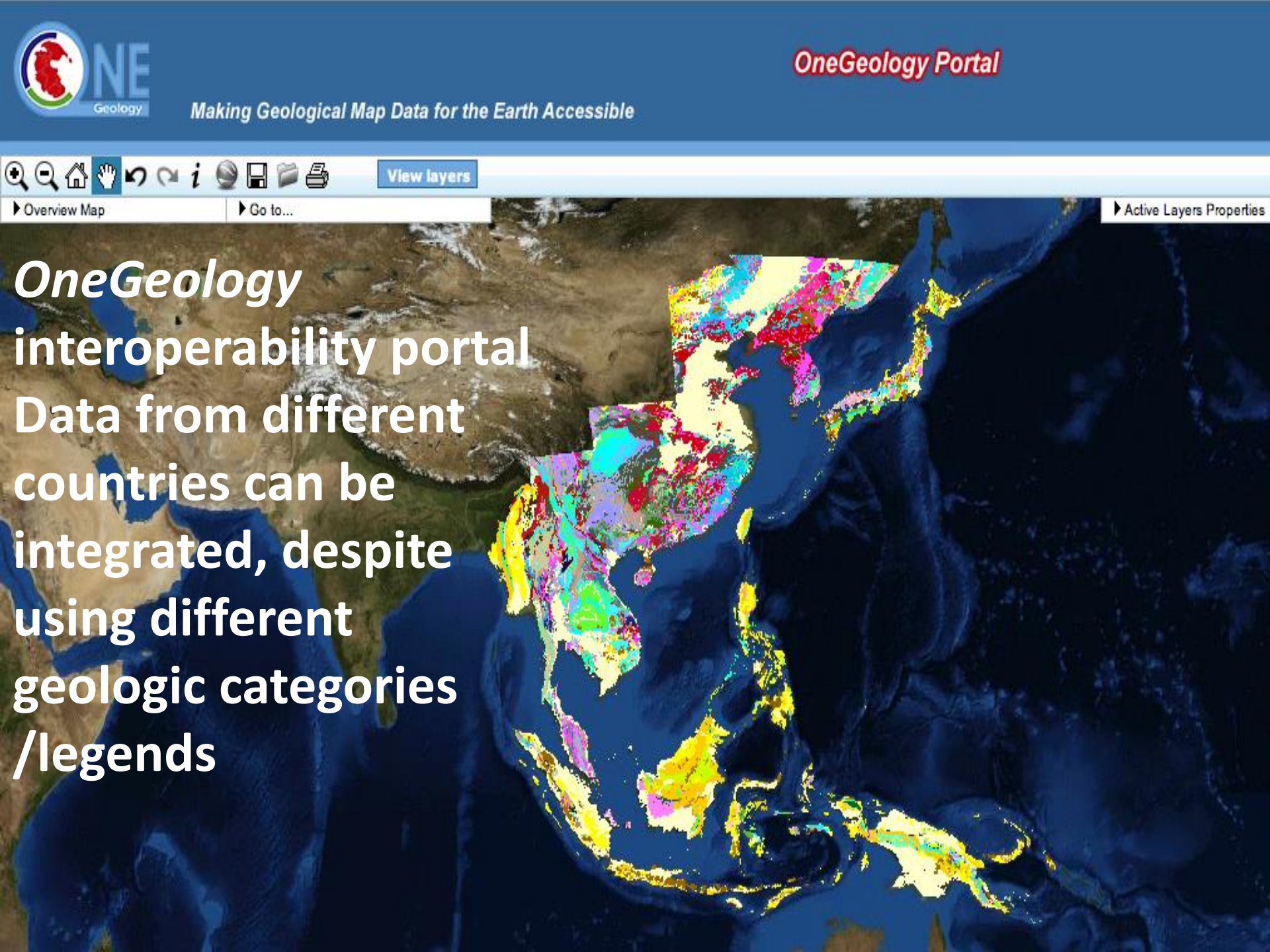
Flight cancellations average 1-3% daily, which translates into 150 to 500 cancelled flights or around 25,000 to 90,000 stranded passengers each day. With its big data and geo-based mobile marketing campaigns Red Roof Inn achieved a 10% business increase from 2013 to 2014.

Pizza Chain Earns More Dough in Bad Weather

Somewhat similar to the previous example, a pizza chain uses a mobile app and mobile marketing techniques to deliver coupons based on bad weather or where power outages leave consumers unable to cook. This mobile and location-based marketing campaign achieves a 20% response rate.

Big Data Examples:

https://public.tableau.com/s/
http://www.nytimes.com/interactive/2009/11/06/business/economy/unemployment-lines.html?_r=0
http://www.informationisbeautiful.net/play/snake-oil-supplements/
http://www.axiis.org/examples/BrowserMarketShare.html
http://live.idashboards.com/hoops/



OneGeology **interoperability portal Data from different countries can be integrated, despite using different geologic categories /legends**