DATA ECOSYSTEMS

FROM VERY LARGE DATA BASES TO BIG DATA INFRASTRUCTURES

Timos Sellis

School of CS & IT

RMIT UNIVERSITY

Big Data – What is it?

Most commonly accepted definition, by Gartner (the 3 Vs)

"***Big data*** *is* ***high-volume***, ***high-velocity*** *and* ***high-variety*** *information assets that demand* ***cost-effective***, ***innovative*** *forms of information processing for* ***enhanced insight*** *and* ***decision making***."

Big Data – some stats

- ***high-volume***, ***high-velocity*** *and* ***high-variety***

Every minute…

(http://www.domo.com/blog/blog/2012/06/08/how-much-data-is-created-every-minute/)

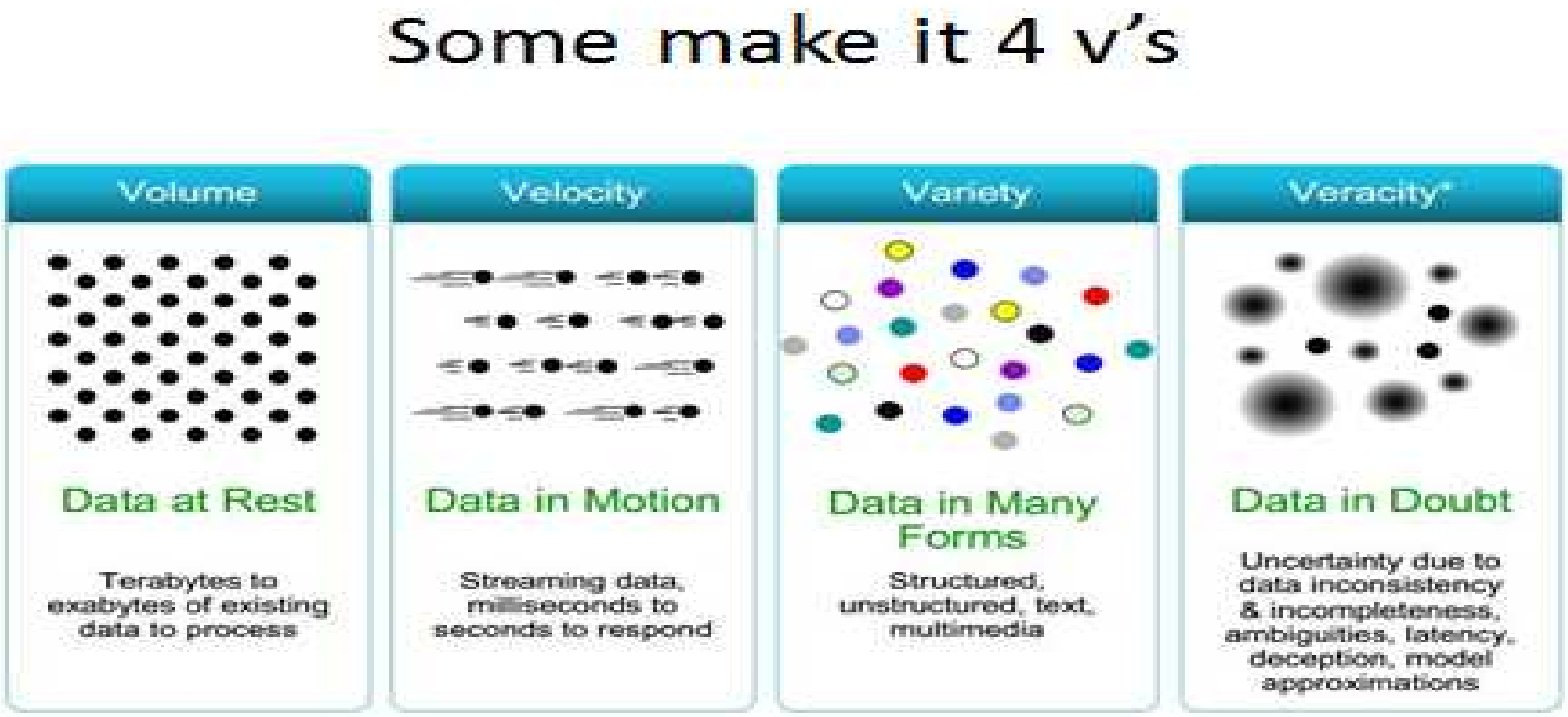
Big Data – Is it a new wave?

- Yes and no
- **Yes**, it is a **different type** of data wave: one needs to put together many sources of information, coming through many different channels, throwing away what is not important, working under time constraints, serving analysts and end users
- **No**, most of these problems have been in the focus of data management research for years
- The main issue is to **put all this together**, using innovative technology, serving users' needs

Big data: 3 V’s

- Three main dimensions- volume, velocity and variety.
- **Volume**: Machine generated data is produced in larger quantities compared to traditional data.
- **Velocity**: The speed of data flowing in.
- **Variety**: Large variety of input data which in turn generates large amount of data as output.

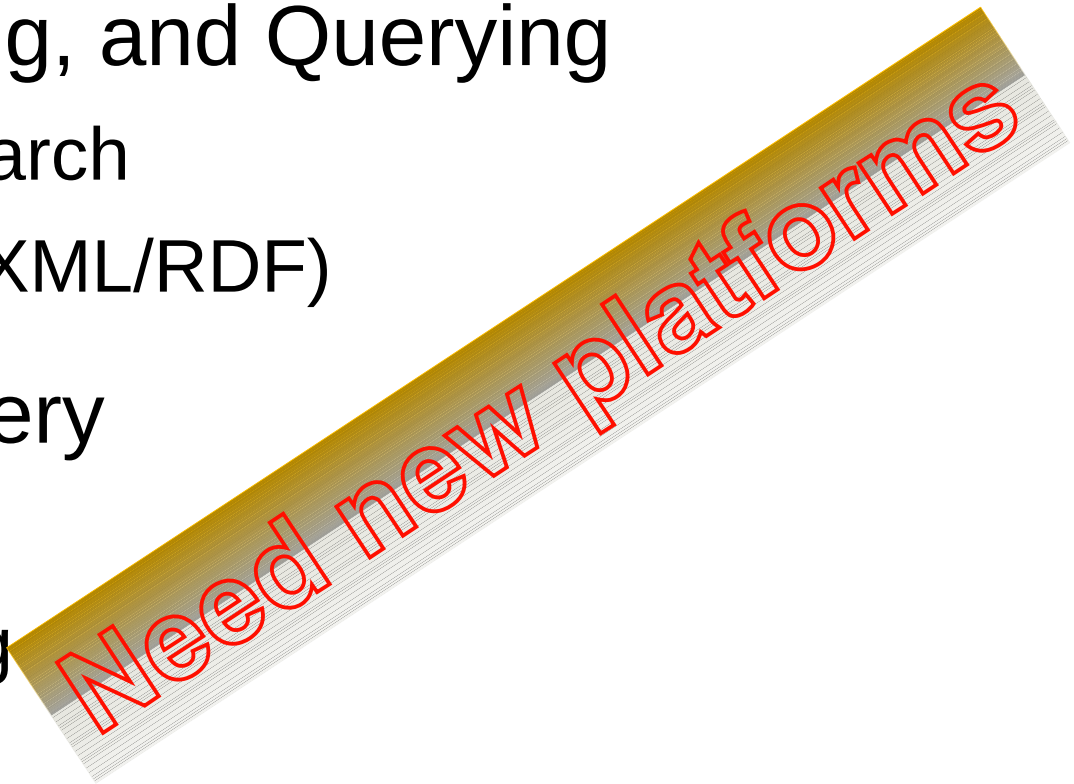
There are actually more Vs



Source: http://www.slideshare.net/kunalkhanna33/big-data-and-hadoop-overvew

What to do with this data?

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

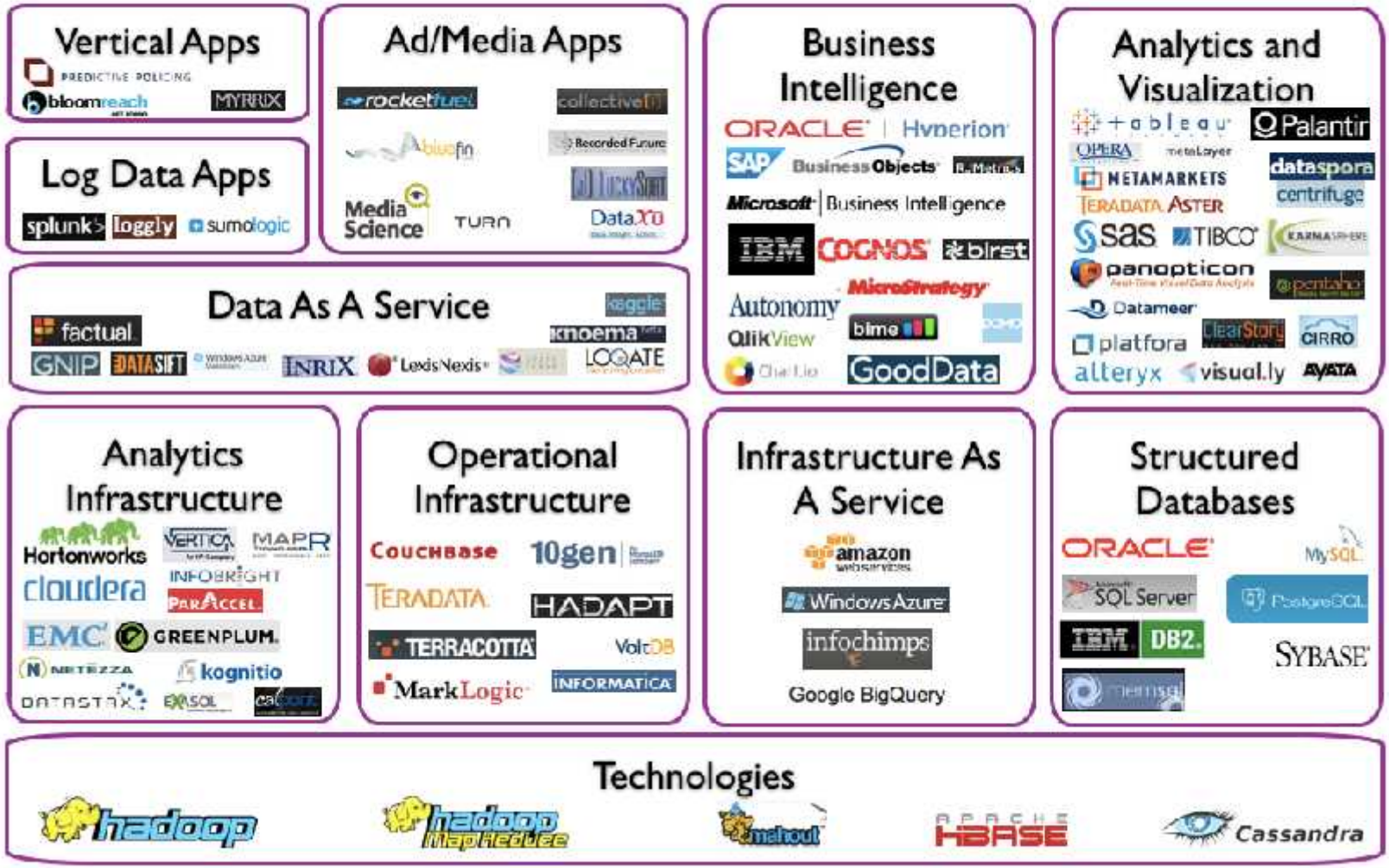


Big Data Platforms – an example

IBM's big data platform

- **Hadoop-based analytics:** Process and analyze any data type across commodity server clusters.
- **Stream Computing:** Drive continuous analysis of massive volumes of streaming data with sub-millisecond response times.
- **Data Warehousing:** Deliver deep operational insight with advanced in-database analytics.

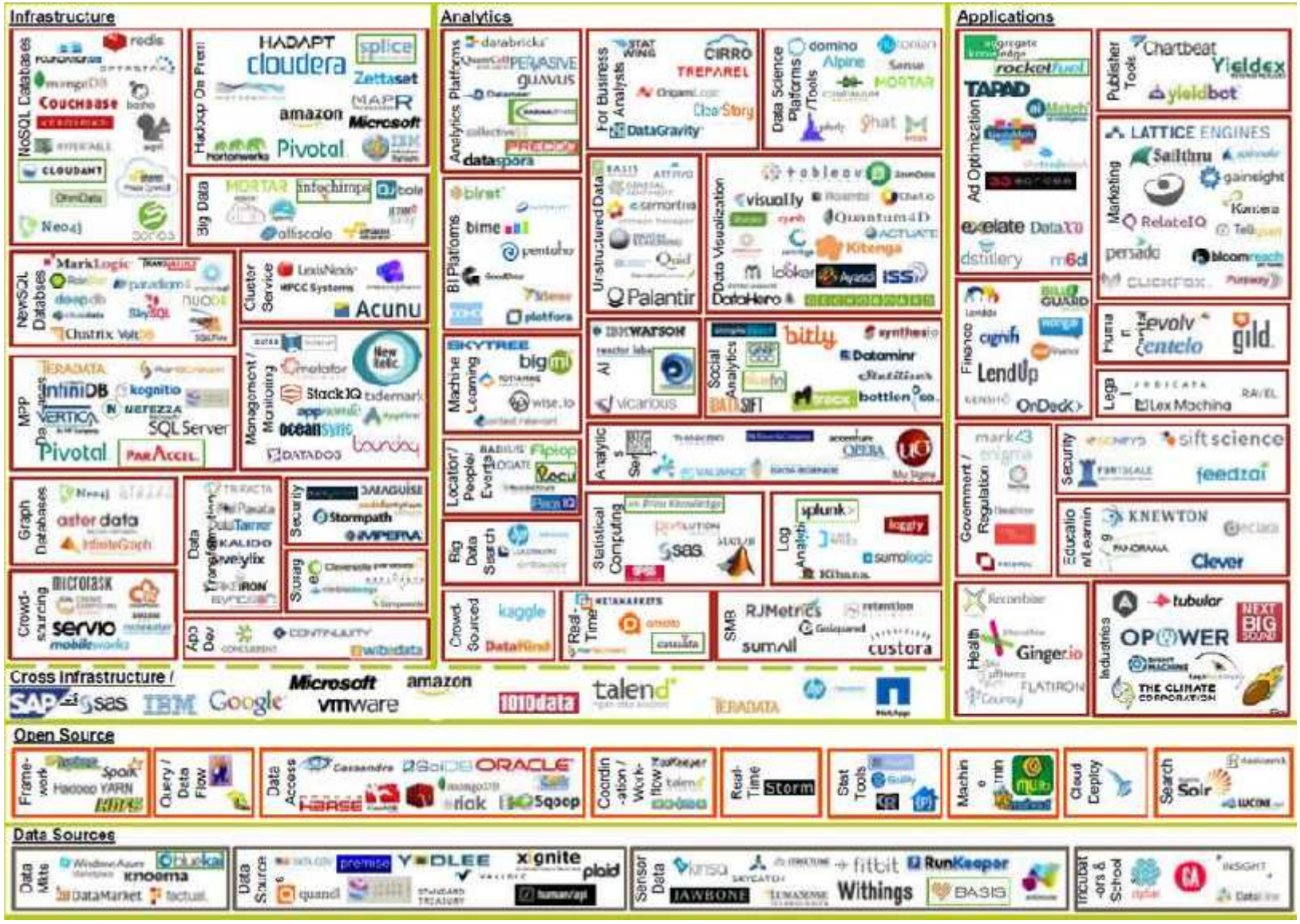
Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

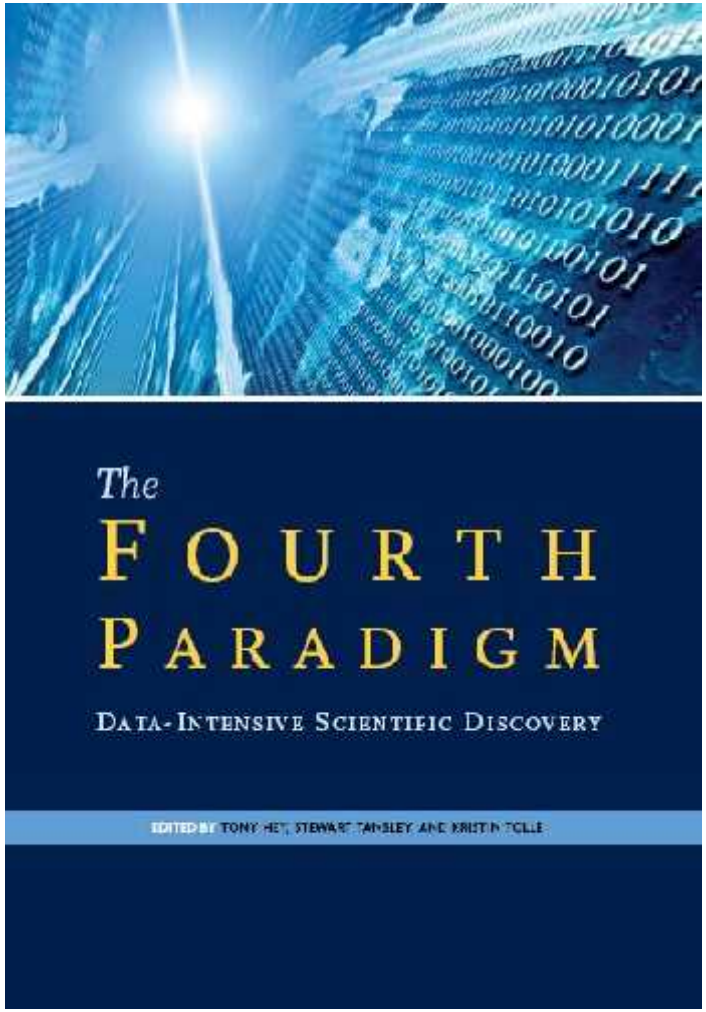
2 years later



http://www.slideshare.net/mjft01/big-data-landscape-matt-turck-may-2014

A paradigm shift - Science

- The 4th Paradigm of Science
  - Data-Intensive Scientific Discovery
- From eScience to dScience



A paradigm shift - Science

Novartis New Drug Research

Developing new drugs

"*Big data was the game changer*," says one of the team leaders, J. Szustakowski, head of Bioinformatics in Biomarker Development at the Novartis Institutes for BioMedical Research (NIBR) in Cambridge, Mass.

To make sense out of this wave of data, scientists are developing sophisticated ways to store, retrieve and analyze it. A new breed of "data scientist" is working to re-invent the traditional drug research team. Instead of biologists, chemists and clinicians working in silos, pharmaceutical companies such as Novartis are assembling collaborative, cross-disciplinary teams. These teams include data scientists, drawing on their expertise in computer science and statistics to sift through information and attempt to extract answers to pressing questions. They collaborate with biologists and clinicians to develop a clear hypothesis and then put it to the test.

http://www.novartis.com/stories/discovery/2013-10-big-data.shtml

A paradigm shift - Business

Danish firm Vestas uses supercomputers and a big data modelling solution to pinpoint the optimal location for its wind turbines to maximize power generation and reduce energy cost.

Vestas Wind Energy Turbine Placement and Maintenance

Incorporates data from global weather systems with data collected from its existing turbines. The wind library holds nearly **3 Petabytes** of data.

Parameters include temperature, barometric pressure, humidity, precipitation, wind direction and velocity from the ground level up to 300 feet, and the company's recorded historical data. **The company expects to analyze even more diverse and bigger weather data sets reaching 20-plus petabytes over the next four years** as Vestas plans to add global deforestation metrics, satellite images, historical metrics, geospatial data and data on phases of the moon and tides.

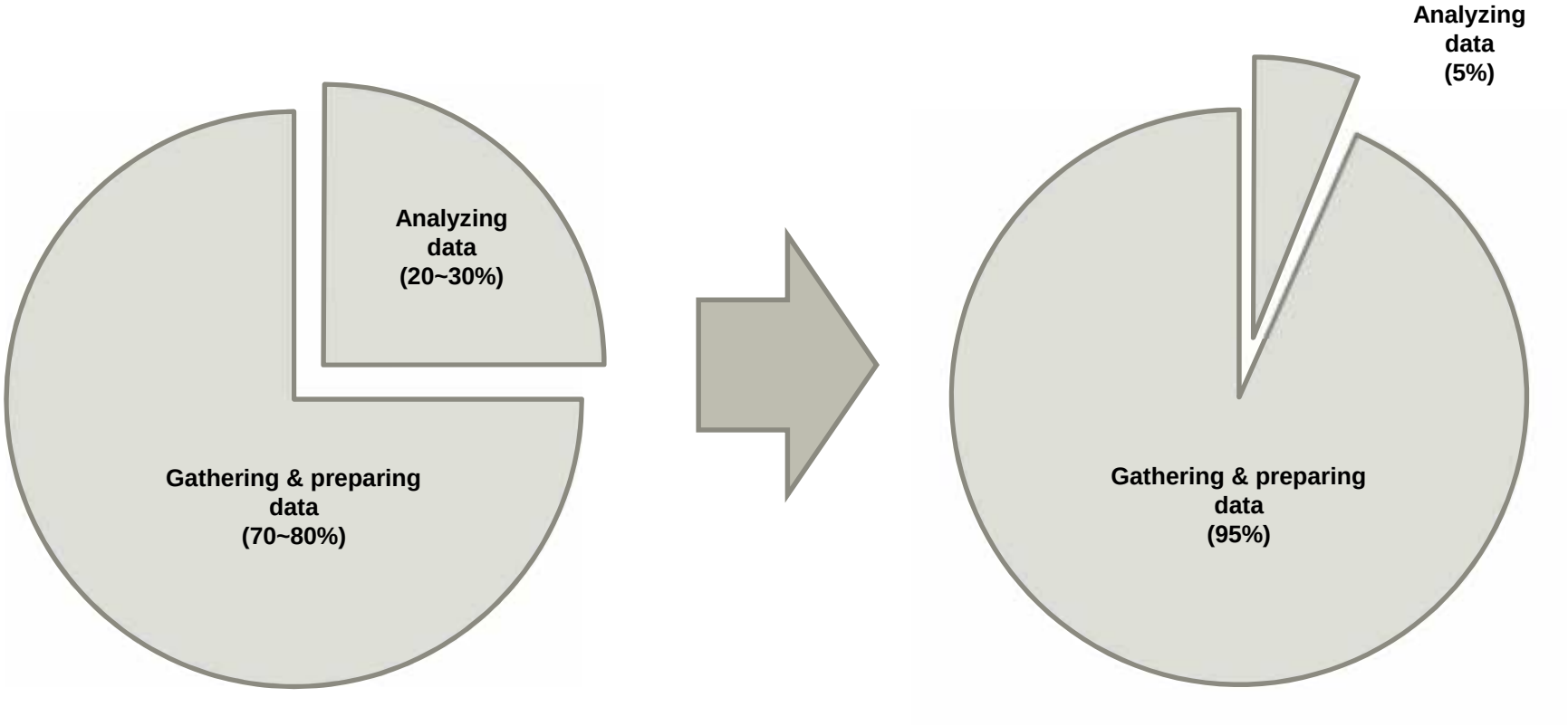
Main Issue - Big Data Analysis

- Complex math operations (machine learning, clustering, trend detection, ….)
- Need for new data structures (eg. support for arrays)
- Lots of intensive computations
  - Matrix multiplication
  - QR decomposition
  - Singular Value Decomposition (SVD) decomposition
  - Linear regression

Main Issue - Exploring Big Data

The time for developing an analysis

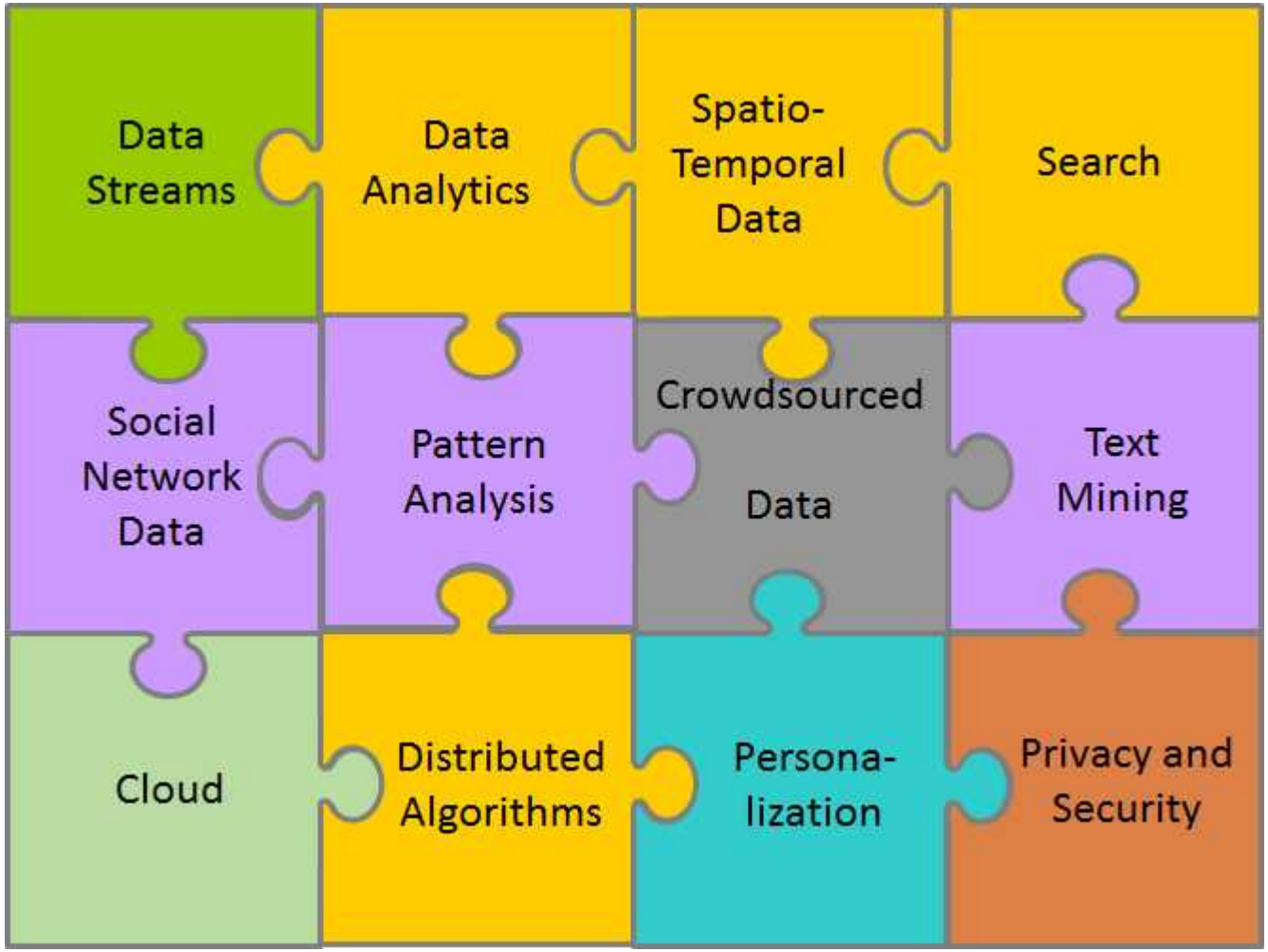
The time for developing an analysis (Initially working with big data)



Unleashing the power of data

Big Data
Infra-structures

Beyond the 3 Vs
Volume
Velocity
Variety



Big Data Infrastructures …….

- The main factor for the **data economy**.
- *The emerging economy in which organizations succeed or fail based in large part on **their ability to leverage data and analytics to improve operational efficiencies, to make better tactical and strategic decisions, and to create innovative products, services and business models to meet & exceed customer expectations**.* [EU]

….. Supporting Data Ecosystems

- Leaving the era of databases and moving to the era of **dataspaces** i.e. a set of loosely interrelated information containers.
- An information container is a resource that holds information and can be referred to via an identifier that is unique to the dataspace.
  - –Examples of such resources include databases, database relations, database tuples, files, records in files, data streams, tuples in data streams, documents, parts of texts, maps, trajectories, etc.

The DataEco view …..Breaking news

BBC has some story on UNICEF's new report on child deprivation,

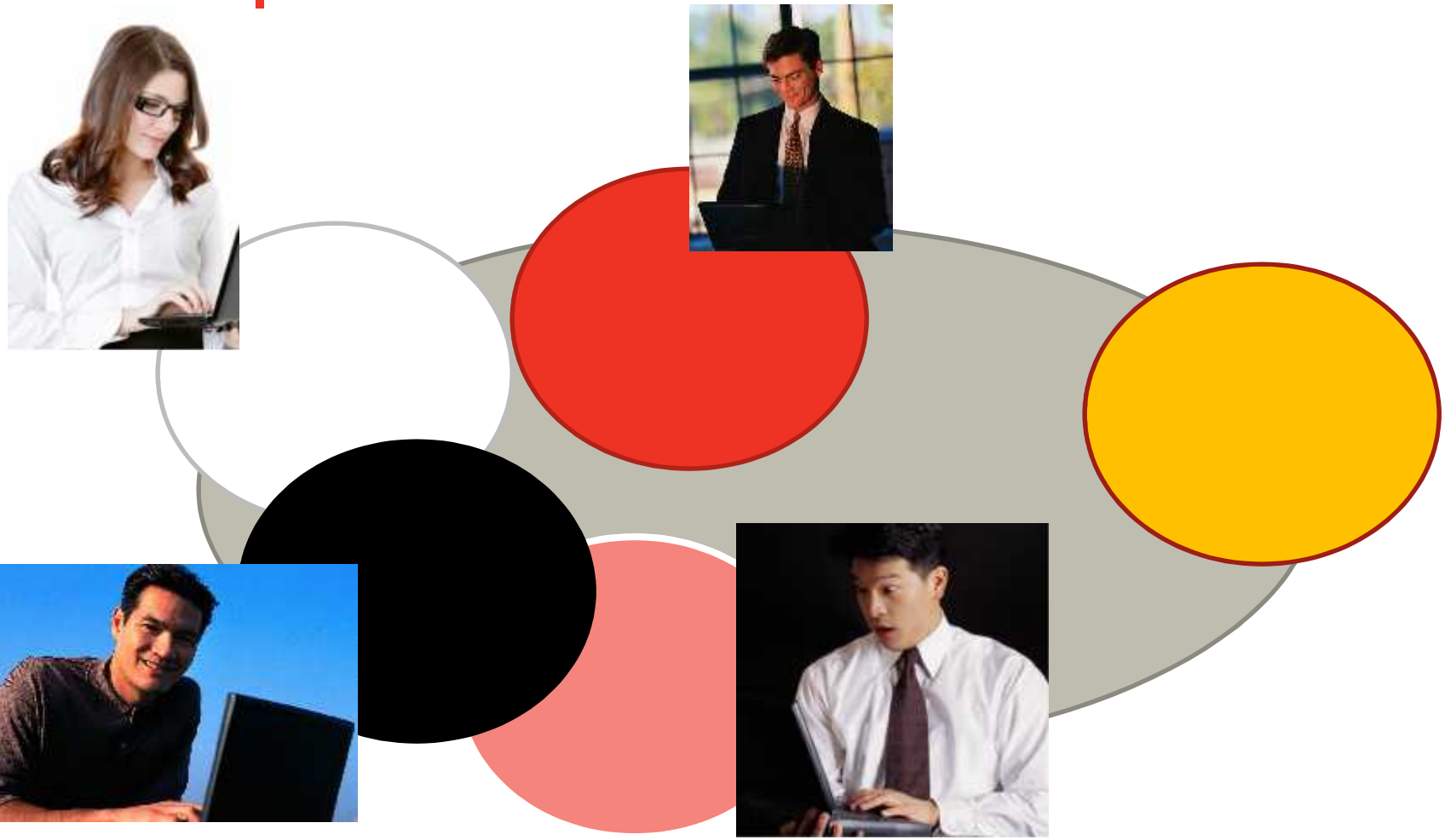
Maria: ministry expert on child poverty in Melbourne

Alert: must create a report!

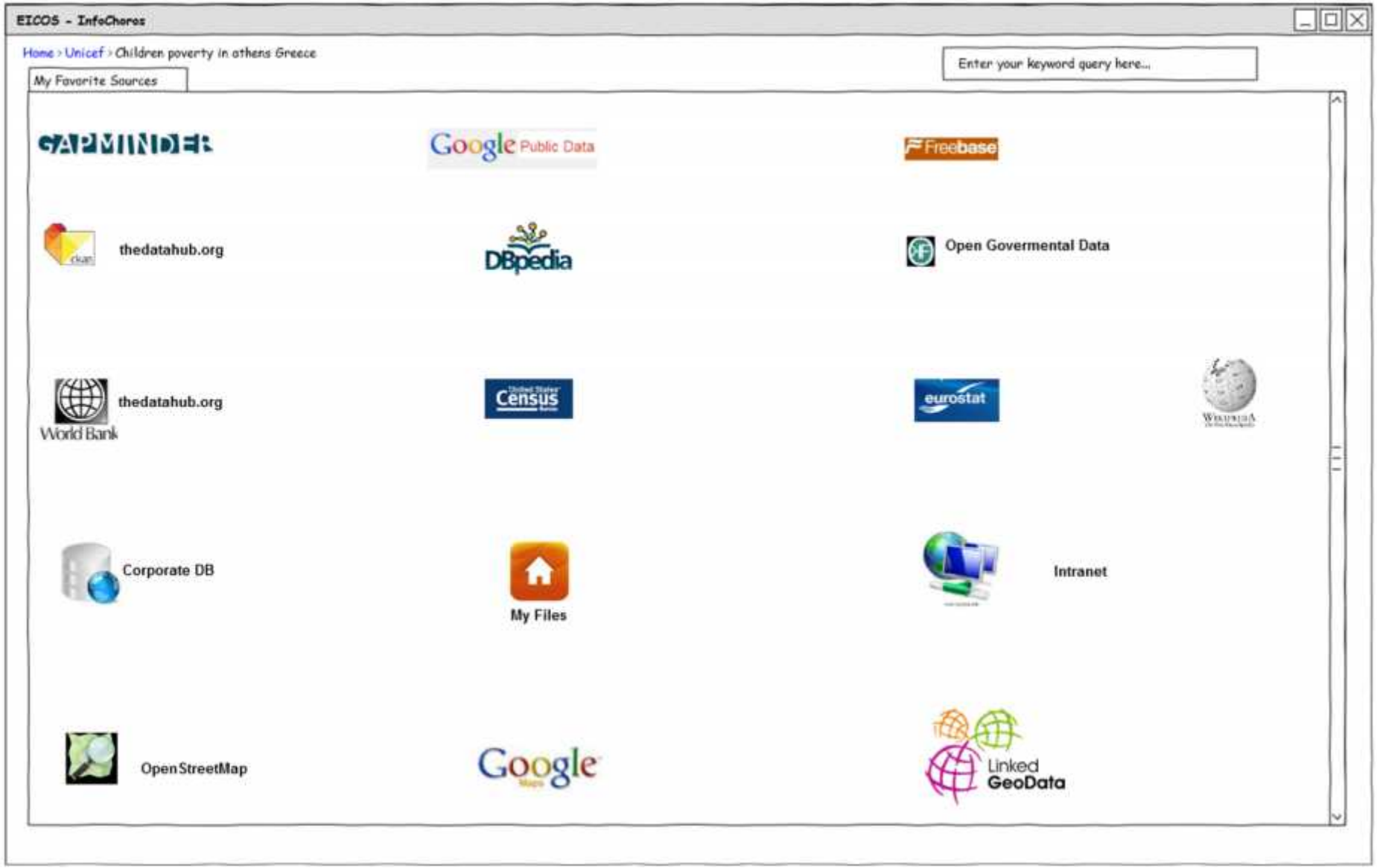
ToDo: create a report on child poverty during the past 5 years

George: veteran on matters of child welfare (Maria's boss)

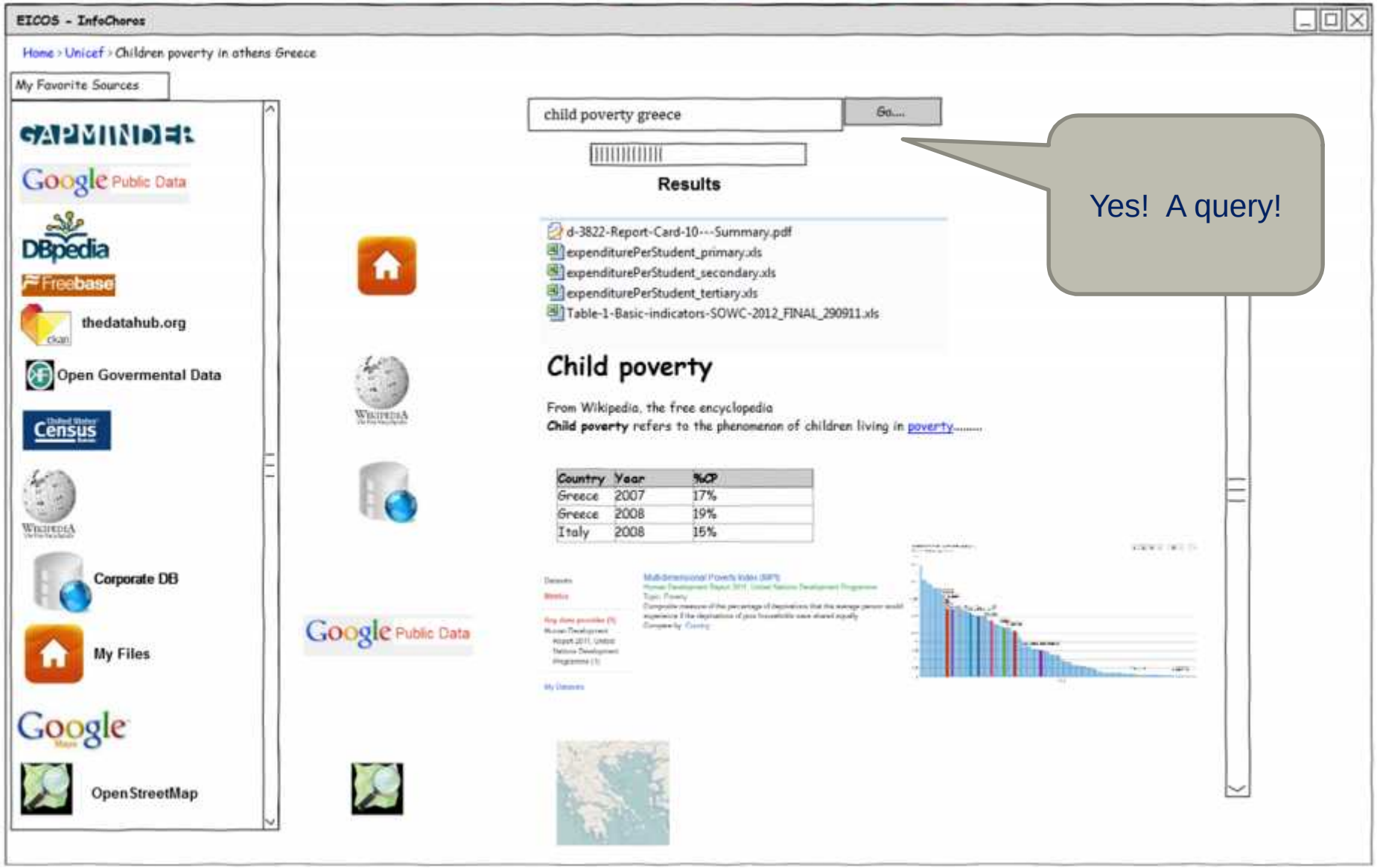
Enter InfoChoros: the ministry's dataspace



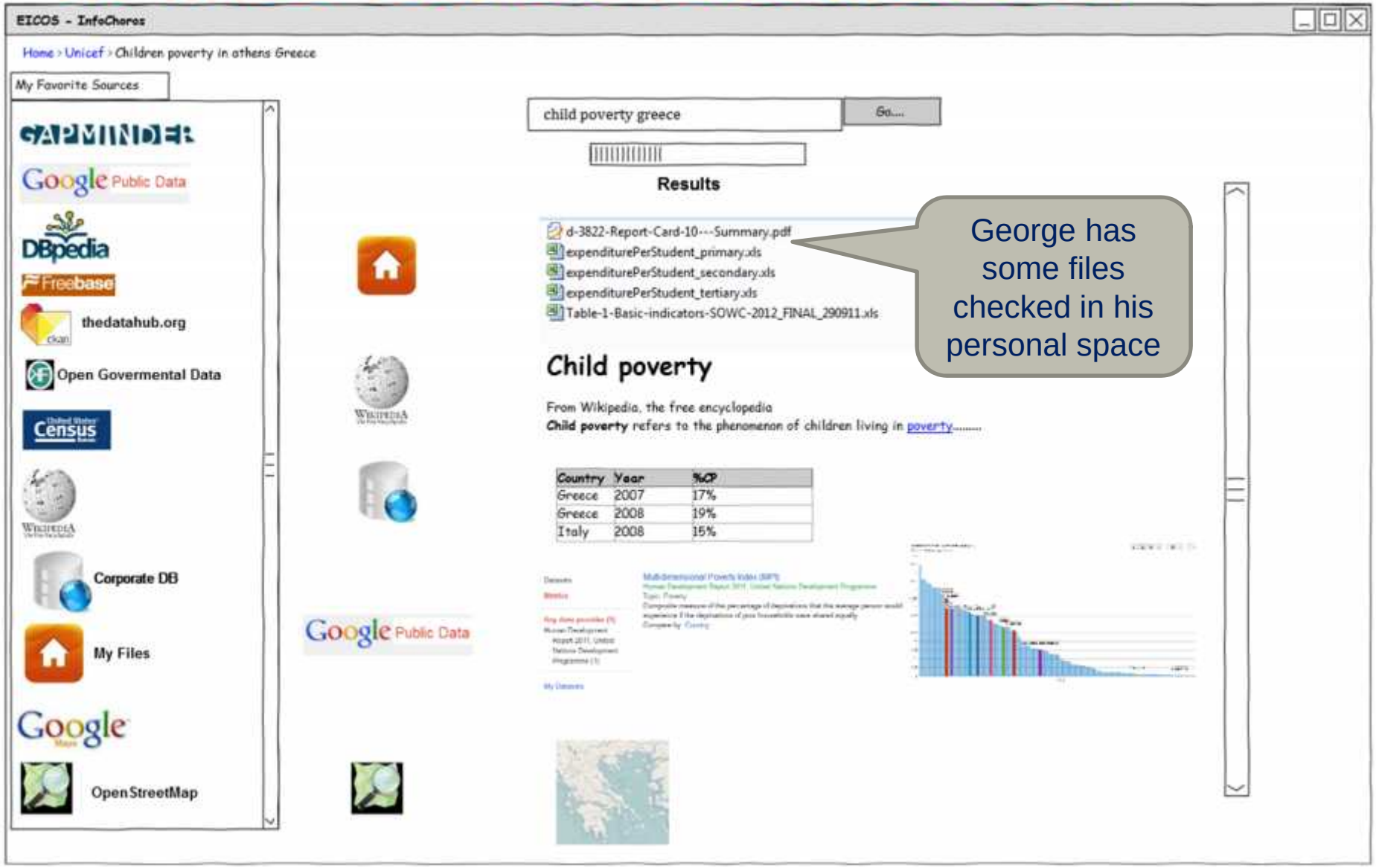
George logs in the system…Infochoros



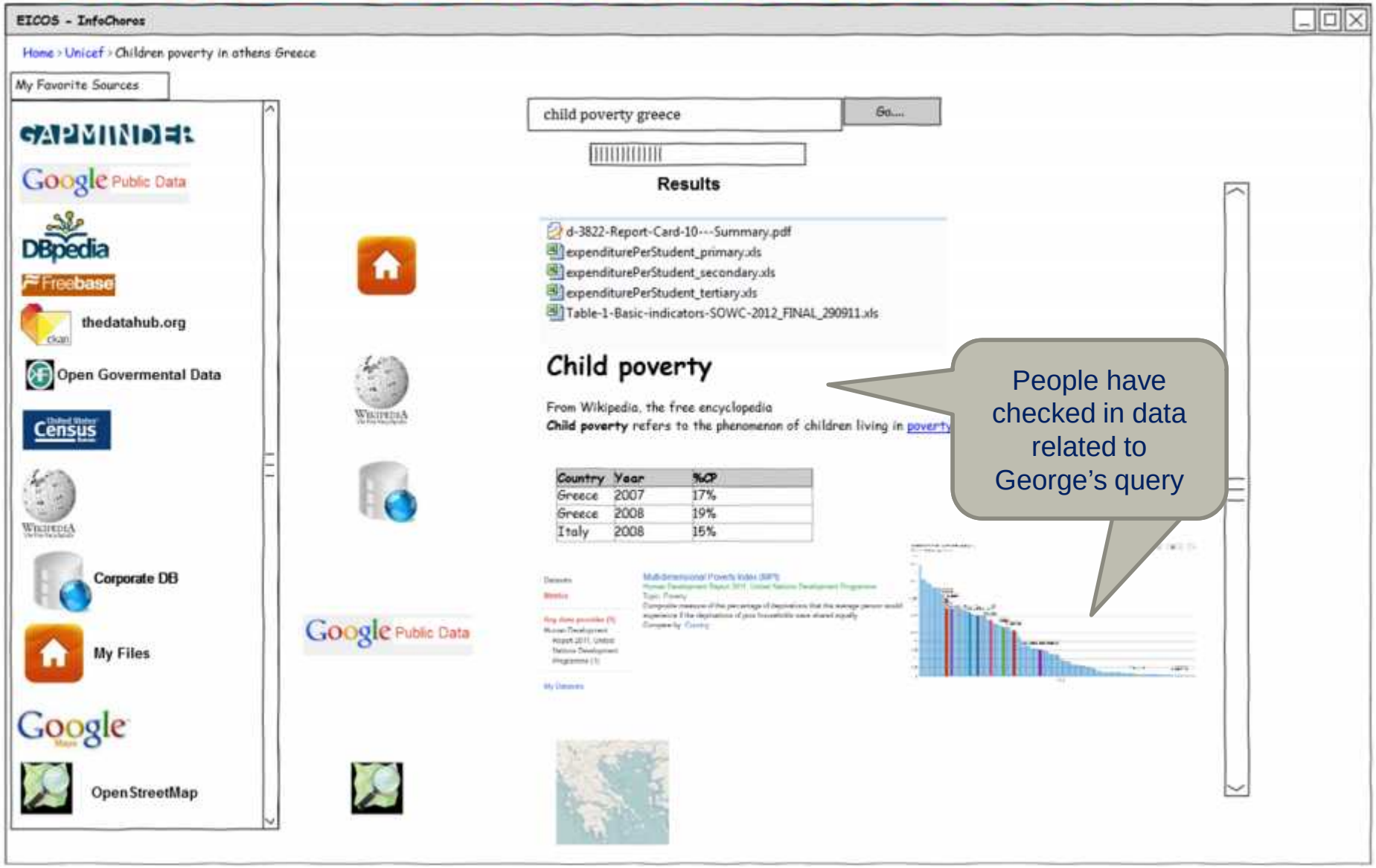
… poses a query & (magically) gets answers



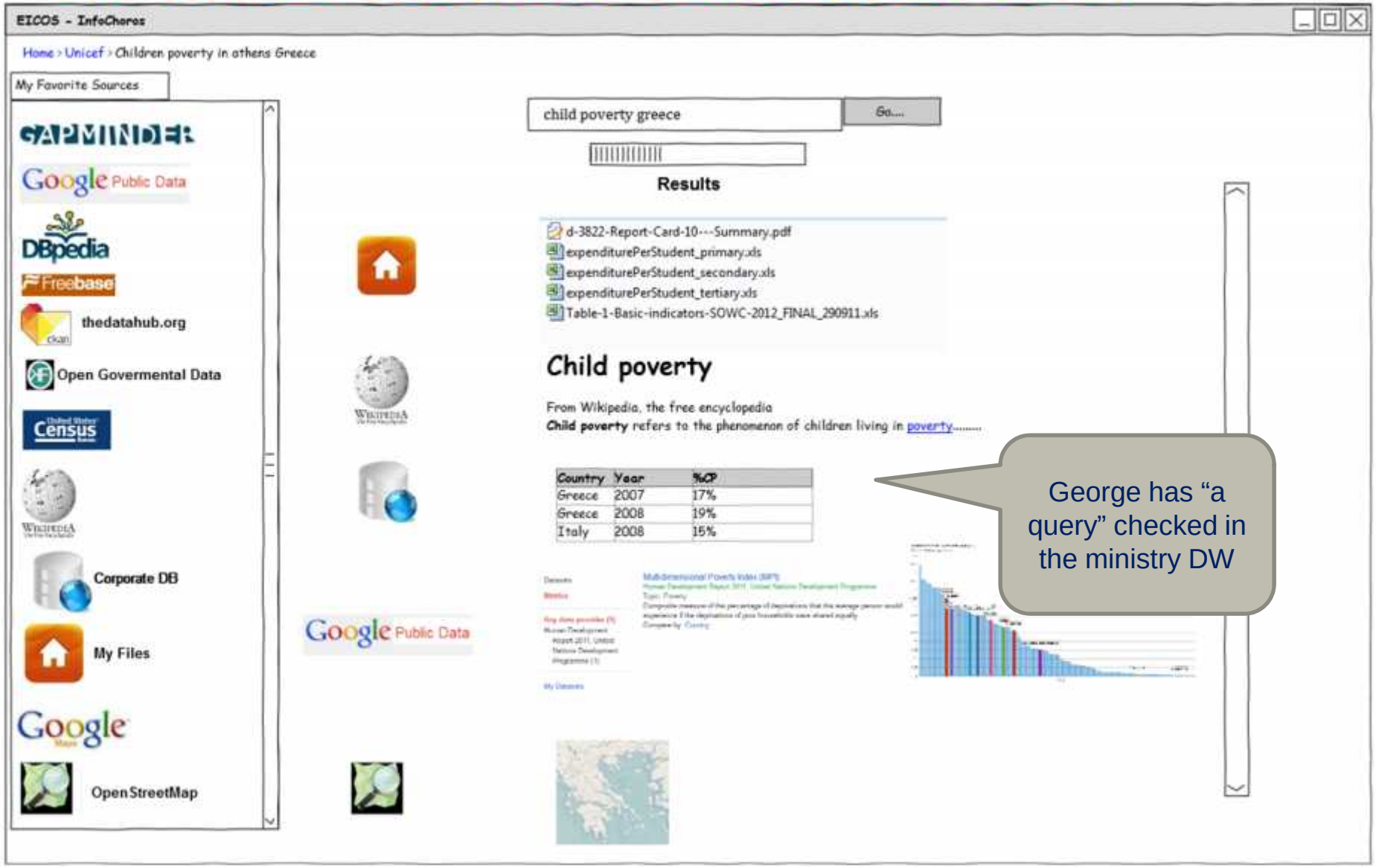
… finds related resources …



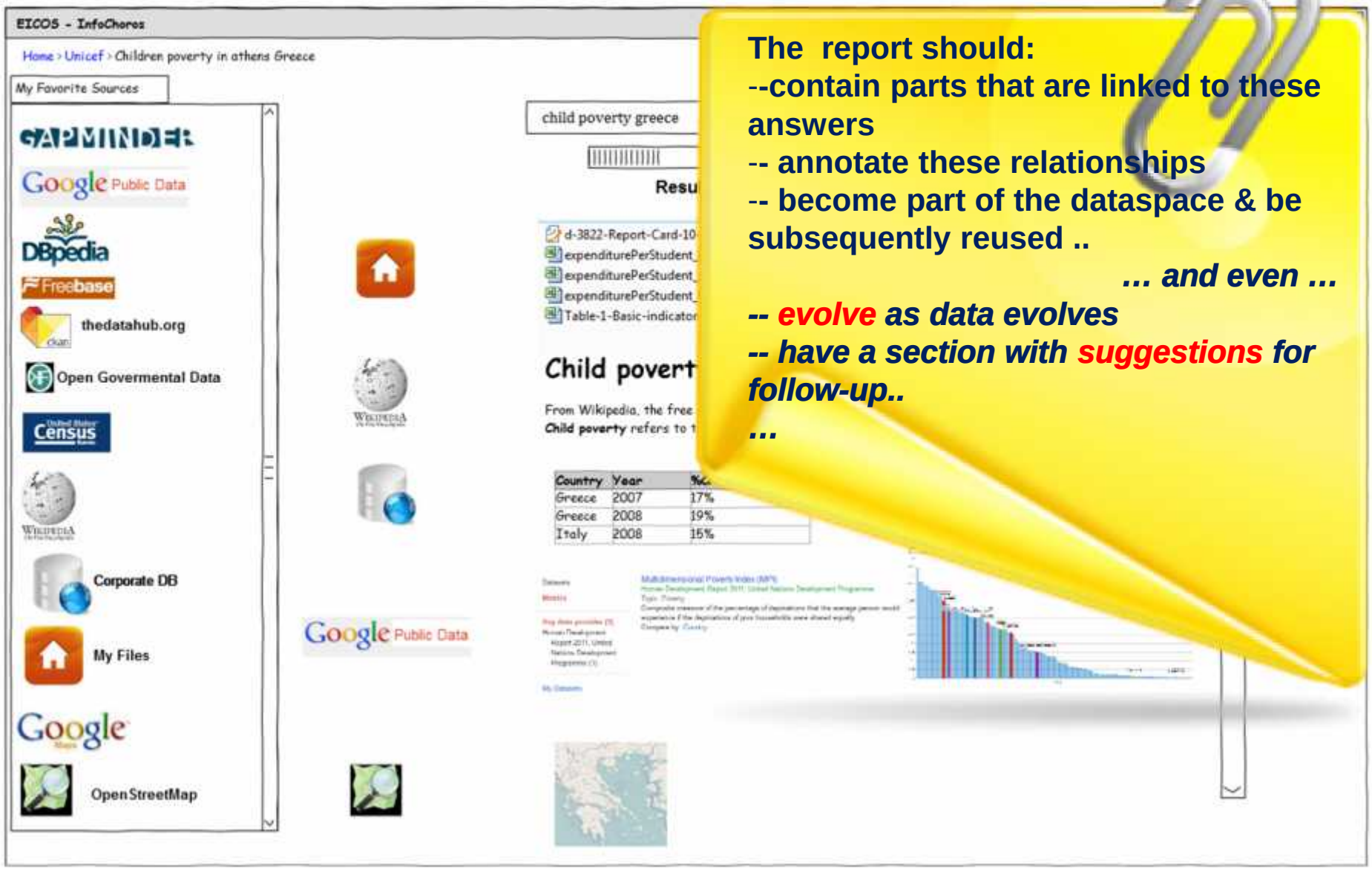
… finds related resources …



… finds related resources …



… poses a query & gets answers



Big Data @RMIT

RMIT UNIVERSITY

Data Analytics Lab

- Aims to open up this opportunity to Australia business and government partners, building on RMIT's existing track record of successful collaborations with partners
- Benefit partners in a diverse range of industries including manufacturing, utilities, transport and logistics, health, established and start-up ICT companies, as well as government agencies.
- Foster and train a new generation of researchers and research fellow experts in big data and data analytics and promote an environment of networking with other research centres, labs, and industry partners, at a national and international level (incl. Barcelona!)

Research Issues (1)

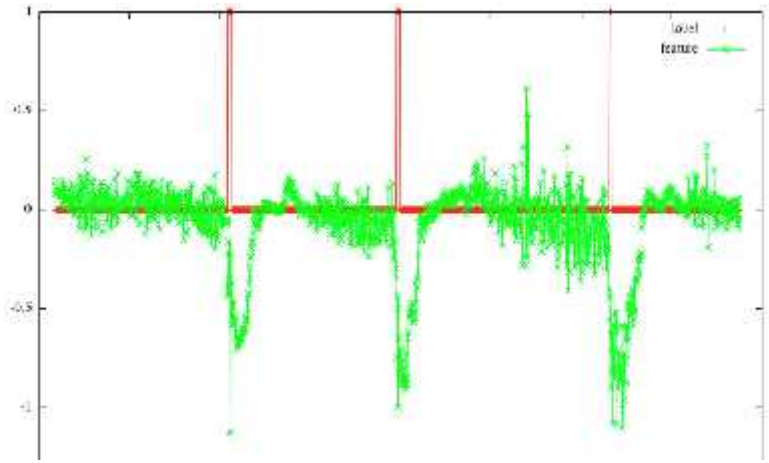
- Main stream
  - **Infrastructure and Architectures** (New large scale data architectures, Cloud architectures)
  - **Models** (Data representation, storage, and retrieval) and
  - **Data Access** (Query processing and optimization, Privacy, Security)

Research Issues (2)

- **Complex Data Analytics**
  - Computational, mathematical, statistical, and algorithmic techniques for modelling high dimensional data, large graphs, and complex (interrelated) data
  - Learning, inference, prediction, and knowledge discovery for large volumes of dynamic data sets
  - Data retrieval and data mining to facilitate pattern discovery, trend analysis and anomaly detection
  - Dimensionality reduction, sparse data

Research Issues (3)

- **Highly Streaming Data**
  - Positional streams
  - Social network data
  - Mobile app data
  - Game data



Excessive acceleration and deceleration

Research Issues (4)

- **Data Integration**
  - Findability and search
  - Information fusion of multiple data sources
  - Semantic integration
  - Recommendation systems

Research Themes

- **Situation Awareness** applications (Disaster Management, Transport)
- **Mobile/Social net analytics** applications (Disaster Management, Health, Design)
- **Financial analytics** applications (Trends, Fraud detection)
- **Smart Cities** applications (Energy, Design)

Future …. what to expect

Word Cloud "Big Data" © Ben Chams

- Interesting research issues
- Innovation
- New perspectives .. ***data science***
- Impact on several areas (health, manufacturing, energy, climate, etc)