

Big Data Infrastructures: Exploiting the Power of Big Data

Timos Sellis

School of CS & IT

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

Where is the big data?



US health care

- \$300 billion value per year
- ~0.7 percent annual productivity growth



Europe public sector administration

- €250 billion value per year
- ~0.5 percent annual productivity growth



Global personal location data

- \$100 billion+ revenue for service providers
- Up to \$700 billion value to end users



US retail

- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth



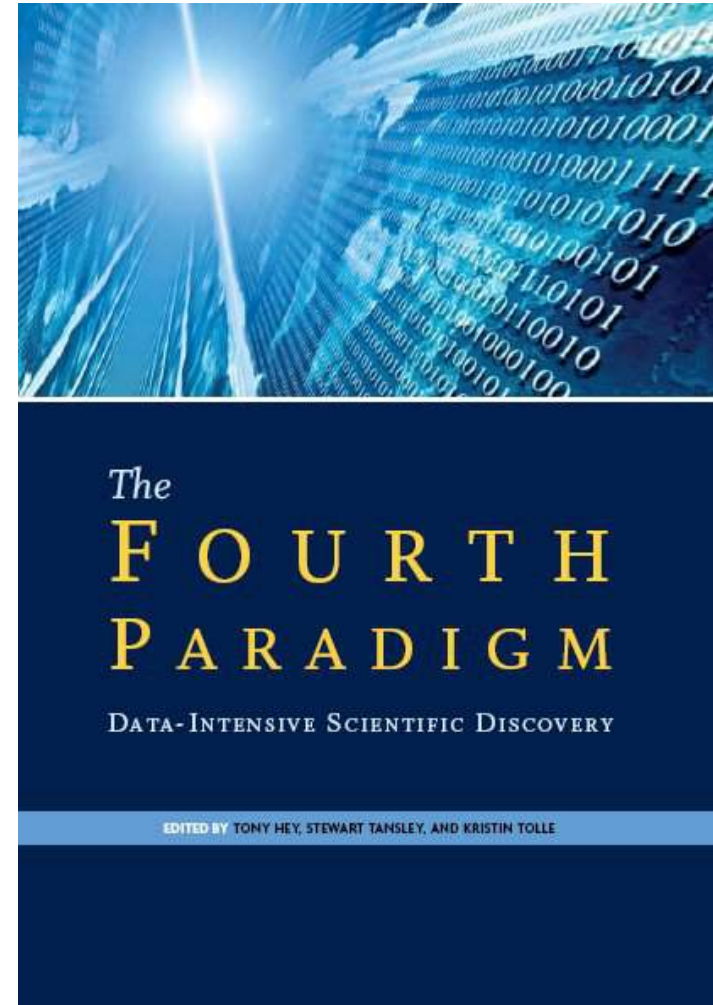
Manufacturing

- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

SOURCE: McKinsey Global Institute analysis

A paradigm shift - Science

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



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.

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.

Vestas Wind Energy Turbine Placement and Maintenance



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.

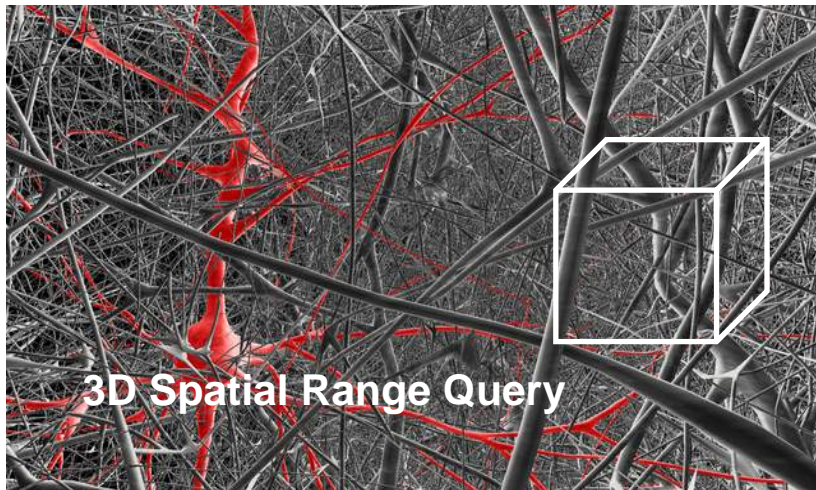
To make sense out of this wave of data, scientists are developing sophisticated ways to store, retrieve and analyze it. A new breed of “data scientist” is working to re-invent the traditional drug research team. Instead of biologists, chemists and clinicians working in silos, pharmaceutical companies 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>



Is there power in the data?

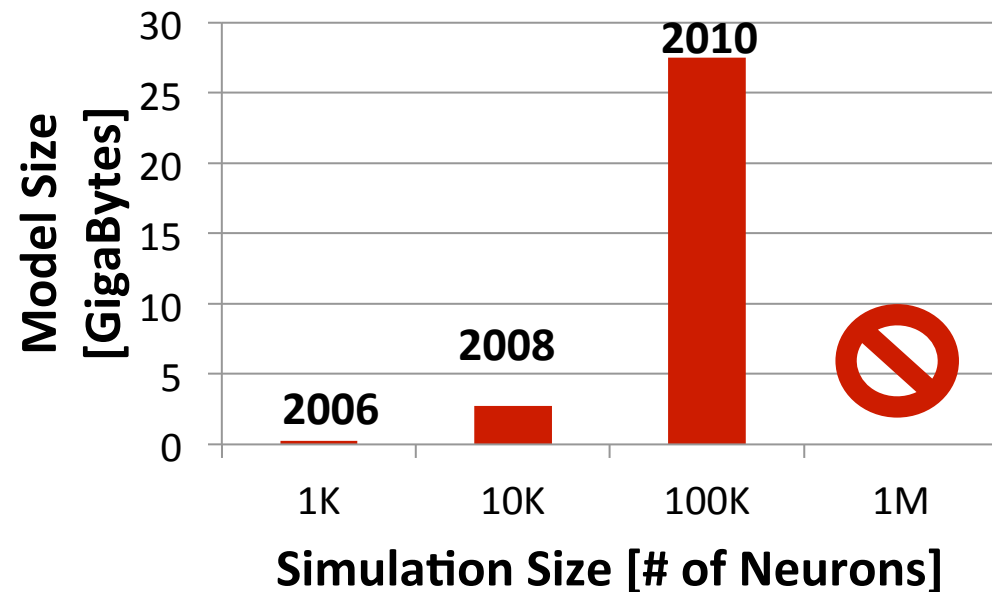
Human Brain Project (EU), <http://www.humanbrainproject.eu/>



86 billion of neurons
100 trillion of synapses

develop platform to **simulate human brain!**

Bottleneck in Spatial Analysis



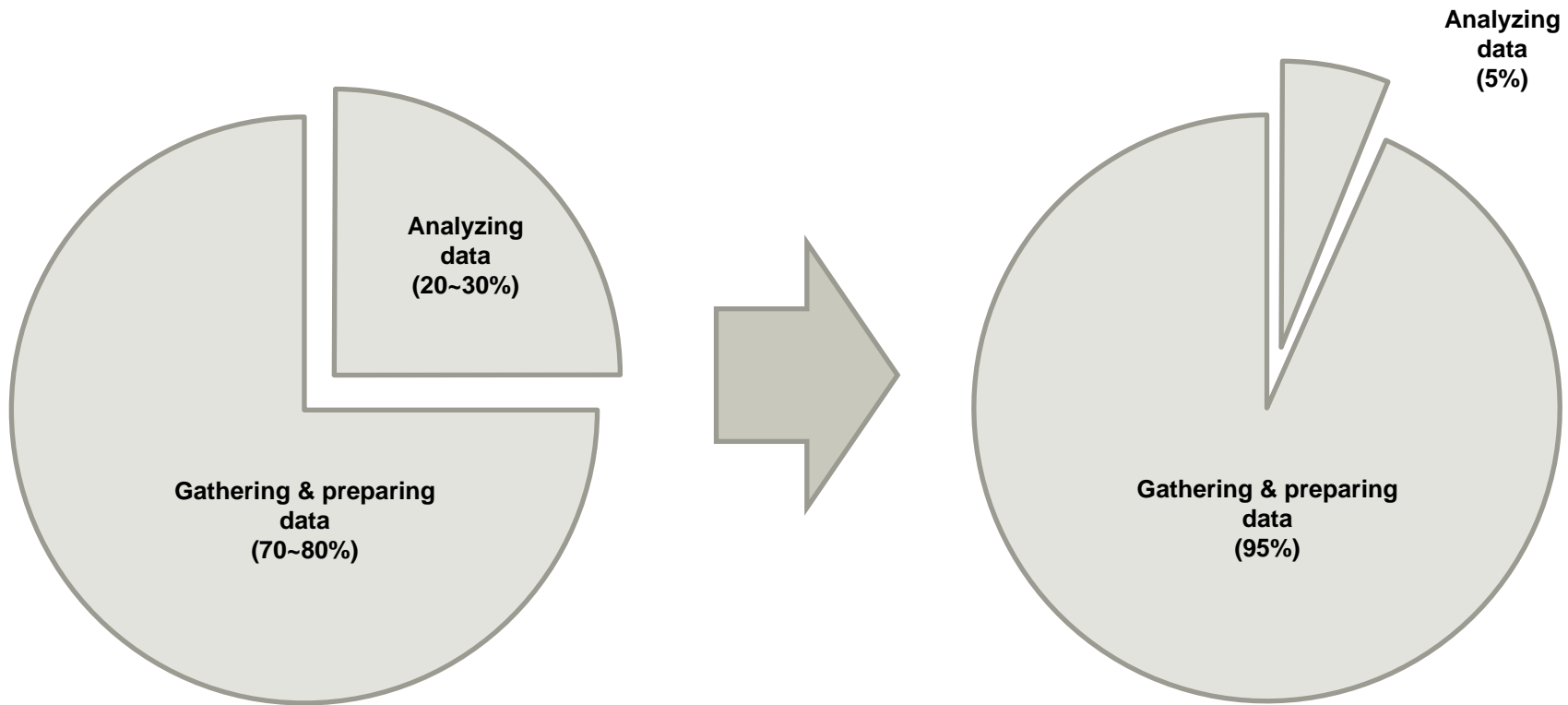
Main Issue - Big Data Analysis

- 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

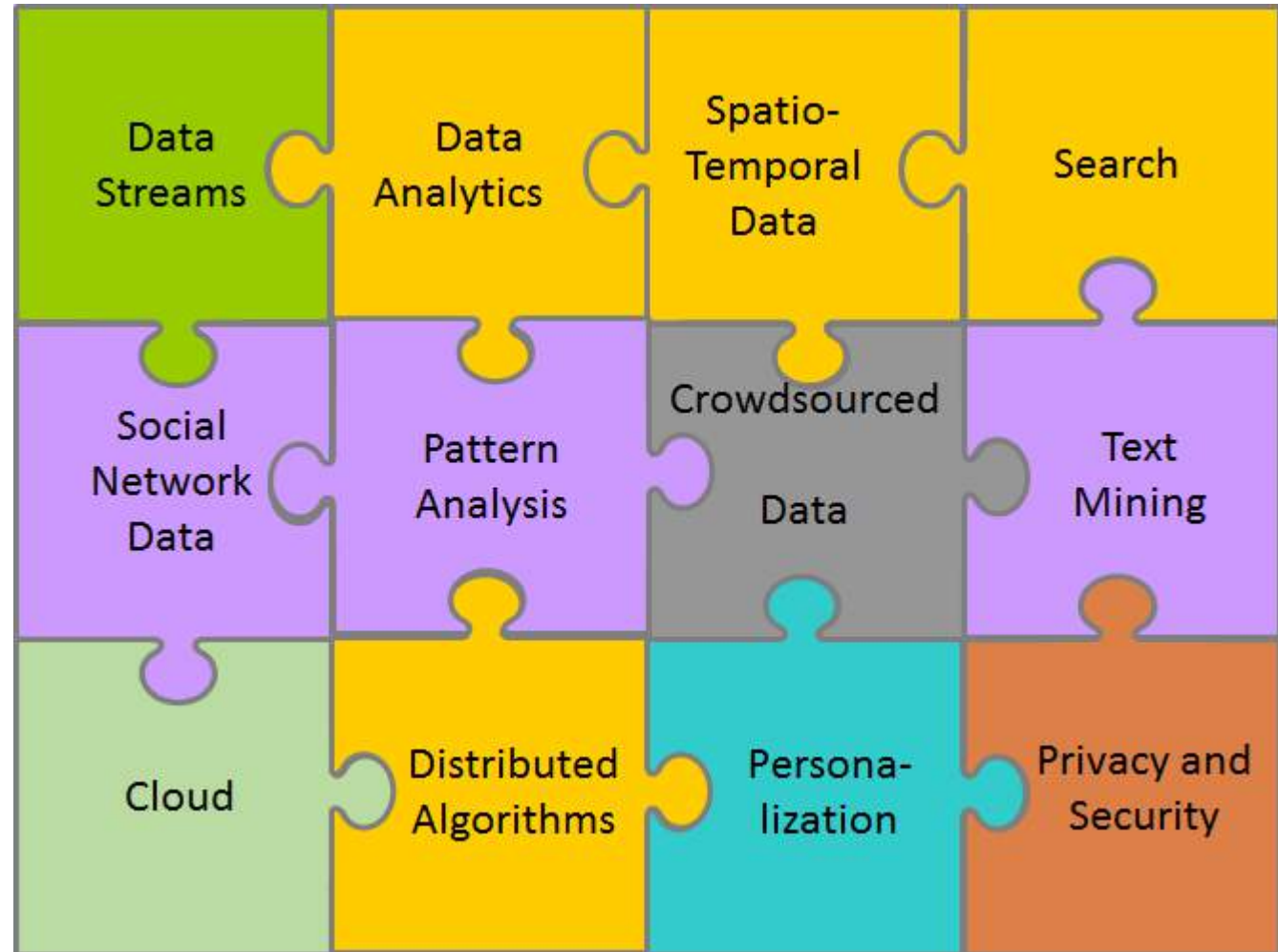
The time for developing an analysis (Initially working with big data)



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 viewBreaking news



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

Maria.: ministry expert on child poverty in Catalonia

Alert: must create a report!



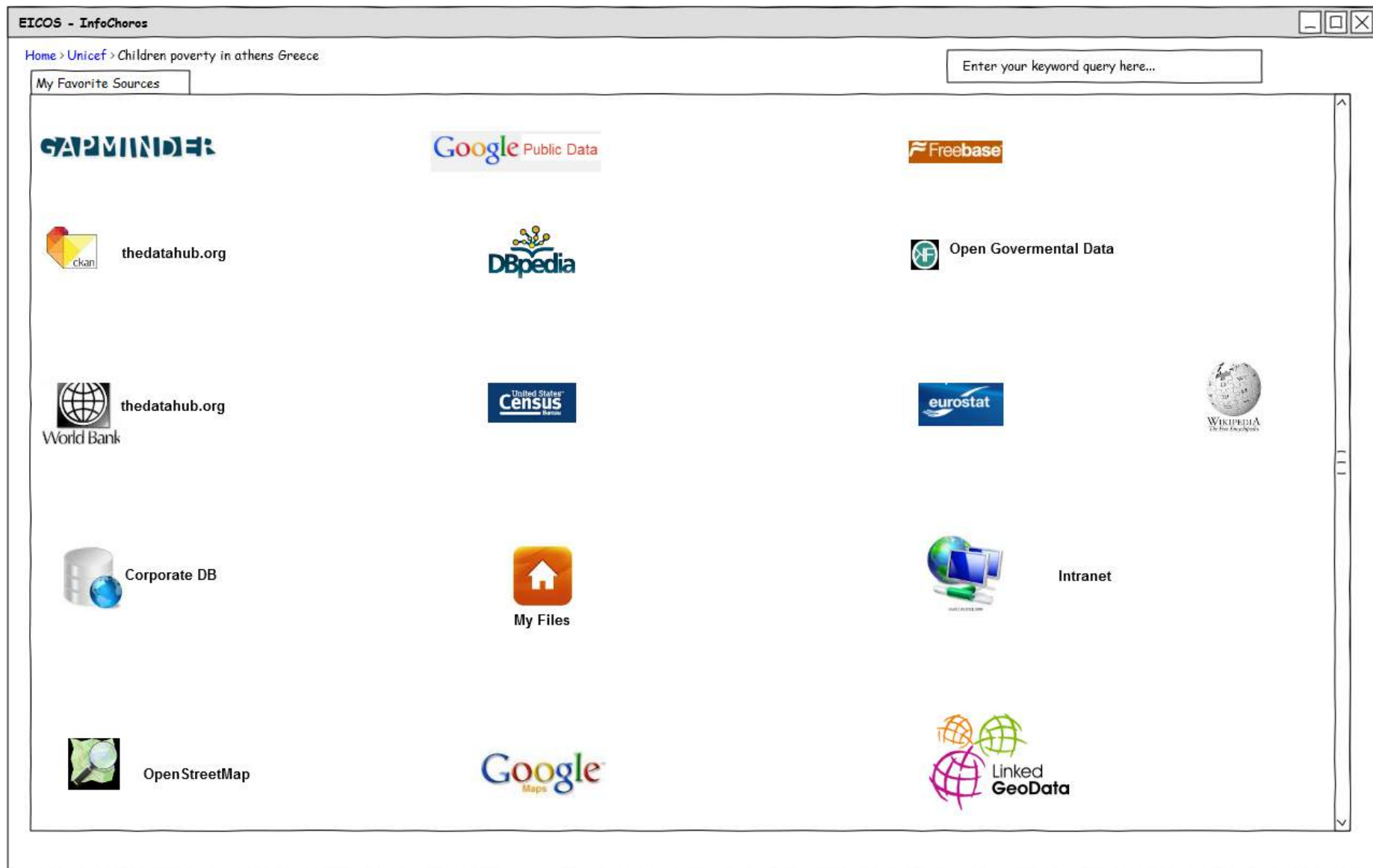
Jorge: veteran on matters of child welfare (Maria's boss)



Enter InfoChoros: the ministry's dataspace



Jorge logs in the system...Infochoros



... poses a query & (magically) gets answers

The screenshot shows the EICOS - InfoChoros web application. At the top, the breadcrumb trail reads "Home > Unicef > Children poverty in athens Greece". A search bar contains the text "child poverty greece" and a "Go..." button. Below the search bar is a progress indicator and the word "Results".

The search results list several files:

- d-3822-Report-Card-10---Summary.pdf
- expenditurePerStudent_primary.xls
- expenditurePerStudent_secondary.xls
- expenditurePerStudent_tertiary.xls
- Table-1-Basic-indicators-SOWC-2012_FINAL_290911.xls

The main result is titled "Child poverty". It includes a Wikipedia snippet: "From Wikipedia, the free encyclopedia Child poverty refers to the phenomenon of children living in poverty.....".

Country	Year	%CP
Greece	2007	17%
Greece	2008	19%
Italy	2008	15%

Below the table, there are sections for "Metadata", "Metrics", and "Any data provider (1)". A bar chart on the right shows data for various countries, with Greece highlighted. A map of Greece is visible at the bottom of the results area.

A speech bubble on the right side of the interface says "Yes! A query!".

The left sidebar, titled "My Favorite Sources", lists various data sources with icons: GAPMINDER, Google Public Data, DBpedia, Freebase, thedatahub.org, Open Governmental Data, United States Census, Wikipedia, Corporate DB, My Files, Google Maps, and OpenStreetMap.

... finds related resources ...

EICOS - InfoChoros

Home > Unicef > Children poverty in athens Greece

My Favorite Sources

- GAPMINDER
- Google Public Data
- DBpedia
- Freebase
- thedatahub.org
- Open Governmental Data
- United States Census
- WIKIPEDIA
- Corporate DB
- My Files
- Google
- OpenStreetMap

child poverty greece

Go...

Results

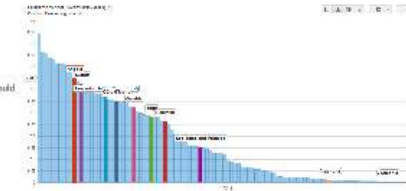
- d-3822-Report-Card-10---Summary.pdf
- expenditurePerStudent_primary.xls
- expenditurePerStudent_secondary.xls
- expenditurePerStudent_tertiary.xls
- Table-1-Basic-indicators-SOWC-2012_FINAL_290911.xls

Child poverty


From Wikipedia, the free encyclopedia
Child poverty refers to the phenomenon of children living in [poverty](#).....

Country	Year	%CP
Greece	2007	17%
Greece	2008	19%
Italy	2008	15%

Dataset: Multidimensional Poverty Index (MPI)
Human Development Report 2011, United Nations Development Programme
Topic: Poverty
Composite measure of the percentage of deprivations that the average person would experience if the deprivations of poor households were shared equally.
Compare by: Country



My Datasets



Jorge has some files checked in his personal space

... finds related resources ...

EICOS - InfoChoros

Home > Unicef > Children poverty in athens Greece

My Favorite Sources

- GAPMINDER
- Google Public Data
- DBpedia
- Freebase
- thedatahub.org
- Open Governmental Data
- United States Census
- WIKIPEDIA
- Corporate DB
- My Files
- Google
- OpenStreetMap

child poverty greece

Go...

Results

- d-3822-Report-Card-10---Summary.pdf
- expenditurePerStudent_primary.xls
- expenditurePerStudent_secondary.xls
- expenditurePerStudent_tertiary.xls
- Table-1-Basic-indicators-SOWC-2012_FINAL_290911.xls


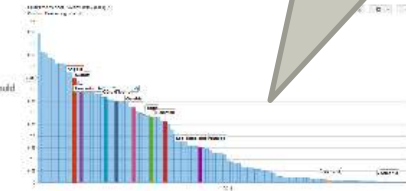
Child poverty

From Wikipedia, the free encyclopedia
Child poverty refers to the phenomenon of children living in poverty

Country	Year	%CP
Greece	2007	17%
Greece	2008	19%
Italy	2008	15%

People have checked in data related to Jorge's query

Dataset: Multidimensional Poverty Index (MPI)
Human Development Report 2011, United Nations Development Programme
Topic: Poverty
Composite measure of the percentage of deprivations that the average person would experience if the deprivations of poor households were shared equally.
Compare by: Country



... finds related resources ...

EICOS - InfoChoros

Home > Unicef > Children poverty in athens Greece

My Favorite Sources

- GAPMINDER
- Google Public Data
- DBpedia
- Freebase
- thedatahub.org
- Open Governmental Data
- United States Census
- WIKIPEDIA
- Corporate DB
- My Files
- Google
- OpenStreetMap

child poverty greece

Go...

Results

- d-3822-Report-Card-10---Summary.pdf
- expenditurePerStudent_primary.xls
- expenditurePerStudent_secondary.xls
- expenditurePerStudent_tertiary.xls
- Table-1-Basic-indicators-SOWC-2012_FINAL_290911.xls

Child poverty

From Wikipedia, the free encyclopedia
Child poverty refers to the phenomenon of children living in [poverty](#).....

Country	Year	%CP
Greece	2007	17%
Greece	2008	19%
Italy	2008	15%

Datasets

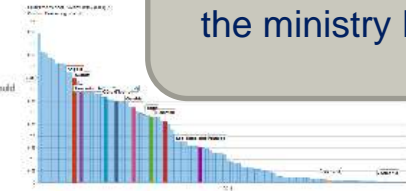
Metrics

Any data provider (1)

Human Development Report 2011, United Nations Development Programme (1)

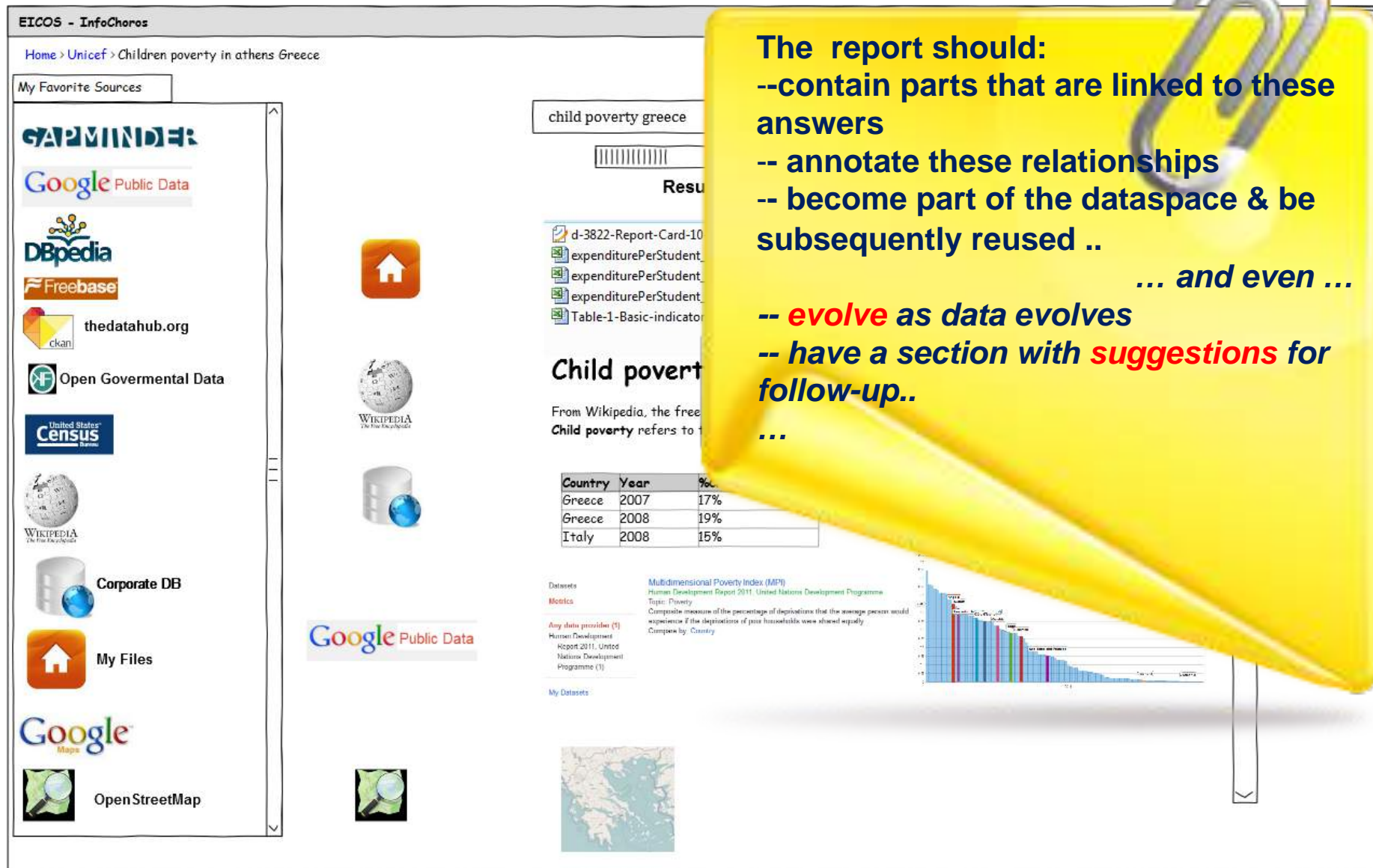
My Datasets

Multidimensional Poverty Index (MPI)
Human Development Report 2011, United Nations Development Programme
Type: Poverty
Composite measure of the percentage of deprivations that the average person would experience if the deprivations of poor households were shared equally.
Compare by: Country



Jorge has "a query" checked in the ministry DW

... poses a query & gets answers



The screenshot shows the EICOS - InfoChoros interface. The search bar contains 'child poverty greece'. The results list includes 'd-3822-Report-Card-10', 'expenditurePerStudent', and 'Table-1-Basic-indicator'. A table titled 'Child poverty' shows data for Greece in 2007 (17%), Greece in 2008 (19%), and Italy in 2008 (15%). A yellow sticky note is overlaid on the right side of the interface, containing the following text:

The report should:

- contain parts that are linked to these answers
- annotate these relationships
- become part of the dataspace & be subsequently reused ..

... and even ...

- *evolve* as data evolves
- have a section with *suggestions* for follow-up..

...

Country	Year	%
Greece	2007	17%
Greece	2008	19%
Italy	2008	15%

The sticky note also features a paperclip icon at the top right and a vertical line at the bottom right.

Big Data @RMIT

www.rmit.edu.au

Data Analytics Lab



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)

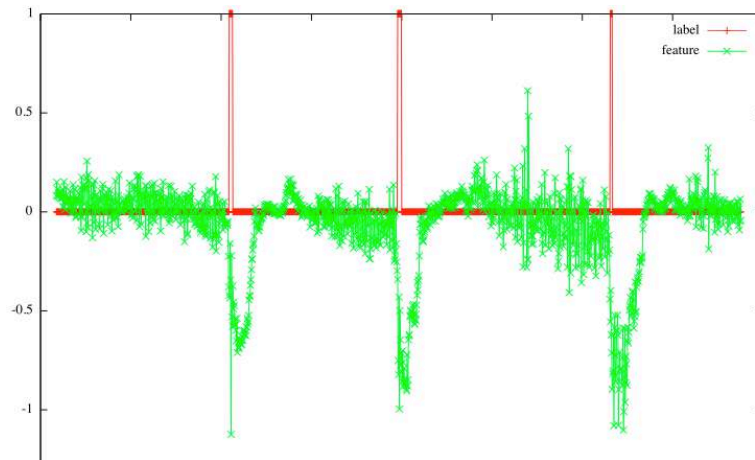
- **Complex Data Analytics**

- 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

Where is that document?



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)

My RMIT Vision

- Establishing a new research Center on Big Data Infrastructures (BDI)
- Target areas: Research, Social, Government Data
- Good collective experience at RMIT
- Target multiple stakeholders: researchers, industry, government
- Aligned with grand challenges internationally (US, EU)

