Big Data – Retos y Oportunidades

Raúl Ramos Pollán

Líder Big Data & Large Scale Machine Learning
Laboratorio de Supercomputación y Cálculo Científico
Universidad Industrial de Santander – Bucaramanga, Colombia

http://sc3.uis.edu.co

Avalancha de datos – *Data Deluge*

McKinsey Global Institute
June 2011
Big data: The ne
for innovation
and product
The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
What is Data Science?
The future belongs to the companies and people that turn data into products
An O'Reilly Radar Report
By Mike Loukides
COMMUNICATIONS OF THE ACM
CACM.ACM.ORG
12/08 VOL.51 NO.12
Surviving the Data Deluge
Open Informa
Ext
from
The Center for Information Technology Policy Presents
BIG DATA
Public Policy and the Exploding Digital Corpus
A One Day Conference
November 30, 2010, 8am–5pm
Friend Center Convocation Room
nature
BIG DATA
SCIENCE IN THE PETABYTE ERA
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
NATUREJOBS
Minnesota musings

In the same way that past Federal investments in information-technology R&D led to dramatic advances in supercopmuting and the creation of the Internet, the initiative we are launching today promises to transform our ability to use **Big Data** for scientific discovery, environmental and biomedical research, education, and national security

Dr. J Holdren,
Director of the White House Office of Science and Technology ddd2012

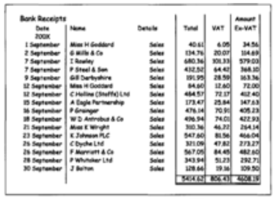
The World Economic Forum convened in Switzerland in January, 2012 highlighted **Big Data** as a new economic asset comparable to currency and gold

Rethinking Personal Data: Strengthening Trus
World Economic Forum. May, 2012

Today, we are experiencing a major shift in decision making driven by several factors: **unprecedented amounts of data** from a variety of sources […] on a variety of socio-economic, technological, and ecological systems, our vastly increasing ability to store and perform computation over very large data sets …

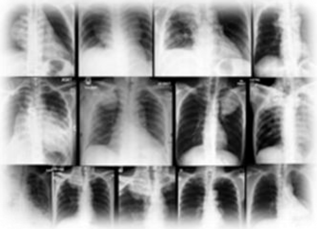
EU Digital Agenda, May 2013

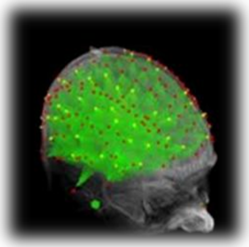
Sources for Big Data

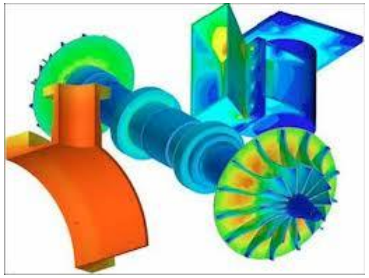


Bank Receipts

Date 200X	Name	Details	Total	VAT	Amount Ex-VAT
1 September	Miss H Goddard	Sales	40.61	6.05	34.56
2 September	G Mills & Co	Sales	134.76	20.07	114.69
7 September	I Rowley	Sales	680.36	101.33	579.03
7 September	P Steel & Son	Sales	432.52	64.42	368.10
9 September	Gill Darbyshire	Sales	191.95	28.59	163.36
12 September	Miss H Goddard	Sales	84.60	12.60	72.00
12 September	C Hollins (Stoffs) Ltd	Sales	484.57	72.17	412.40
15 September	A Eagle Partnership	Sales	173.47	25.84	147.63
16 September	P Grainger	Sales	476.14	70.91	405.23
18 September	W D Antrobus & Co	Sales	496.94	74.01	422.93
21 September	Miss K Wright	Sales	310.36	46.22	264.14
23 September	K Johnson PLC	Sales	547.60	81.56	466.04
26 September	C Dyche Ltd	Sales	321.09	47.82	273.27
26 September	F Marriott & Co	Sales	567.05	84.45	482.60
28 September	P Whitaker Ltd	Sales	343.94	51.23	292.71
30 September	J Bolton	Sales	128.66	19.16	109.50
			5414.62	806.43	4608.19







consumer logs

mobile devices

social networks

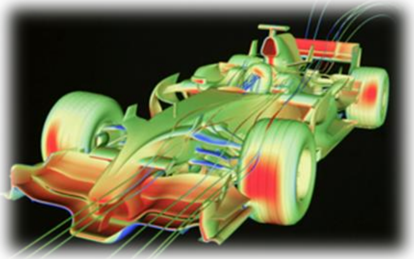
industry

patient registries

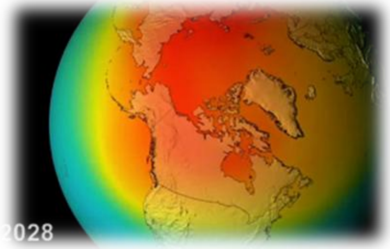
document repositories

bank transactions

web logs



government



sensors, instrumentation

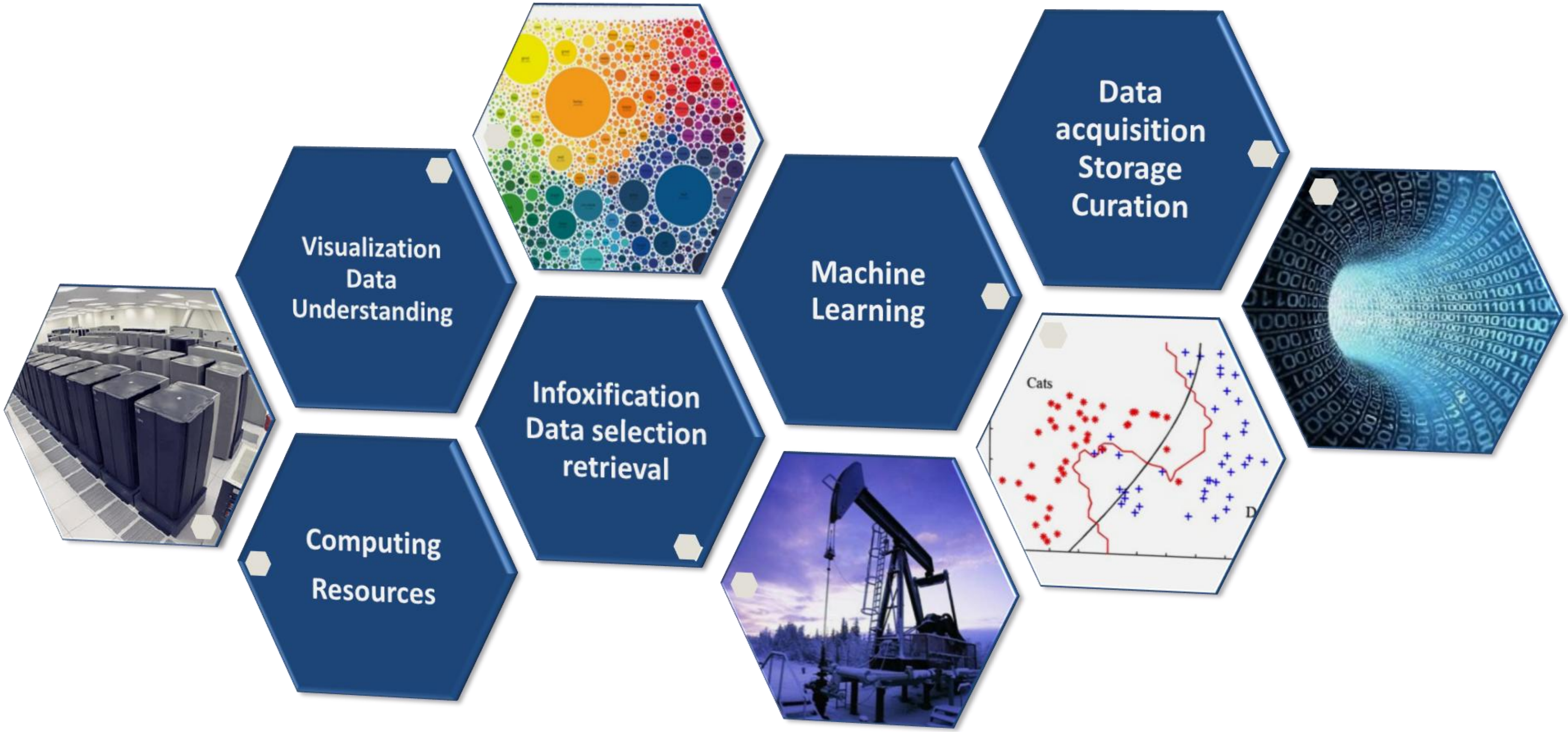
simulations

science

image acquisition

biological data

Data Science / Big Data



Visualization Data Understanding

Data acquisition Storage Curation

Machine Learning

Infoxification Data selection retrieval

Computing Resources

Bussiness Intelligence → managing existing data for monitoring management concerns

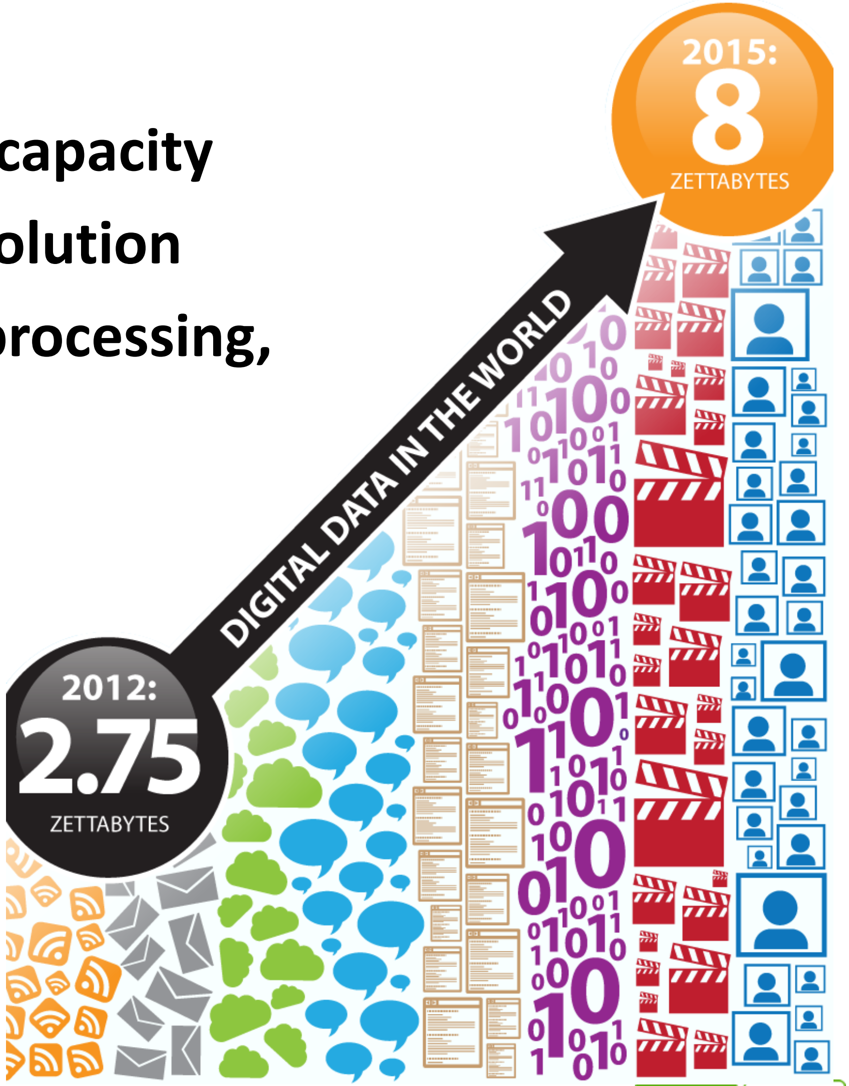
Data Science → discovering knowledge

Big Data

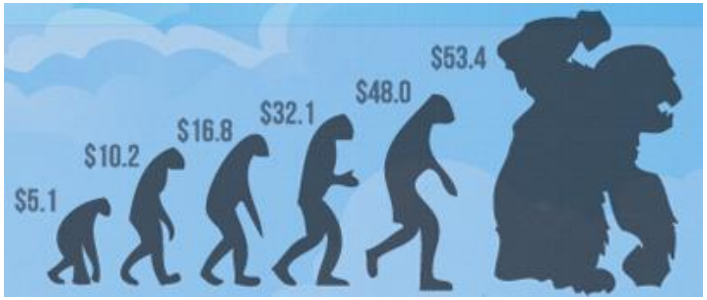
datasets size >> traditional DB capacity

datasets groth >> tecnology evolution

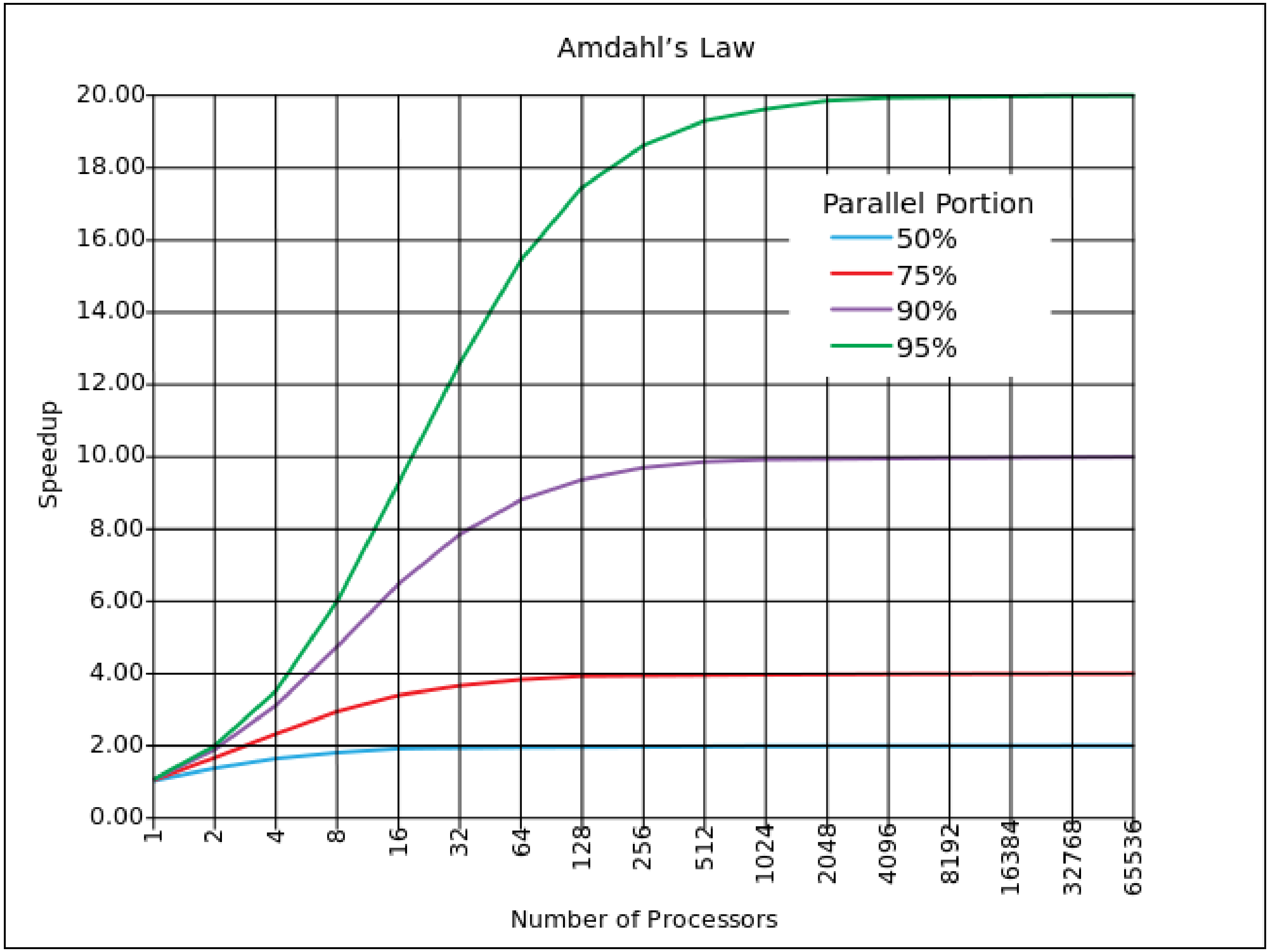
new models & techniques for processing, analyzing, visualizing



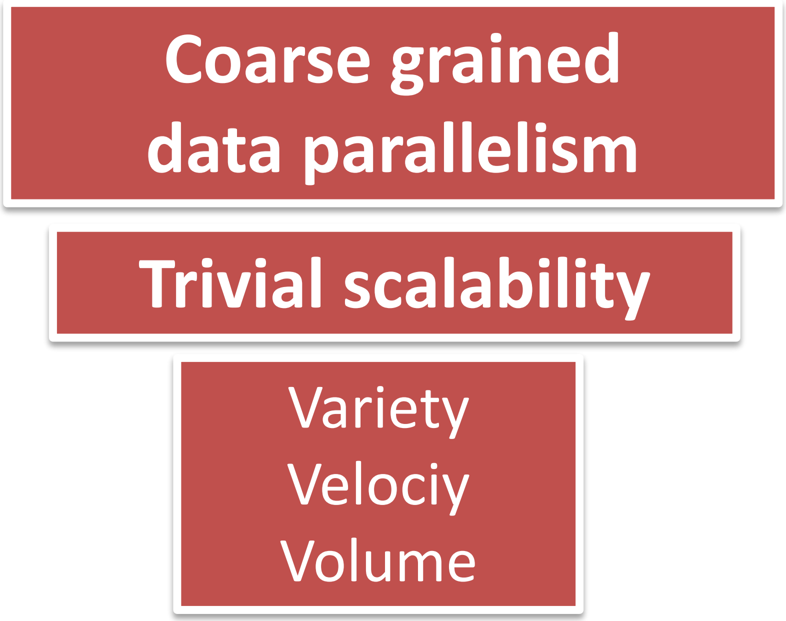
BIG DATA MARKET FORECAST (US BILLIONS)



Ley de Amdahl



Big Data



Coarse grained data parallelism

Trivial scalability

Variety
Velociy
Volume

Big Data

mongoDB

amazon webservices

Amazon DynamoDB

hadoop

HIVE

Coarse grained data parallelism

APACHE HBASE

POWERED BY cloudera IMPALA

Trivial scalability

Cassandra

Variety

Velociy

Volume

tableau SOFTWARE

Spark

Lightning-Fast Cluster Computing

pentaho

OpenNebula.org

The Open Source Toolkit for Cloud Computing

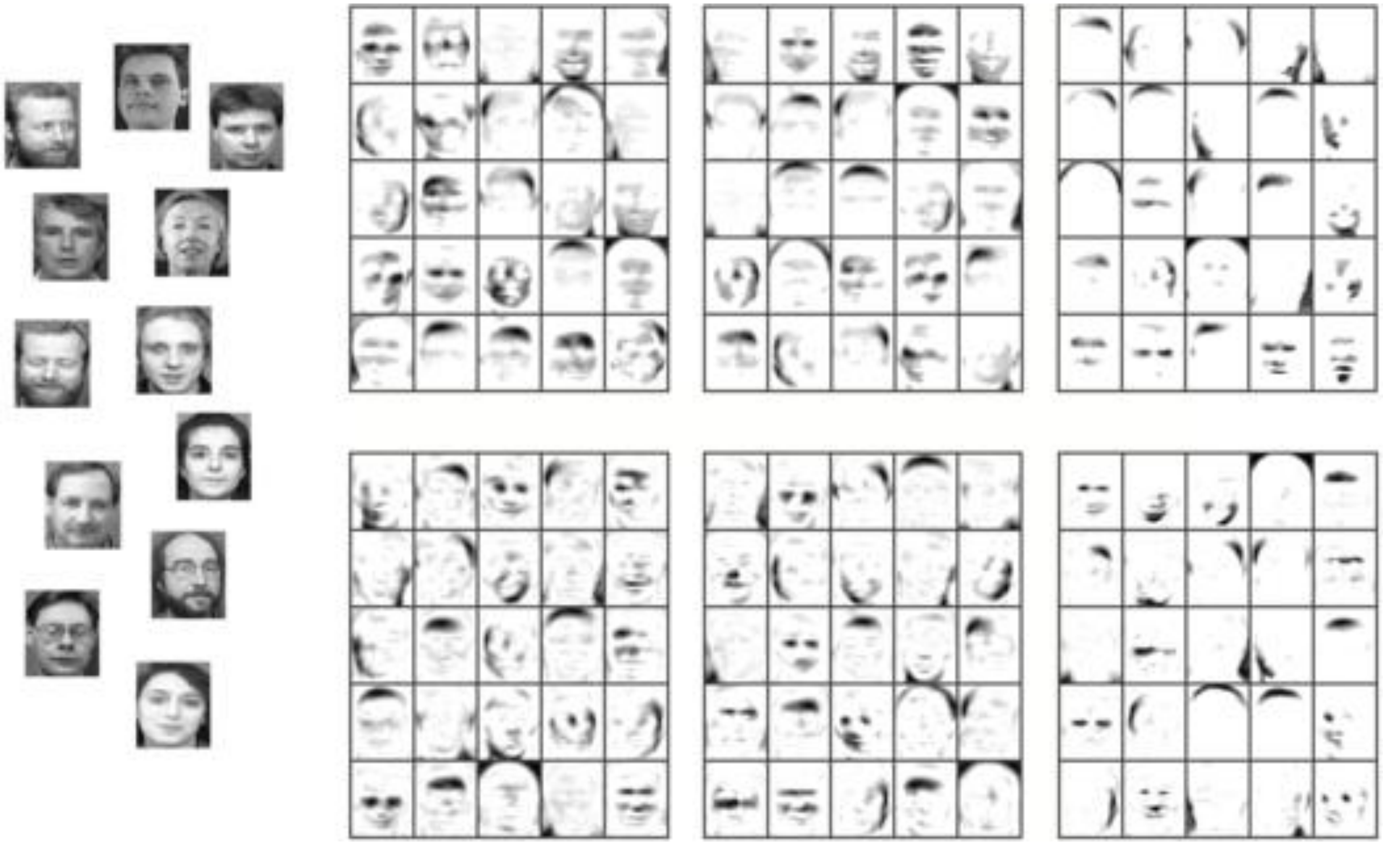
Zoomdata

ClearStory DATA

Now You See It™

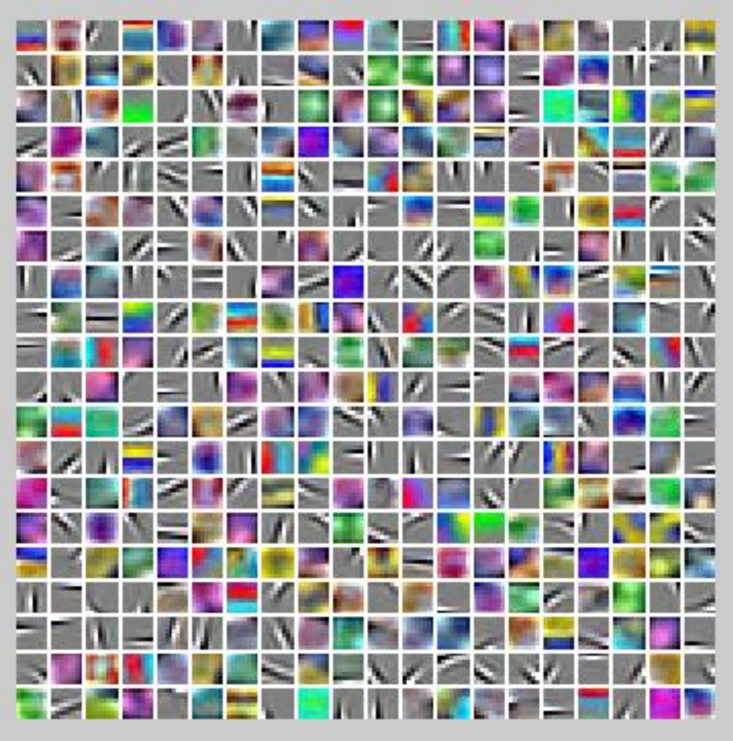
Learning representations

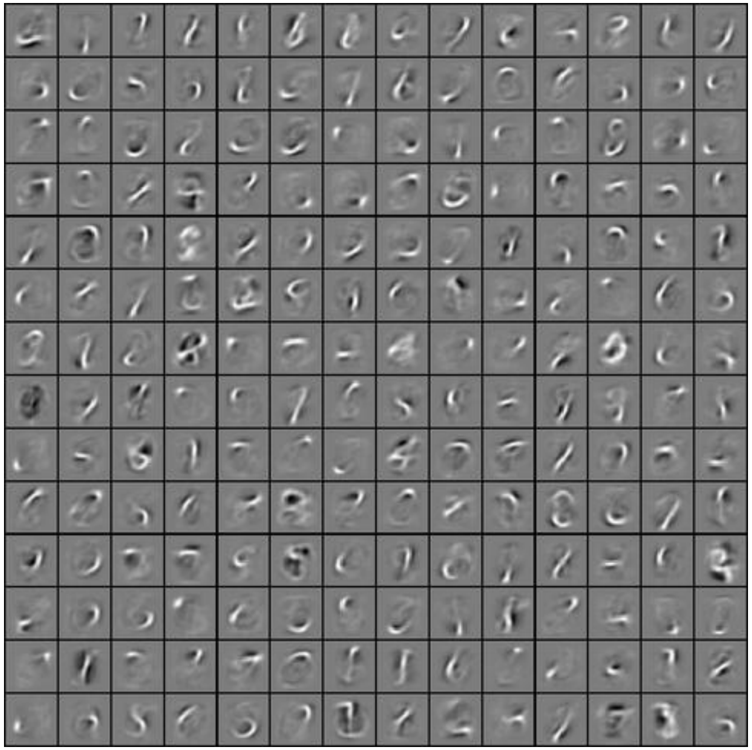
Latent semantics



Learning representations

Bag of Features / Latent semantics





How to crunch 1PB?

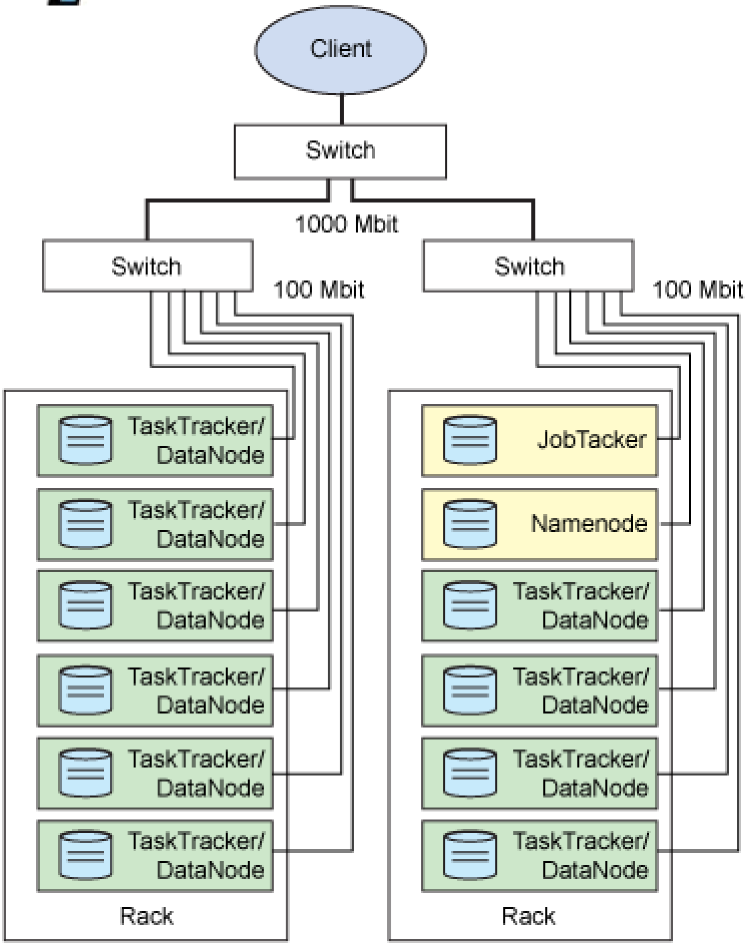
Lots of disks spinning all the time
Redundancy, since disks die
Lots of CPU cores, working all the time
Retry, since network errors happen

Design Qualities

Scalable – many servers with cores and disks
Reliable – redundant storage
Fault-tolerant – auto retry, self healing

Computation to Data

Very simple computing model → Map-Reduce
Each computing node is also a storage node
HDFS → on top of ext3, fixed 64MB file blocks
write once, read many



NoSQL

- Expressivity SQL vs. Scalability

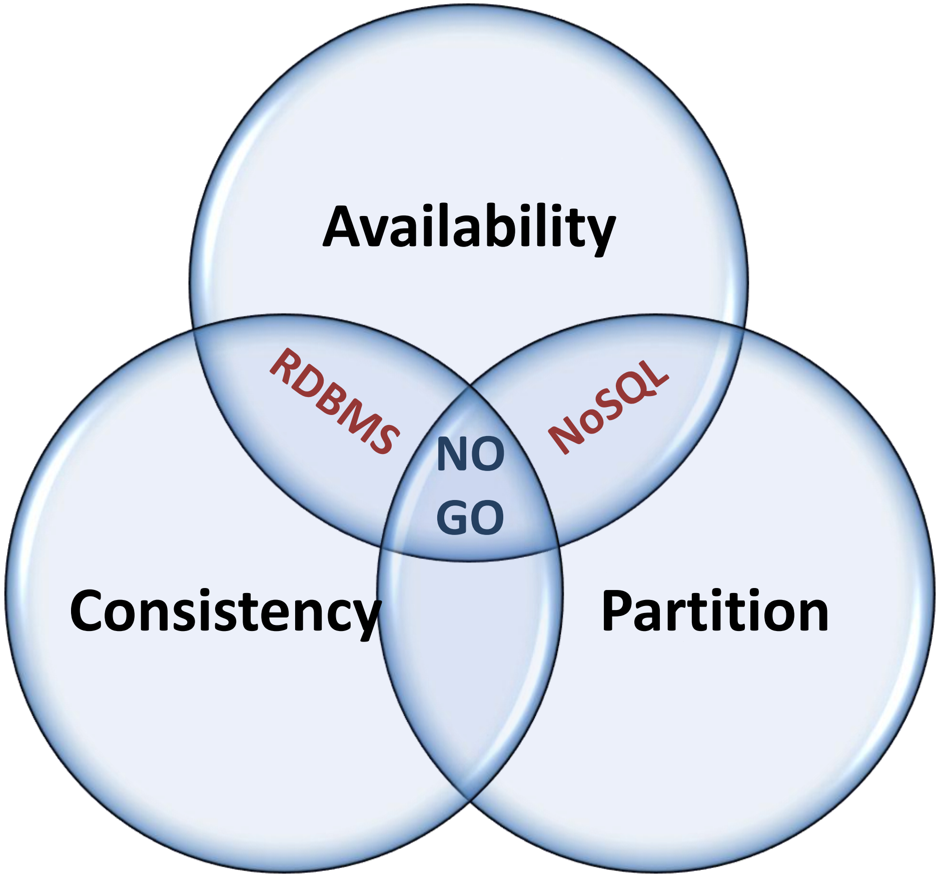
- Simpler data model (key, values)
- Simpler operations

    Scan/access per key, basic transactions (check&put)

    No joins, no SQL language
- Simple failover and scale up
- Big table, Hbase, DynamoDB, Azure, Cassandra, etc.

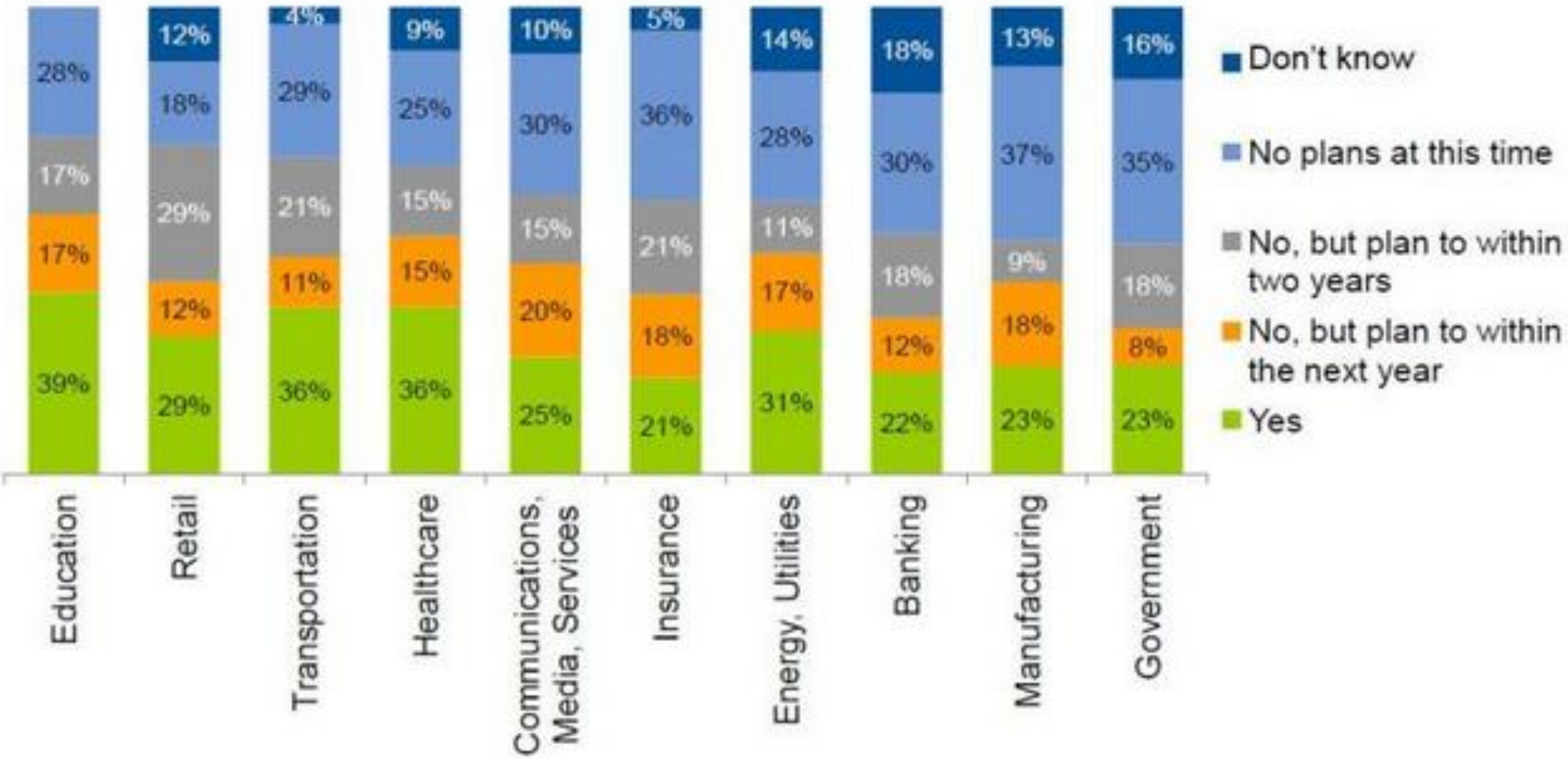
→ more work for programmers!!!

NoSQL Eventual Consistency



Big Data Investments by Industry

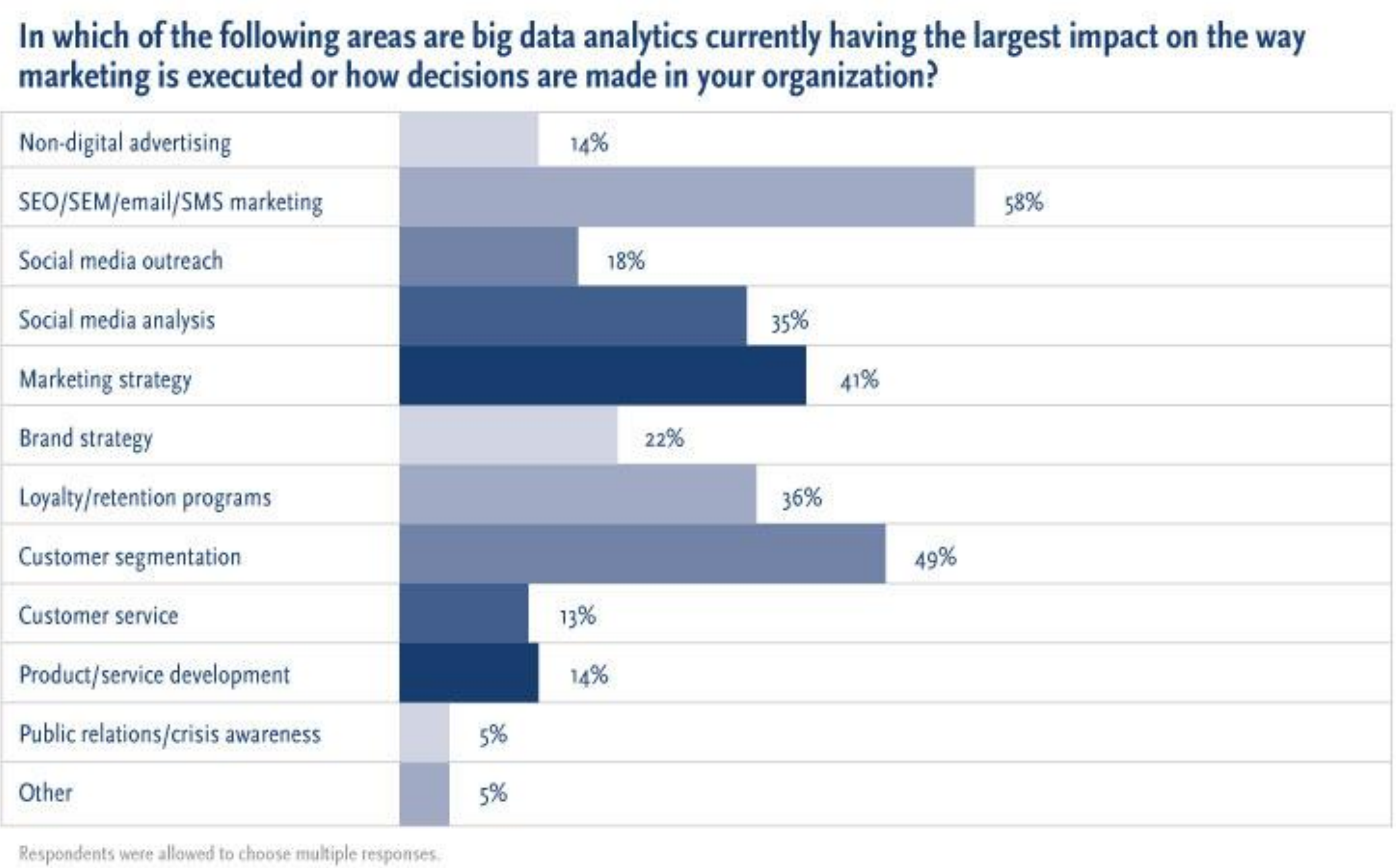
Has your organization already invested in technology specifically designed to address the big data challenge?



Source: Gartner (July 2012)

Big Data in Marketing

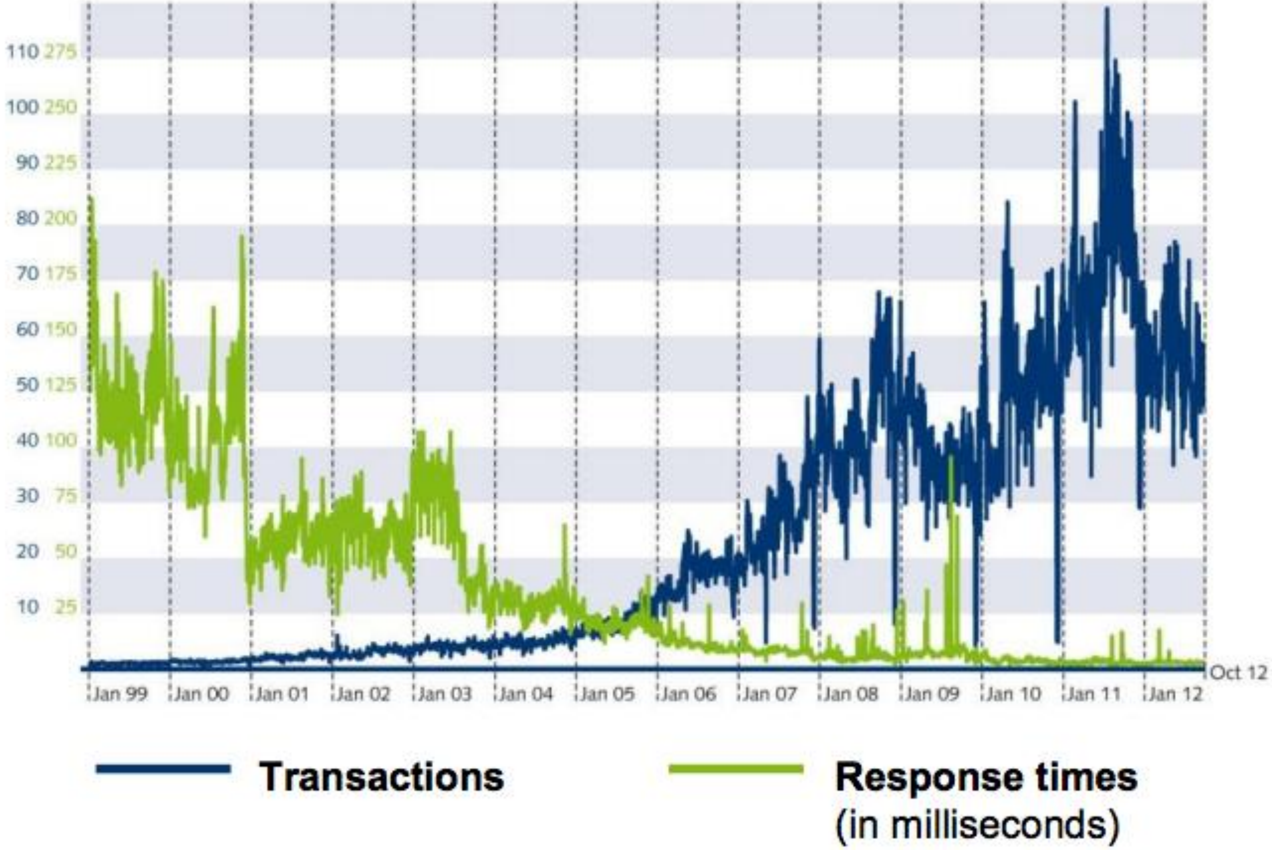
Spencer Stuart survey
171 US marketing executives



In which of the following areas are big data analytics currently having the largest impact on the way marketing is executed or how decisions are made in your organization?

Respondents were allowed to choose multiple responses.

Number of processed transactions at Eurex Exchange & response times



Transactions

Response times (in milliseconds)

Key points for Big Data

→ Integration of different data sources

→ Continuous prediction models from historical + real time data

→ (Semi)automatic knowledge discovery is now possible

→ Computing infrastructure

→ Multidisciplinary teams (in house, w/ academy)

Multidisciplinary teams

The world of computing is flat, and anyone can do it. What will distinguish us from the rest of the world is **our ability to do it better and to exploit new architectures** we develop before those architectures become ubiquitous.

There is a clear and urgent need for a **new, modern approach to educating and training the next generation of researchers** in high performance computing specifically, and in modeling and simulation generally, for scientific discovery and engineering innovation.

Inadequate **education and training of the next generation of computational scientists** threatens global as well as U.S. growth of SBE&S […] unless we prepare researchers to develop and use the next generation of algorithms and computer architectures, we will not be able to exploit their game-changing capabilities.

There are clear and urgent opportunities for **industry-driven partnerships with universities** and national laboratories to hardwire scientific discovery and engineering innovation through SBE&S.

www.wtec.org/sbes 2009

Challenges → Focus at UIS-SC3

- Identify data assets within your organization
- Identify diversity and sources of data
- Assess data links to support bandwidth
- Assess data quality and private data
- Understand gap data collection-interpretation
- Building specialized skills
- Building multidisciplinary teams

Universidad Industrial de Santander

Supercomputación y Cálculo Científico

Guane I

60 TFlops

GPU Powered

Launched 4/2012

Doubling soon

Gas Oil Energy

Big Data

Agua Recursos

Medio Ambiente

Astrofísica

Biología

http://sc3.uis.edu.co