
To boldly go where no
man has gone before

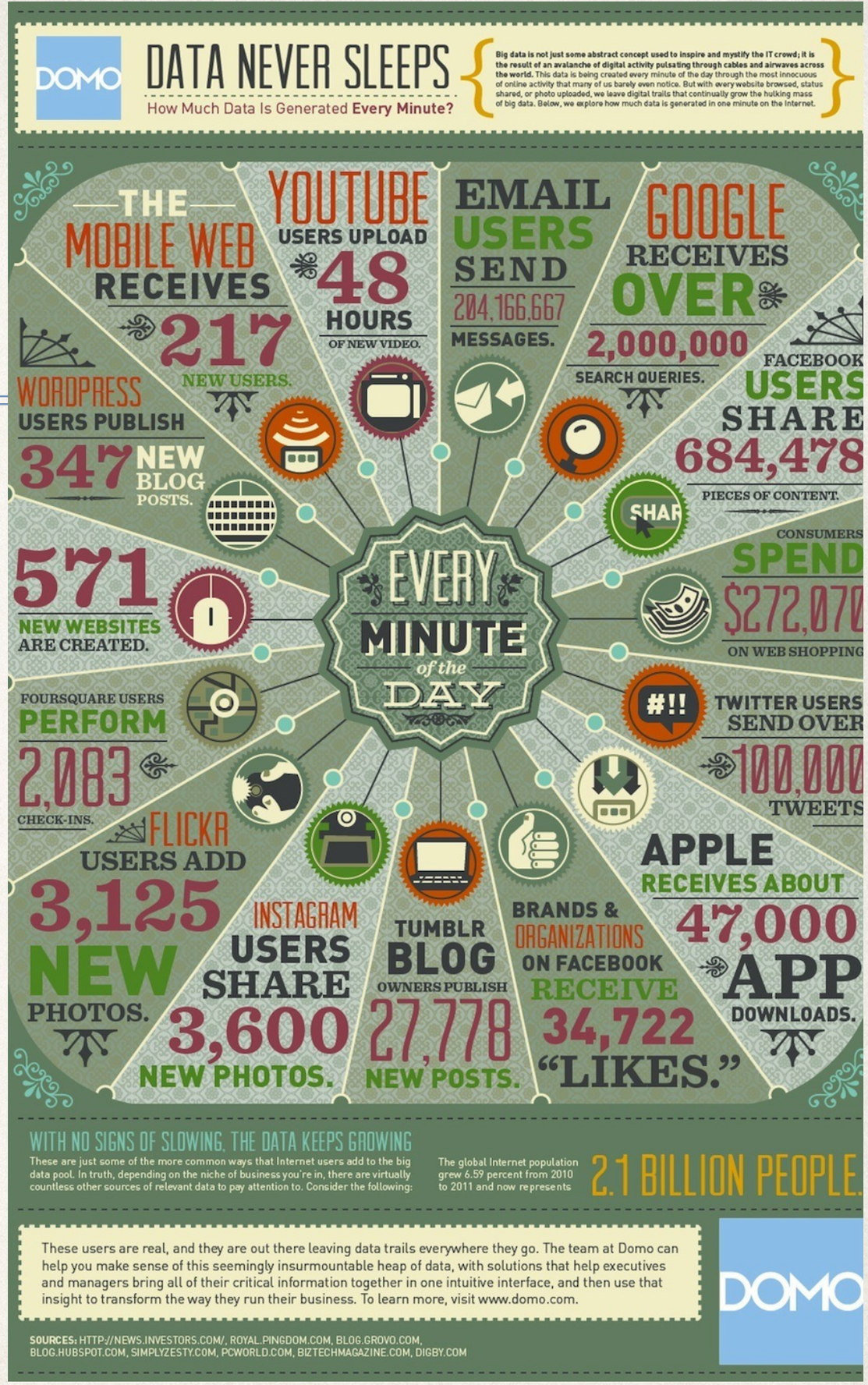
Big Data and the Data Revolution

Chris Stephens
C3-Centro de Ciencias de la Complejidad y Instituto de Ciencias Nucleares, UNAM

Primer Escuela de Verano de Modelación para la Sostentabilidad
LANCIS 24/06/2015

There's been a data revolution...

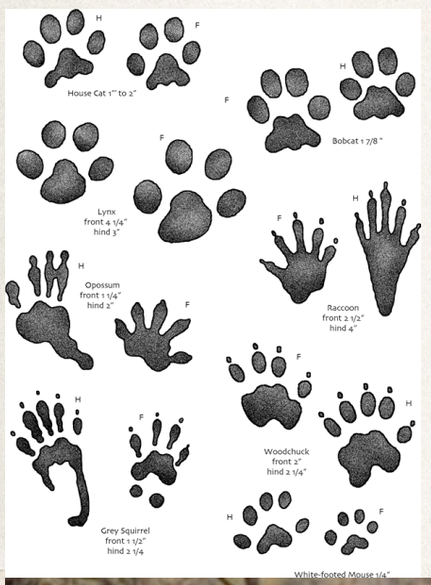
But just what's revolutionary?



Camille Kubie, Estuary, 2014

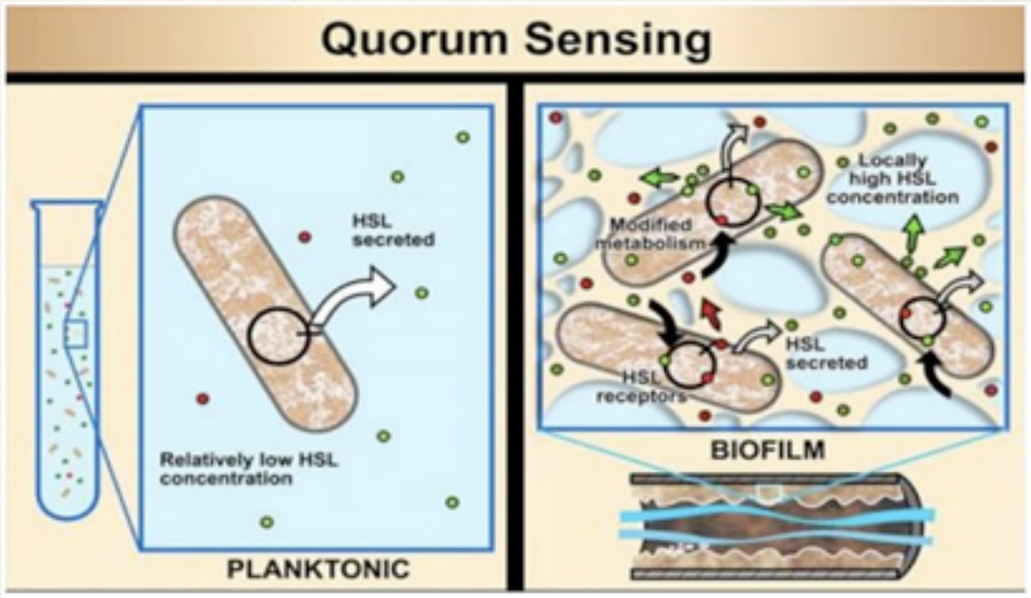
What's revolutionary?

Data types? No.



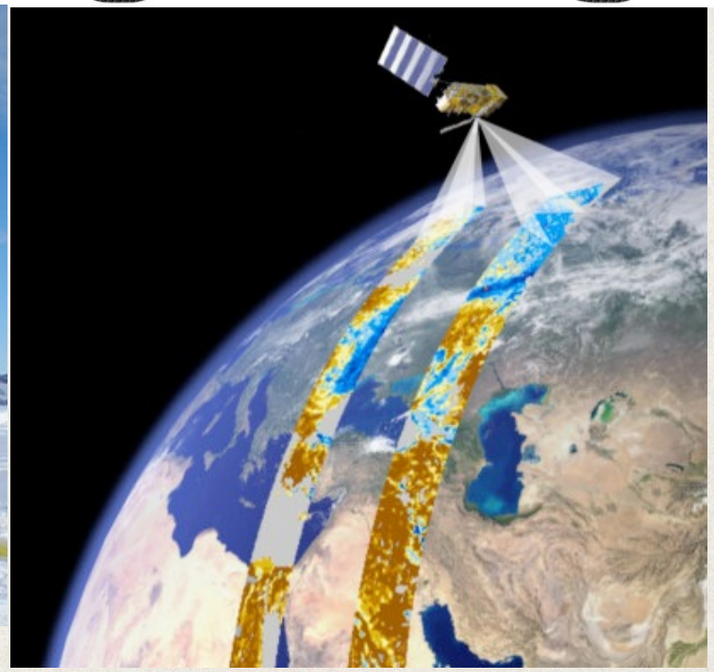
Raw data:
Chemical
Electromagnetic
Acoustic...

as functions of space and time tell us what is going on in the world.



We use data about ***events*** to take ***decisions.***

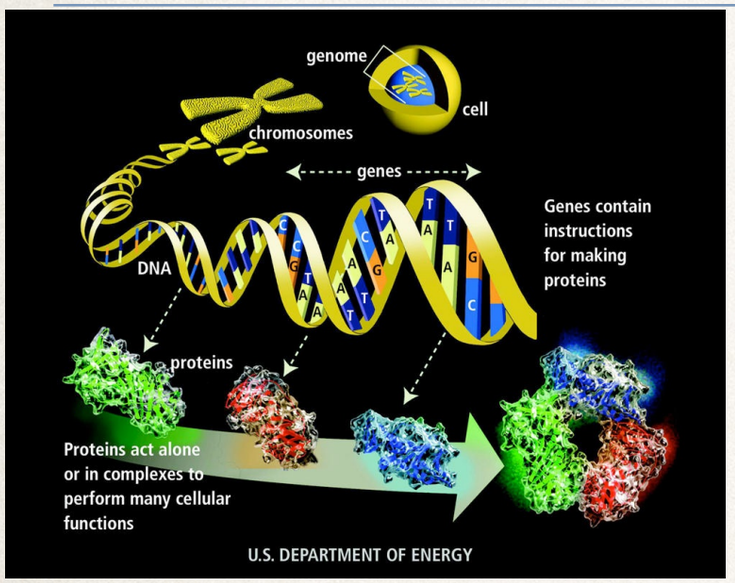
Data sensing? Yes.

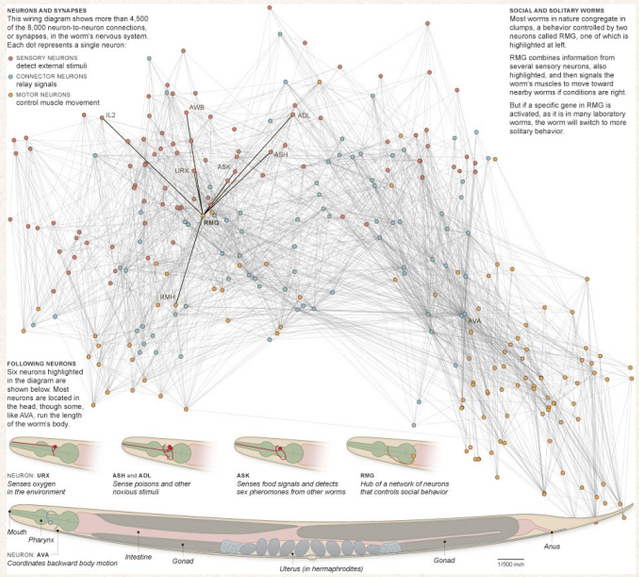


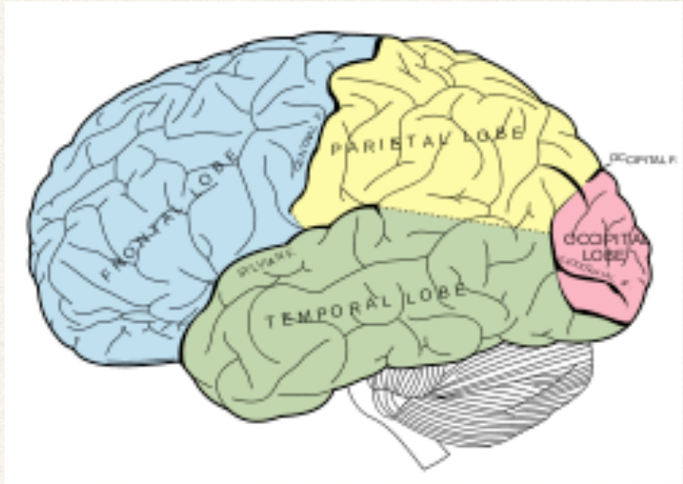
Data storage and processing? Yes.

Human brain

10-100 Terrabytes





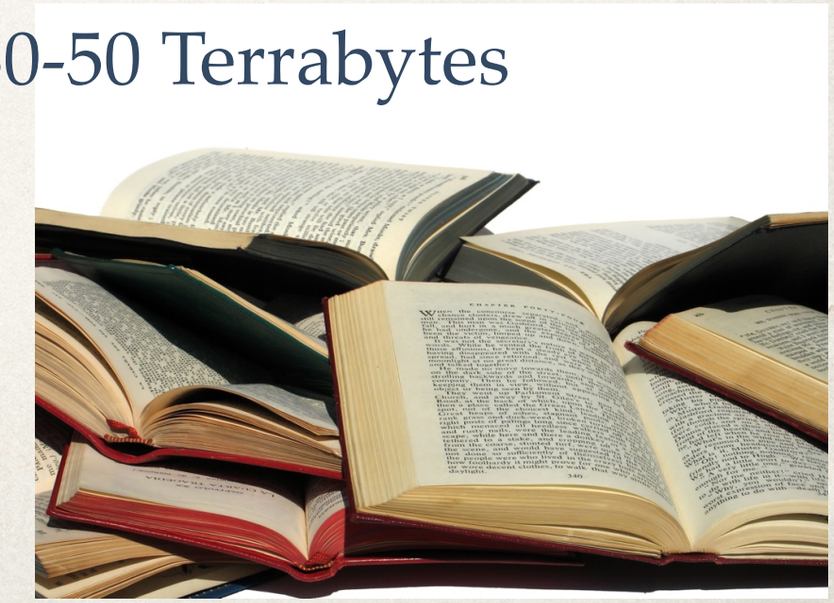


Genomes 1kB - 1.5 GB

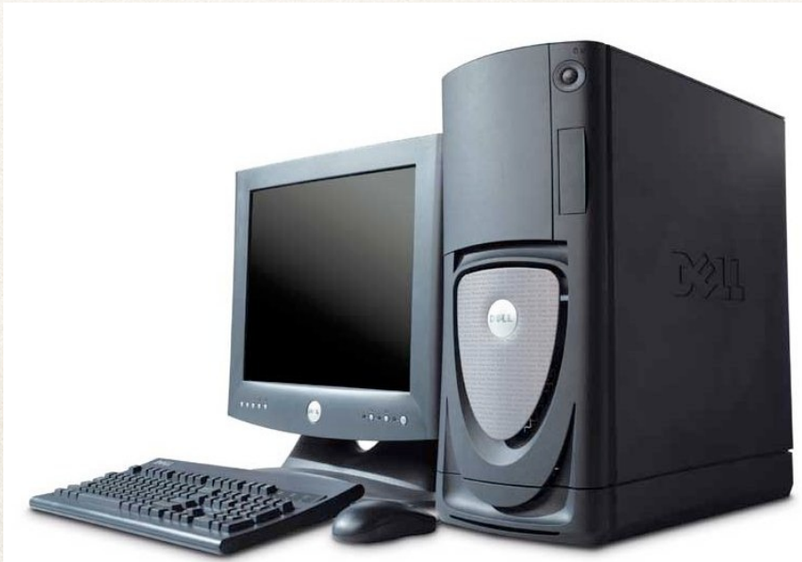
Worm neural network 0.3MB

All the books in the world

30-50 Terrabytes



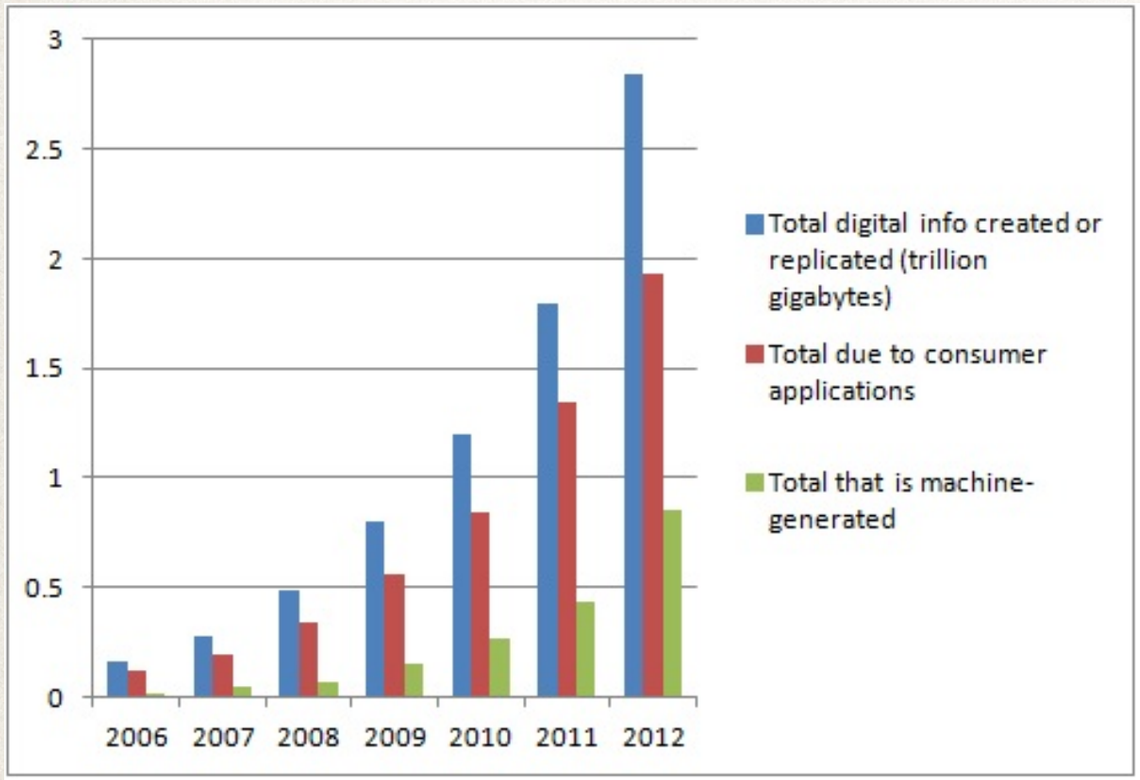
In electronic form 1 zettabyte



Raw data is processed and stored

What's revolutionary?

Data storage and processing? Yes.



Source: IDC

Data growth by type

Some of the world's biggest databases

28 million books
> 20 terrabytes text data

AT&T

323 terrabytes
1.9 trillion call records

amazon.com

> 42 terrabytes

World Meteorological Organization
Global Atmosphere Watch
Organización Meteorológica Mundial
Organisation Météorologique Mondiale
Всемирная Метеорологическая Организация

220 terabytes of web data
6 petabytes of additional data

What's revolutionary?

Data storage and processing? Yes.

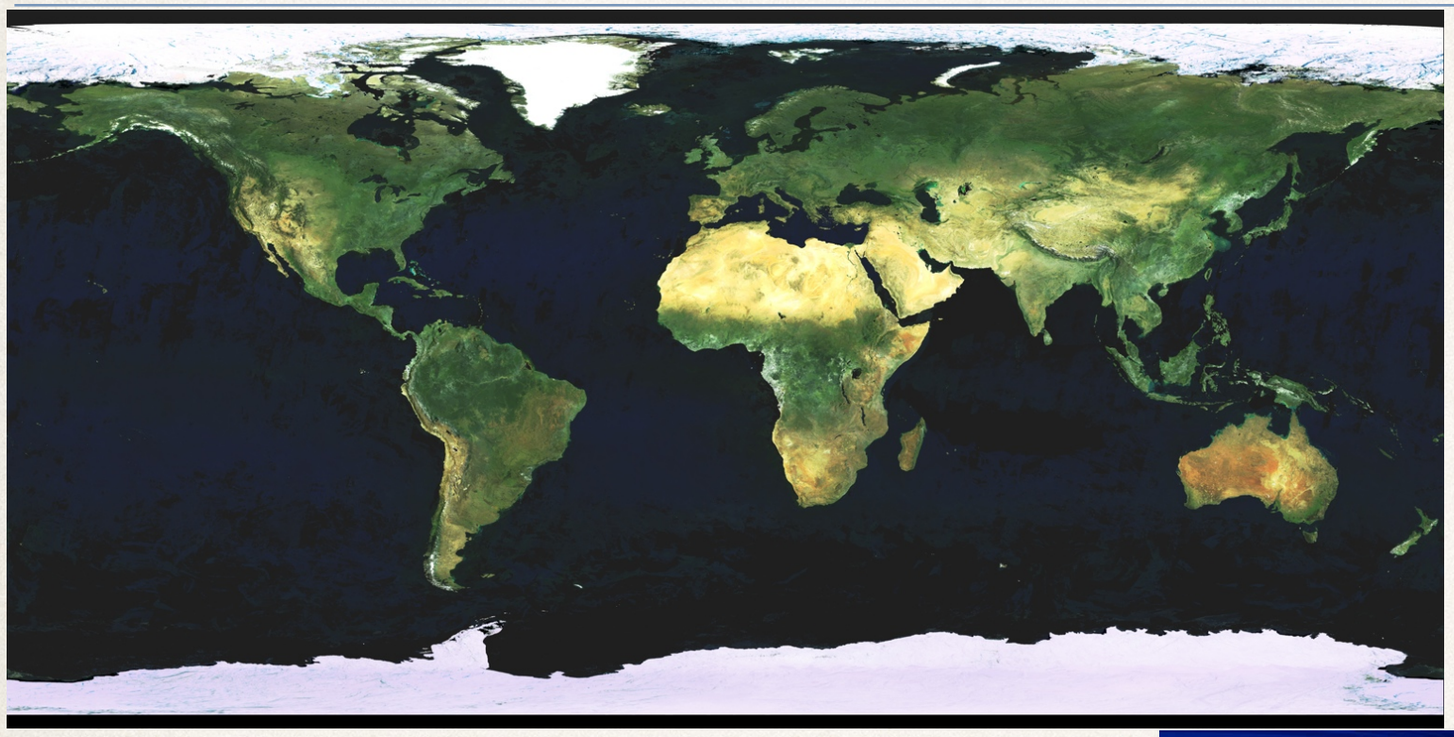
We can now track and record what is happening in the world like never before.

For example, a financial market where...

every transaction that occurs is processed (a summary of relevant information is determined) and then electronically stored.

What's revolutionary?

Data connectivity? Yes.

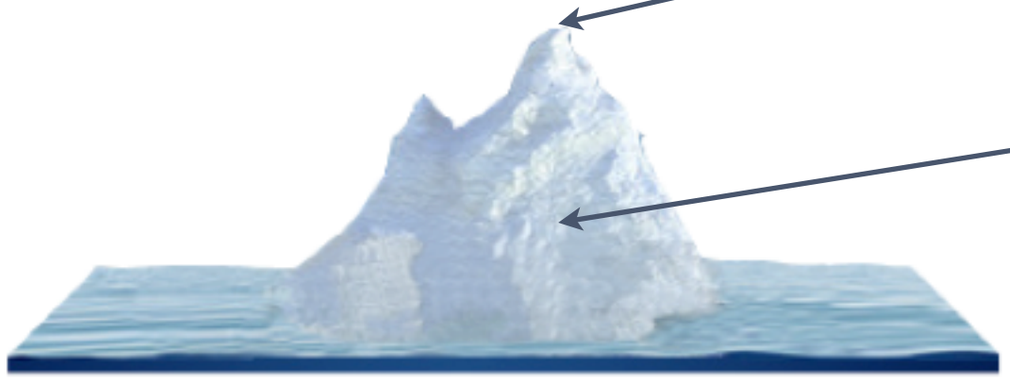


Then

Real space --> cyberspace

Now

Data connectivity? Yes.
But just how connected are we?

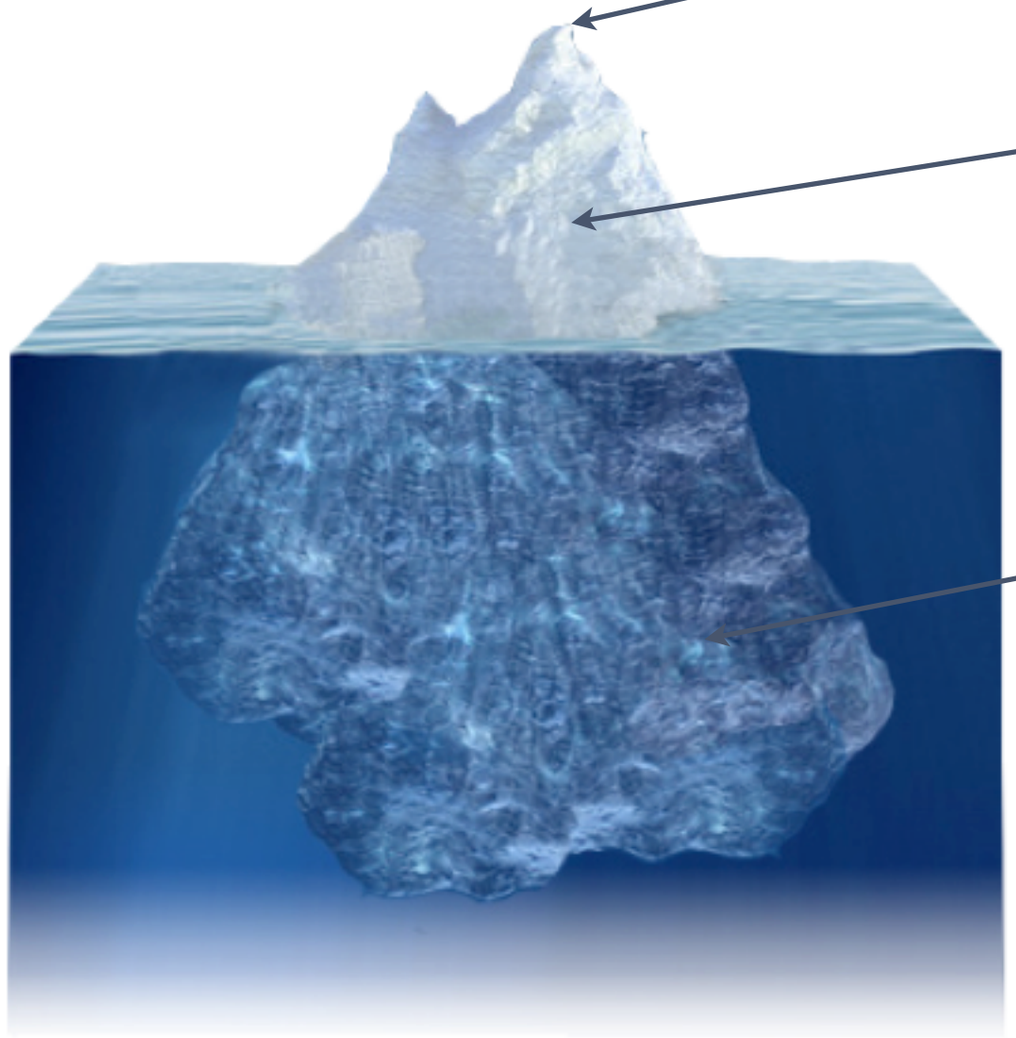


My data:
a snowflake

The data we have access to: the tip of the iceberg

Data connectivity? Yes. But is it that great?

Public versus private data



My data:
a snowflake

The data we have access to: the tip of the iceberg

The data we don't have access to!

What's revolutionary?

Data search capacity? Yes.

Pre-writing:
The first "search engine".
Find the person that knows
what you want to know.

Post writing:
The second "search engine".
Find the text that contains
what you want to know.

Post www:
The third search engine.
A machine does it

Data search capacity? Yes. But just how good is it?

Easy

late etruscan pottery

Web Images More Search tools

About 320,000 results (0.24 seconds)

Etruscan Pottery - The Mysterious **Etruscans**
www.mysterious**etruscans**.com/art/**pottery**.html
Jan 1, 2006 – Most **pottery** found at **Etruscan** burial sites follows very closely on the ...
The shapes and motifs of the mid- to **late** 7th century are derived largely ...

Etruscan Art - Metropolitan Museum of Art
www.metmuseum.org/toah/hd/etru/hd_etru.htm
Greek **potters** and their works influenced the development of **Etruscan** fine ... source of evidence for artistic achievement during the **Late** Classical and Hellenistic ...

Etruscan art - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Etruscan**_art
The **Etruscans** invented the custom of placing figures on the lid which **later** influenced the Romans to do the same. The Hellenistic period funerary urns were ...

It's fast but not clever!

Web Images News More Search tools

About 224,000,000 results (0.47 seconds)

List of **rivers** by length - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/List_of_**rivers**_by_length
As a result, the length measurements of many **rivers** are only approximations. In particular, there has **long** been disagreement as to whether the Nile or the ...
Definition of length - List of rivers longer than 1000 km

The Longest **Rivers** in the World - Social Studies for Kids
www.socialstudiesforkids.com/articles/.../**long**est**rivers**intheworld.htm
Did you know that the longest **river** in the world is the Nile? Egypt's greatest **river** is 4,135 miles **long**! In fact, Africa has two of the ten longest **rivers**. The Congo ...

Lengths of major **rivers**, from USGS Water-Science School
ga.water.usgs.gov/edu/**rivers**ofworld.html
Jan 10, 2013 – Ever wonder what rivers are the longest? Look at the graphic below to see our short list of **long rivers**. (It's not so easy to define how long a river ...

Top 9 Longest **Rivers** in the World - UNP
www.unp.me › Chit-Chat › Gapp-Shapp
Aug 23, 2010 – This **long river** can be divided into Ob River and The Irtysh is the major tributary of the Ob. There're several other tributaries for Ob. The water in ...

Top Ten Longest **Rivers** in the World List - Fun Science Facts for Kids
www.sciencekids.co.nz/sciencefacts/topten/**long**est**rivers**.html
4 days ago – Longest Rivers in the World. The world features some amazingly **long rivers** but which are the longest? Check out our list of the top ten longest ...

What are three very **long rivers** - WikiAnswers
wiki.answers.com › ... › Geography › Bodies of Water › Lakes and Rivers
Is this a trick question? Because it can range from 1000 years ago to 100 billion years ago. Which very **long river** in Brazil has its mouth at the Atlantic ocean?

Hard

early victorian educational reforms

Web Images Videos More Search tools

About 4,220,000 results (0.17 seconds)

Towards **Victoria** as a Learning Community
www.education.vic.gov.au › Our Department › Strategic Directions
Mar 22, 2013 – Department of **Education** and **Early** Childhood Development ...
Victoria's Plan for **School** Funding **Reform** · Towards **Victoria** as a Learning ...

Education in **Victoria** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Education**_in_**Victoria**
Education in **Victoria**, Australia is supervised by the Department of **Education** .... responsible for the **reform** policy development process and the **early** stages of its ...

Victorian Legislation: a Timeline - The **Victorian** Web
www.**victorian**web.org/history/legistl.html
Dec 20, 2006 – The first **Education** Act did not reach the Statute Books until 1870. 1834 Poor Law Amendment Act. Following the 1832 **Reform** Act, the PLAA ...

Victoria throws **education reforms** into disarray - The Age
www.theage.com.au › National
Feb 24, 2013 – **Victoria** throws **education reforms** into disarray ... system could be phased in as **early** as next year - and "no school would be worse off".

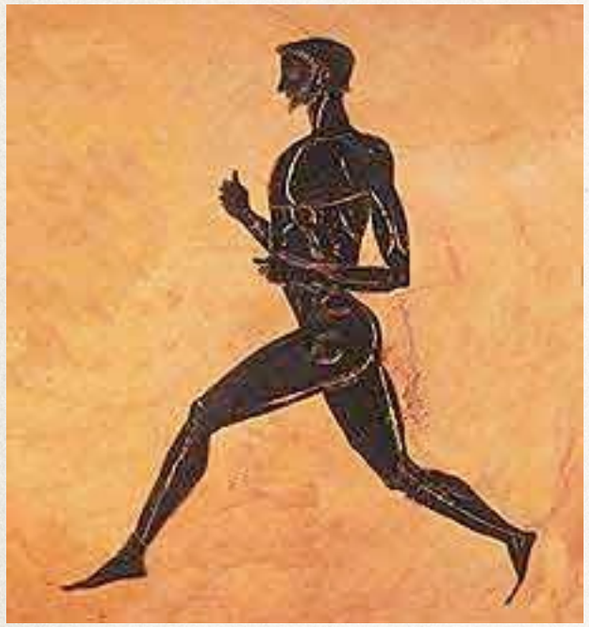
§25. Public **School reform**. XIV. Education. Vol. 14. The **Victorian** ...
www.bartleby.com › ... › The Victorian Age, Part Two › Education
The first steps in a real **reform** of courses of instruction among schools of this type were taken by the **early Victorian** foundations, chiefly proprietary, such as ...

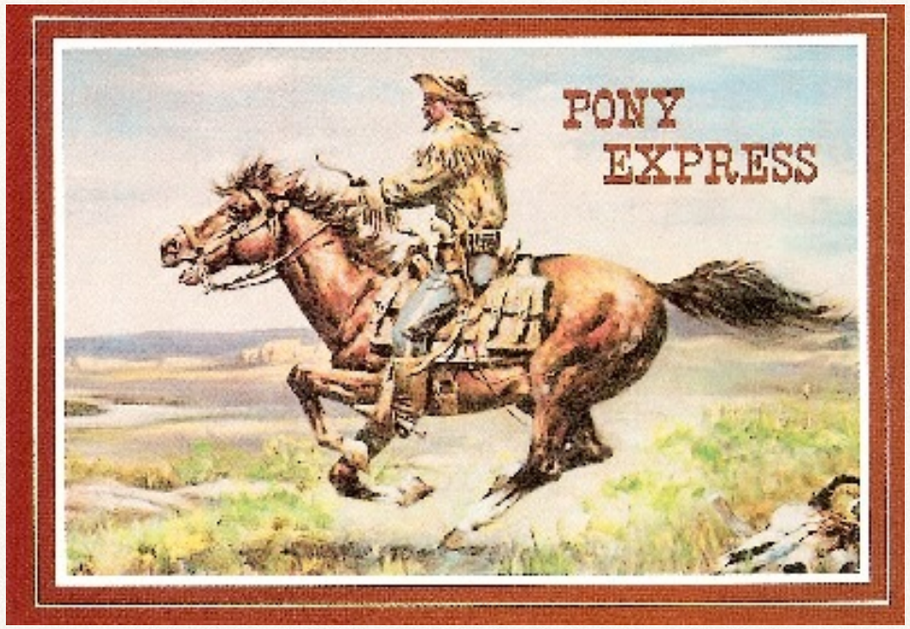
Victorian education reform: Education Act 1870
www.architecture.com/.../**Education**InAModernWor... - United Kingdom
Victorian education reform: Education Act 1870. Perspective view of Harper Street School, New Kent Road, London, 1885. Print Designer: Robert W Edis ...

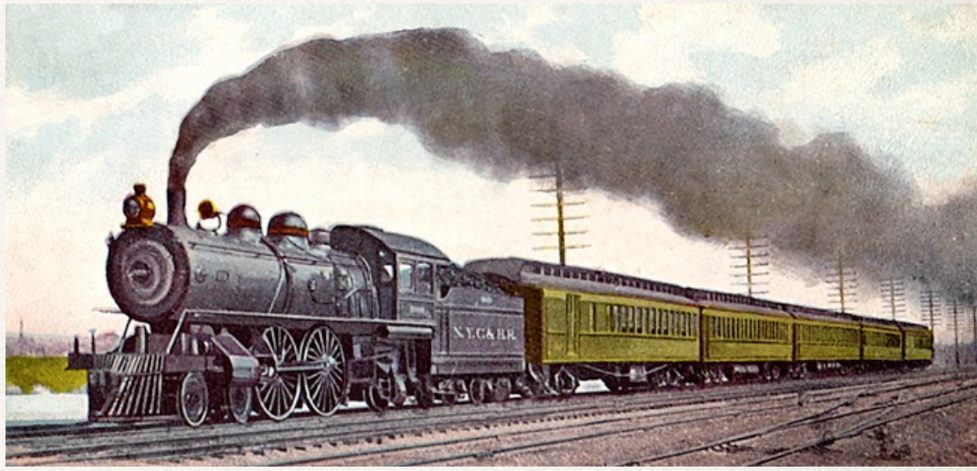
Humans are wonderful at semantics, machines aren't

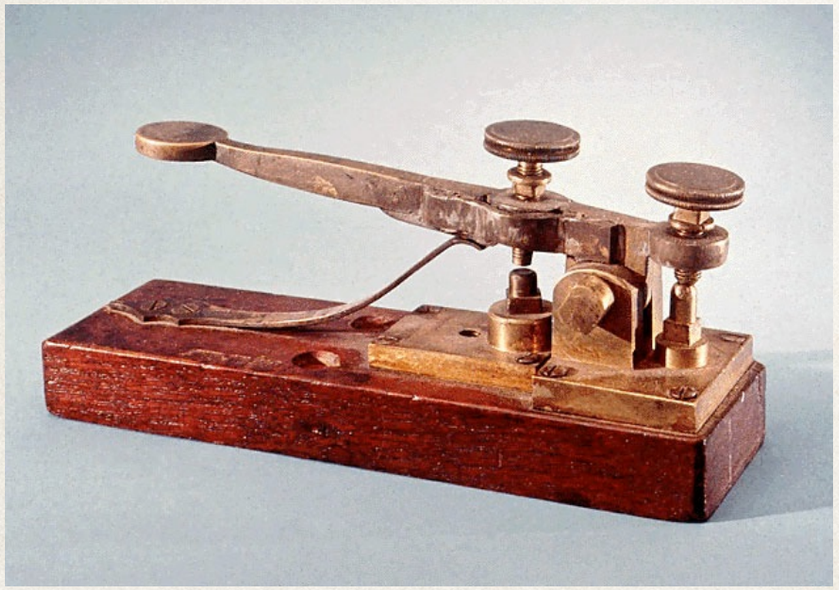
What's revolutionary?

Data communication speed? Yes, but not like you imagine?









What's revolutionary?

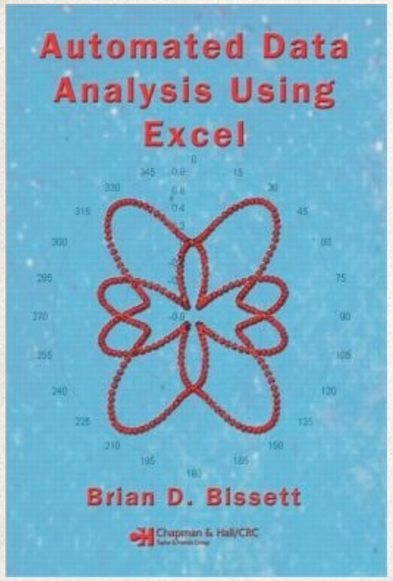
Data purpose? No.

- We have always used data to take decisions
- There has always been an "evolutionary pressure" in favour of those who can record, process, store and analyse more efficiently and more precisely data in order to make better decisions

Data analysis? Yes.

Instinct: Hard-wired data analysis

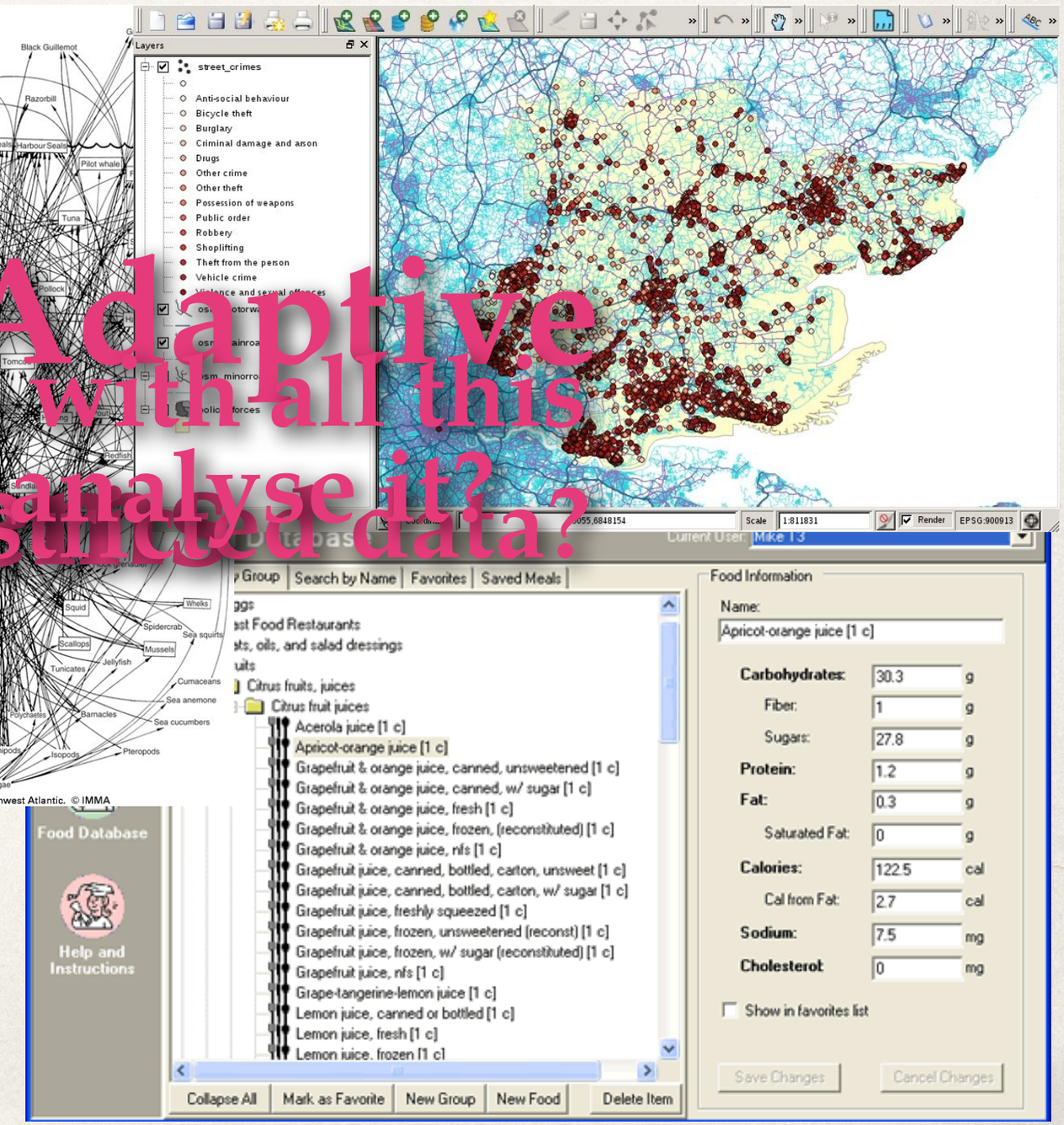
Learning: adaptive data analysis

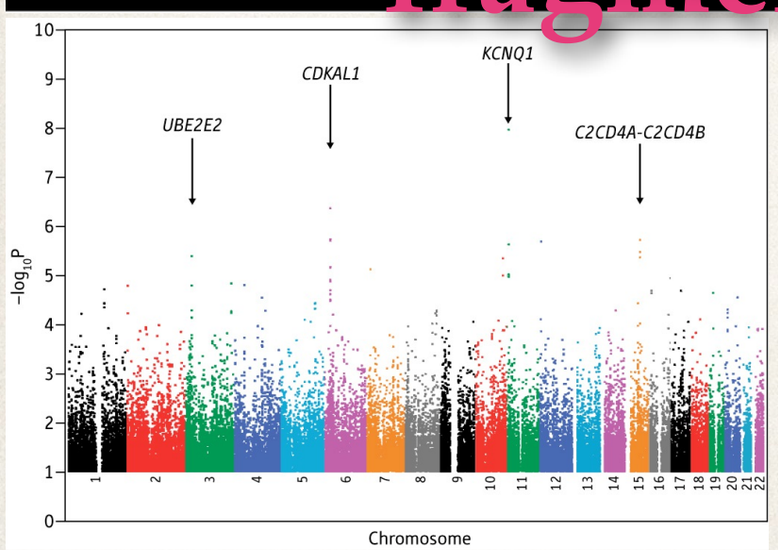


Automated data analysis - data mining.
Can consider large volumes of data with large numbers of variables

What does all the data from the data revolution represent?

A simplified food web for the Northwest Atlantic. © IMMA





Complex Adaptive

What do we do with all this

How do we analyse it?

fragmented, siloed data?

Data mining

- "... the exploration and analysis of data in order to discover patterns, correlations and other regularities."
- There are two main data mining tasks
  - Predicting - "pattern" identification
    - Establishes "causal" statistical relations
  - Profiling - "pattern" description
    - Identifies what are the key drivers associated with a pattern
    - There are three main requirements
    - Data, data processors, inference algorithms

From the science of yesterday to the science of tomorrow...

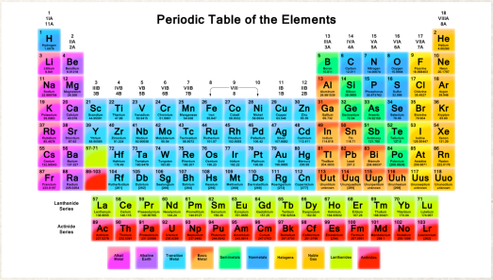
How we do science in a nutshell...

- **The Scientific method:** Systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses

  - **Phenomenology** - a body of knowledge that relates empirical observations of phenomena to each other, in a way that is *consistent* with fundamental theory, but is not directly derived from theory.

  - **Taxonomy** - the practice and science of classification. A classification of things or concepts, as well as to the principles underlying such a classification.

    - Examples: Medicine, astronomy, chemistry, biology, physics,...



  - **Scientific law** - when a particular phenomenon always occurs if certain conditions are present

The worldview of the last 3 centuries:

Universality

We're all equal under the law

But in physics and chemistry...
once you've seen one perfect gas
there's really not a lot to say
In general, you don't need
At any rate, we've seen them places
you have seen them
that much data

In Complex Adaptive Systems however...

t you

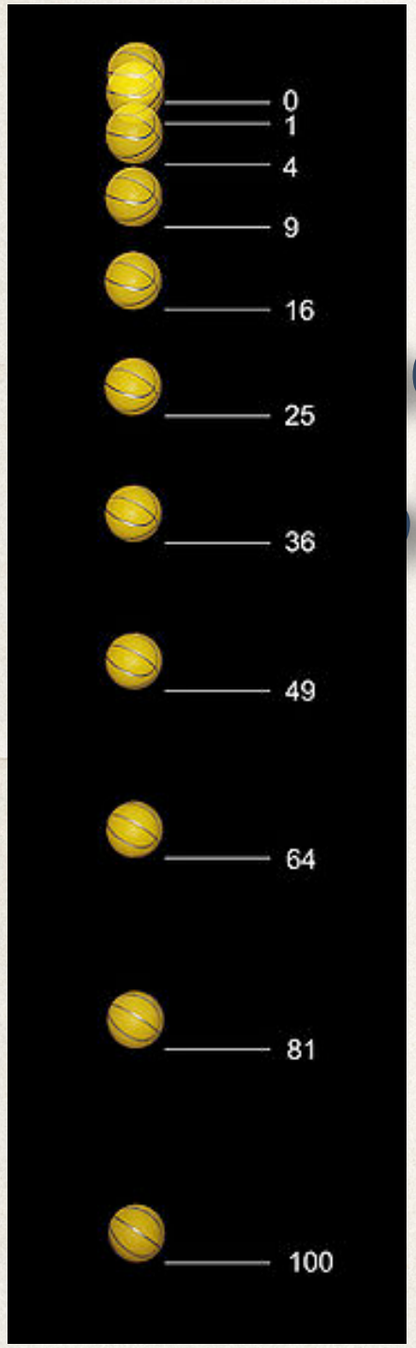
Imagine what you can say about a city

versus

a crystal as big as a city!

The data revolution is revolutionising our ability to study the immensely rich phenomenology of complex systems and construct more appropriate taxonomies

Another conclusion



Mechanistic

Adaptive

The *evolution* of function

e difference between complex an

is the revolution that allowed

le systems is the difference betw

Systems obey the same laws

But it's not a "slave" to them

of physics as it pays.

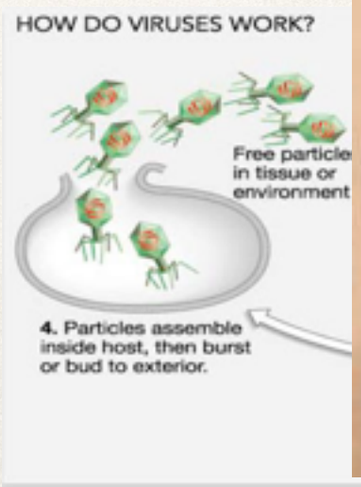
"being" and "doing"

Complexity is a consequence

of that revolution.


BALLOT


joe-ks


HOW DO VIRUSES WORK?
Free particle
in tissue or
environment
4. Particles assemble
inside host, then burst
or bud to exterior.


There are good decisions
and there are bad decisions

Modeling complexity

To make a mathematical model of a dynamical system...

we need a space of states

and update rules that tell us how to get from one state to another

Modeling Education as a Complex Adaptive System

- **Predicting academic success**
  - CUAED distance learning students
  - Success = graduated or covered 100% of course material in 10 semesters or less
    - From 2005-2010
    - 3,697 students in the sample, 370 “graduated”
    - Terminal efficiency 10%
    - 24 predictor variables
    - Calculate P(Success | predictors)

Preliminary analysis with: M. en I. J. Gerardo Moreno Salinas; Clase de Mineria de datos

Modeling Education as a Complex Adaptive System

Variable	Epsilon	Nx	Nxc	N	Nc	Pc	Pxc	Descripción
Edad ingreso	0.516007	418	45	3697	370	10.01%	10.77%	Edad de ingreso [< 20]: 1
Edad ingreso	-2.621298	860	63	3697	370	10.01%	7.33%	Edad de ingreso [20, 24]: 2
Edad ingreso	-1.528292	668	55	3697	370	10.01%	8.23%	Edad de ingreso [24, 28]: 3
Edad ingreso	-1.730235	529	41	3697	370	10.01%	7.75%	Edad de ingreso [28, 32]: 4
Edad ingreso	-0.467363	408	38	3697	370	10.01%	9.31%	Edad de ingreso [32, 36]: 5
Edad ingreso	3.252179	285	45	3697	370	10.01%	15.79%	Edad de ingreso [36, 40]: 6
Edad ingreso	0.971195	226	27	3697	370	10.01%	11.95%	Edad de ingreso [40, 44]: 7
Edad ingreso	3.014808	164	28	3697	370	10.01%	17.07%	Edad de ingreso [44, 48]: 8
Edad ingreso	2.232843	80	14	3697	370	10.01%	17.50%	Edad de ingreso [48, 52]: 9
Edad ingreso	3.511756	59	14	3697	370	10.01%	23.73%	Edad de ingreso [> 53]: 10
Sexo	4.043805	2049	260	3697	370	10.01%	12.69%	Sexo, Femenino: 1
Sexo	-4.509025	1648	110	3697	370	10.01%	6.67%	Sexo, Masculino: 2

Statistically significant effect on success from both age and gender

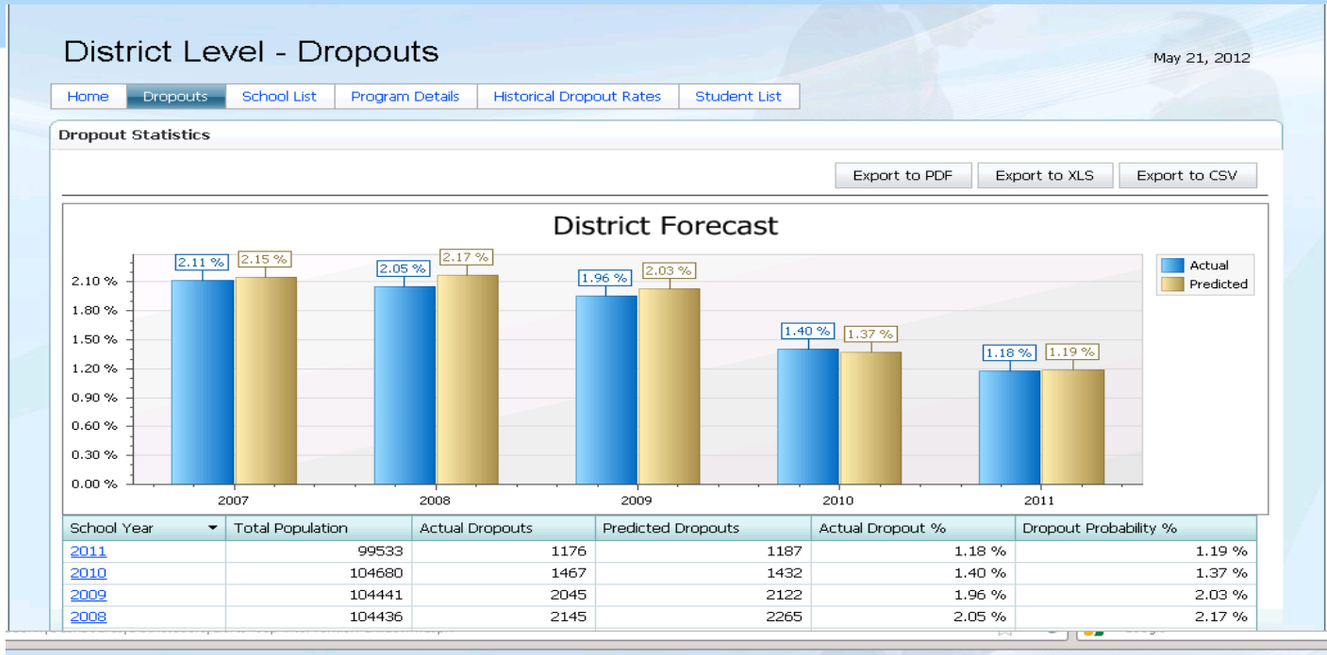
Modeling Education as a Complex Adaptive System

Variable	Epsilon	Nx	Nxc	N	Nc	Pc	Pxc	Descripción
Sostén Económico	1.484220	267	34	3697	370	10.01%	12.73%	Sostén Económico, Cónyuge: 2
Sostén Económico	-0.907795	176	14	3697	370	10.01%	7.95%	Sostén Económico, Madre: 3
Sostén Económico	1.650505	286	37	3697	370	10.01%	12.94%	Sostén Económico, No contesto: 6
Sostén Económico	0.428081	60	7	3697	370	10.01%	11.67%	Sostén Económico, Otro: 7
Sostén Económico	-1.110523	396	33	3697	370	10.01%	8.33%	Sostén Económico, Padre: 8
Sostén Económico	-3.092583	1498	114	3697	370	10.01%	7.61%	Sostén Económico, Tú mismo:9
Escolaridad Máx Madre	-2.063992	416	29	3697	370	10.01%	6.97%	Escolaridad Máx Madre, Carrera técnica: 1
Escolaridad Máx Madre	-2.933960	177	6	3697	370	10.01%	3.39%	Escolaridad Máx Madre, Licenciatura: 2
Escolaridad Máx Madre	-2.828409	269	13	3697	370	10.01%	4.83%	Escolaridad Máx Madre, Media superior o normal: 3
Escolaridad Máx Madre	-1.445546	36	1	3697	370	10.01%	2.78%	Escolaridad Máx Madre, Posgrado: 7
Escolaridad Máx Madre	0.231696	1017	104	3697	370	10.01%	10.23%	Escolaridad Máx Madre, Primaria: 8
Escolaridad Máx Madre	-0.444536	520	49	3697	370	10.01%	9.42%	Escolaridad Máx Madre, Secundaria: 9
Escolaridad Máx Madre	2.441668	237	35	3697	370	10.01%	14.77%	Escolaridad Máx Madre, Sin instrucción: 10

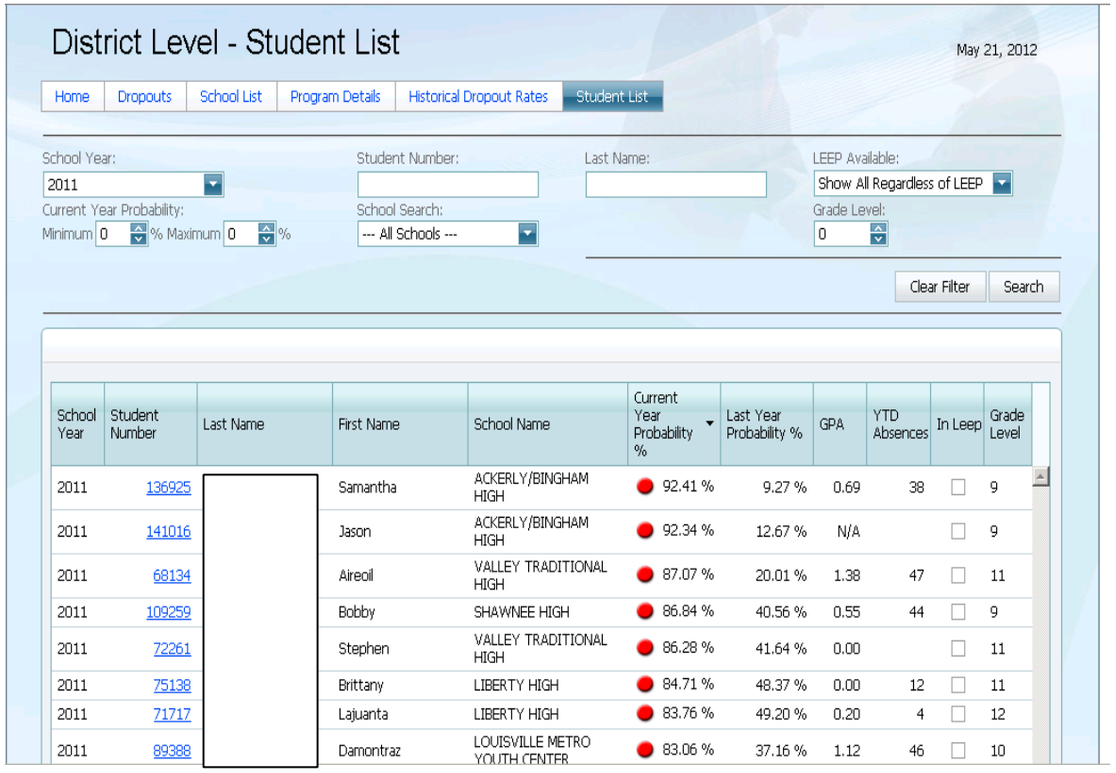
Statistically significant effect on success from economic support and parental scholastic level

How to model a complex world?

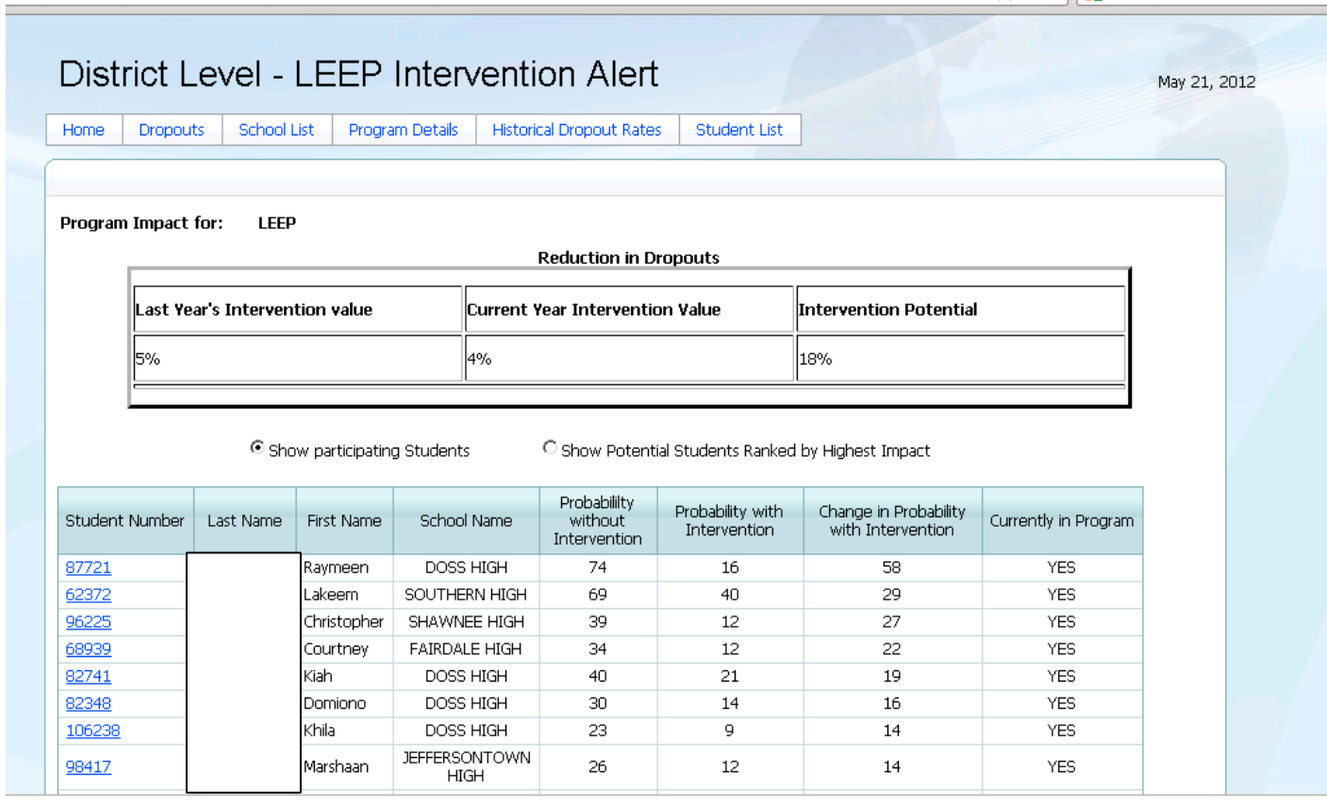
Predicting the dynamics of high school dropouts



School Year	Total Population	Actual Dropouts	Predicted Dropouts	Actual Dropout %	Dropout Probability %
2011	99533	1176	1187	1.18 %	1.19 %
2010	104680	1467	1432	1.40 %	1.37 %
2009	104441	2045	2122	1.96 %	2.03 %
2008	104436	2145	2265	2.05 %	2.17 %



School Year	Student Number	Last Name	First Name	School Name	Current Year Probability %	Last Year Probability %	GPA	YTD Absences	In Leep	Grade Level
2011	136925	[redacted]	Samantha	ACKERLY/BINGHAM HIGH	92.41 %	9.27 %	0.69	38		9
2011	141016	[redacted]	Jason	ACKERLY/BINGHAM HIGH	92.34 %	12.67 %	N/A			9
2011	68134	[redacted]	Aireoil	VALLEY TRADITIONAL HIGH	87.07 %	20.01 %	1.38	47		11
2011	109259	[redacted]	Bobby	SHAWNEE HIGH	86.84 %	40.56 %	0.55	44		9
2011	72261	[redacted]	Stephen	VALLEY TRADITIONAL HIGH	86.28 %	41.64 %	0.00			11
2011	75138	[redacted]	Brittany	LIBERTY HIGH	84.71 %	48.37 %	0.00	12		11
2011	71717	[redacted]	Lajuanta	LIBERTY HIGH	83.76 %	49.20 %	0.20	4		12
2011	89388	[redacted]	Damontraz	LOUISVILLE METRO YOUTH CENTER	83.06 %	37.16 %	1.12	46		10



Student Number	Last Name	First Name	School Name	Probability without Intervention	Probability with Intervention	Change in Probability with Intervention	Currently in Program
87721	[redacted]	Raymeen	DOSS HIGH	74	16	58	YES
62372		Lakeem	SOUTHERN HIGH	69	40	29	YES
96225		Christopher	SHAWNEE HIGH	39	12	27	YES
68939		Courtney	FAIRDALE HIGH	34	12	22	YES
82741		Kiah	DOSS HIGH	40	21	19	YES
82348		Domiono	DOSS HIGH	30	14	16	YES
106238		Khila	DOSS HIGH	23	9	14	YES
98417		Marshaan	JEFFERSONTOWN HIGH	26	12	14	YES

Challenge: How do we model the effects of interventions?

Conclusions: The data revolution

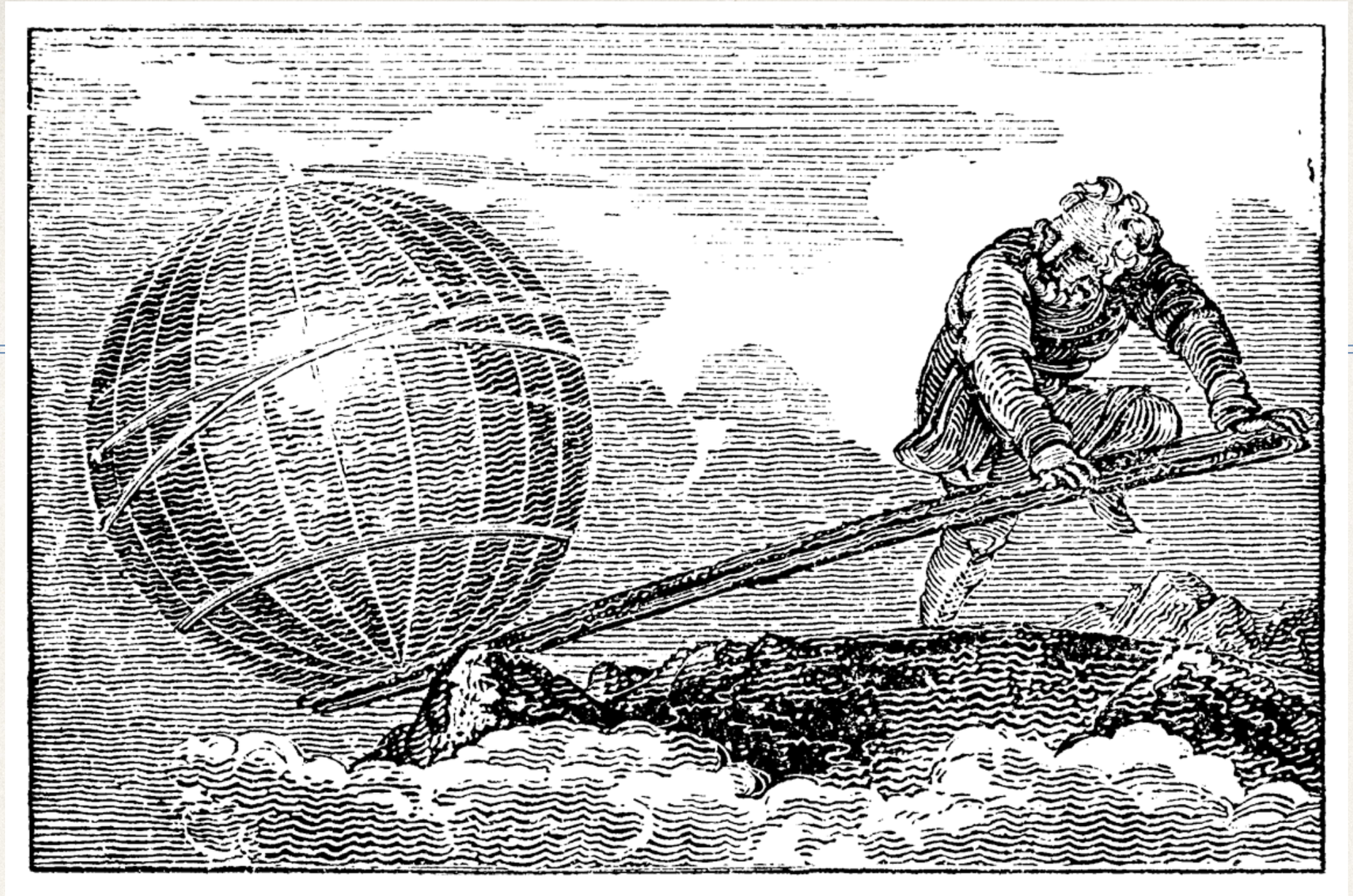
- We are generating 1 Zettabyte of data per year. That's about 1 Terrabyte per person per year. That's more than a million books!
  - Humans can't use or analyse all of that data
  - Should we just dump it or ignore it?
  - There is a huge potential for good (or ill) in much of the data
  - Who should have access to it?
  - Who should decide how its to be used?
- The collection, use and abuse of this data will probably be the most significant factor in our history over the next 100 years
- Data mining will play a more and more important role in the future

Conclusions: Complex Adaptive Systems

- We don't have adequate conceptual or theoretical frameworks in which to understand complex adaptive systems or complexity
  - Physical systems "are", while complex (adaptive) systems "do"
  - Physical systems are described by few relevant variables, for complex adaptive systems there are many that range from the micro to the macro
- Good science starts with phenomenology and taxonomy before moving on to theory
- Basically all the data generated in the data revolution is "non-scientific" and is associated with complex adaptive systems
- Data mining is not only the only way to attack this data, its also the appropriate way to develop a better phenomenological and taxonomic understanding of complex adaptive systems

Conclusions: Education

- Big data and data mining can make the design and provision of education much more effective and efficient
- Education needs to be more interdisciplinary
  - Careful, you can't become an expert in everything!
- Education needs to be more collaborative
  - Need to have much more contact between students from different disciplines collaborating to solve problems that can't be solved individually
- Education needs to be more "computational"
  - Basically all disciplines need to account for and benefit from the data revolution



δῶς μοι πᾶ στῶ καὶ τὰν γᾶν κινάσω

Give me a place to stand on and I´ll move the earth

Give me enough data and I´ll predict anything