
To boldly go where no
man has gone before

Data in Science versus Data Science: What's the difference?

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

Data Mining Course, Universidad de Pamplona
4-6 December 2017

1. **Why do we need data in science?**
2. **Isn't all science "data science"?**
3. **Why do we talk about data science now?**
4. **What's the difference?**

The principal purpose of living systems and the principal purpose of science is to…

Predict

for

Decision making

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

How we do science in a nutshell…

The traditional approach

- **The Scientific method:** DATA , DATA and DATA , and the formulation, DATA and DATA of hypotheses

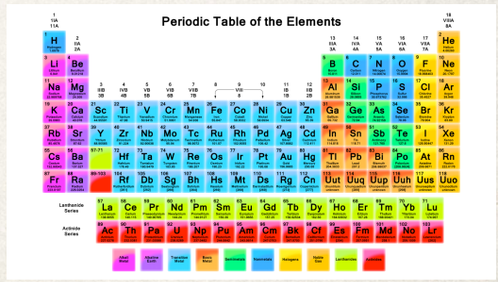
DATA

- **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.

DATA

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



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

DATA

And the "modern" approach? Data Science —> Data Mining

"… the exploration and analysis of data in order to discover patterns, correlations and other regularities."

There are two main datamining tasks

- Predicting – "pattern" identification
  - Establishes "causal" statistical relations
- Profiling - "pattern" description
  - Identifies what are the key drivers associated with a pattern

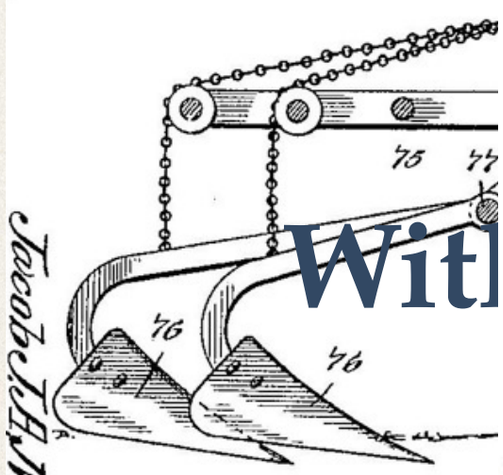
There are three main requirements

- Data; "data processors"; "inference algorithms"

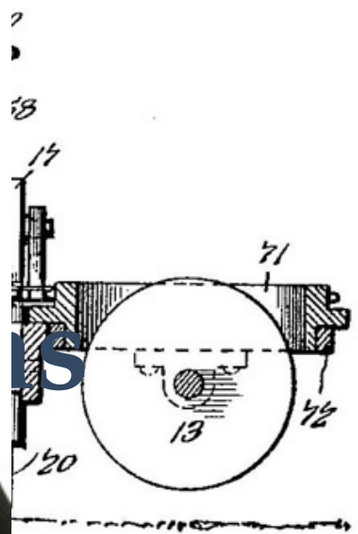
The worldview of the last 3 centuries:

In fact...

How es?



Wit



s

we are slaves of the law

The ne

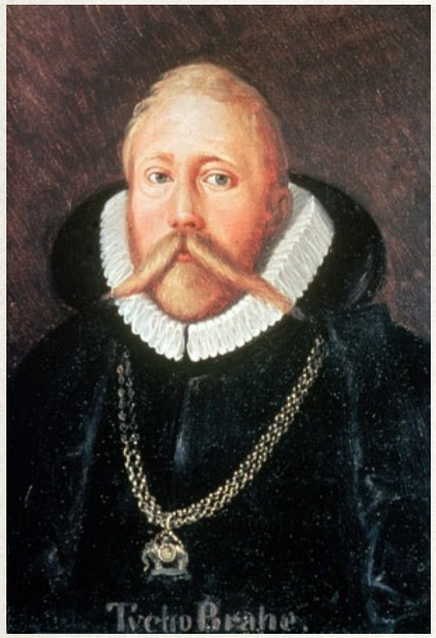
Universality

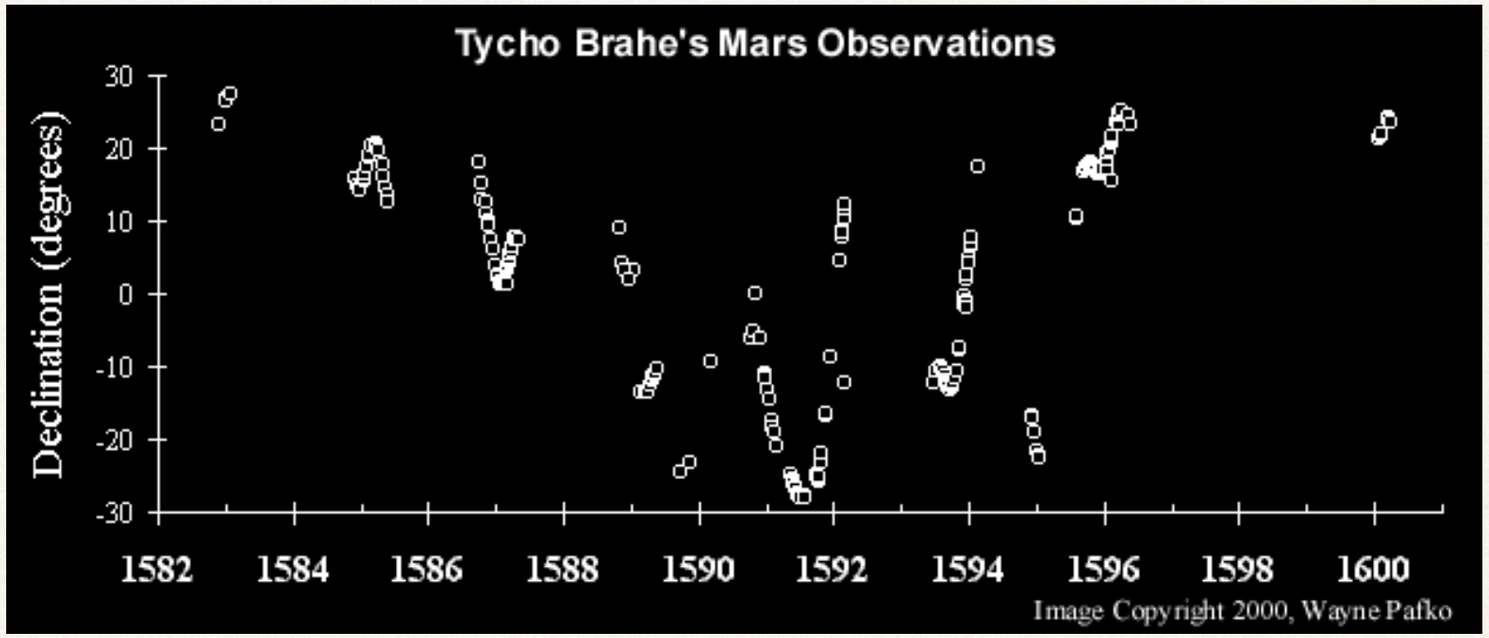
We're all equal under the law

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 to know At any interesting place that much data

Isn't all Science Data Science?

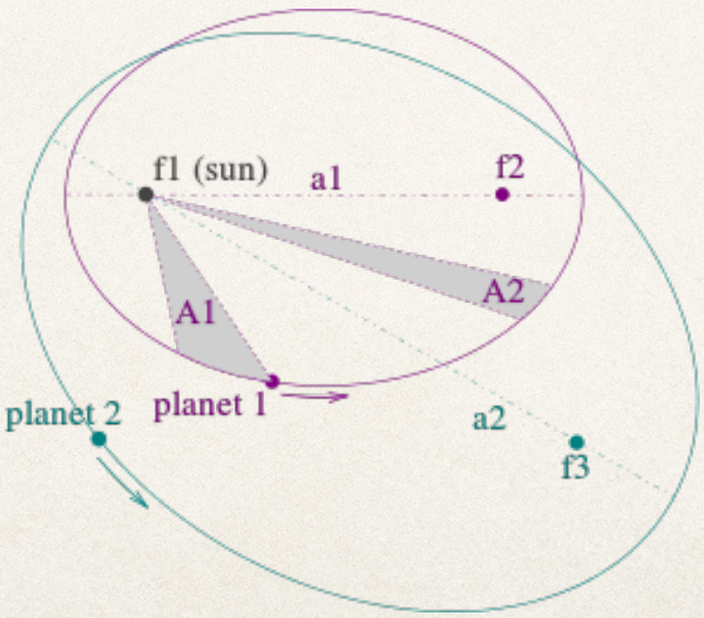
Data —> Phenomenology —> Taxonomy —> Theory





Data

Phenomenology

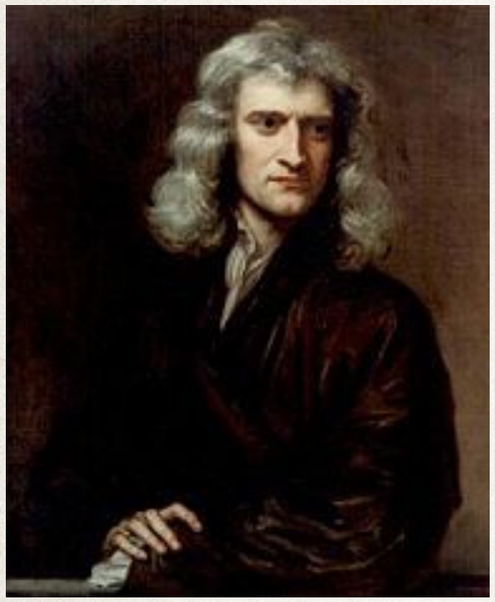


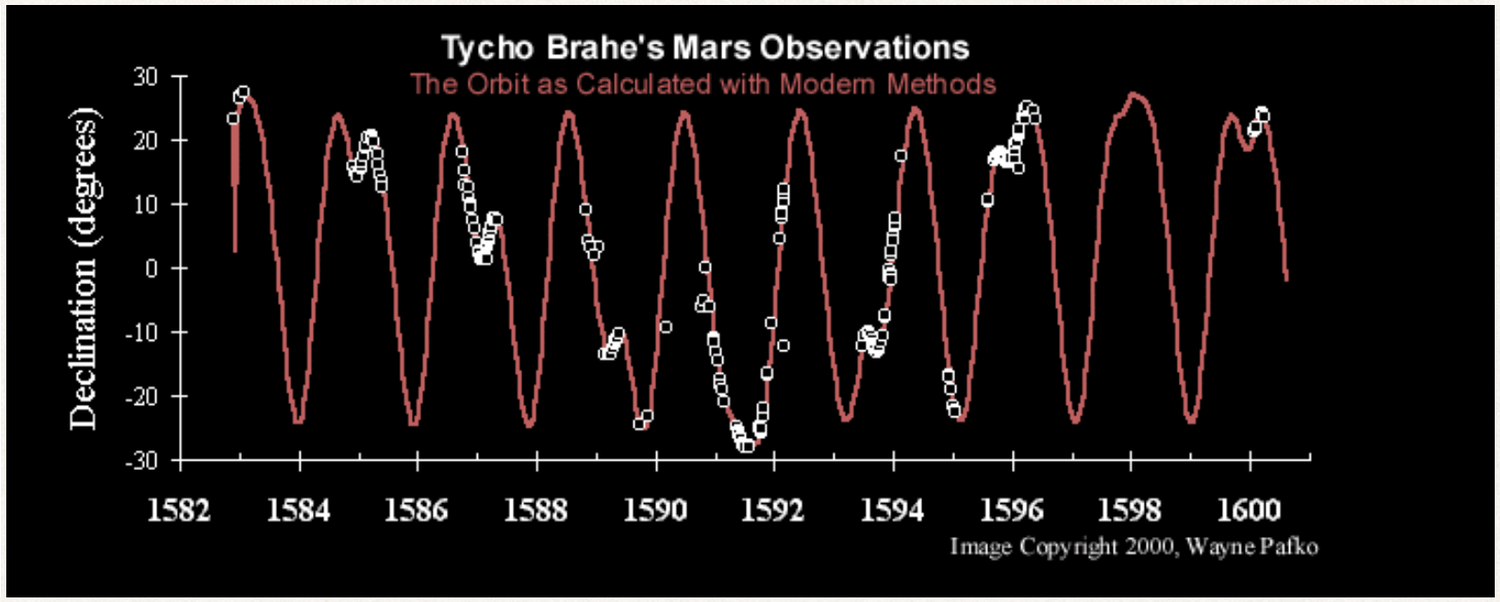
Kepler's Laws

1. The orbit of a planet is an ellipse with the Sun at one of the two foci.
2. A line segment joining a planet and the Sun sweeps out equal areas during equal intervals of time.
3. The square of the orbital period of a planet is proportional to the cube of the semi-major axis of its orbit.

Isn't all Science Data Science?

Data —> Phenomenology —> Taxonomy —> Theory





Theory

$F = ma$

$F = GMm / r^2$

Isaac Newton computed the acceleration of a planet moving according to Kepler's first and second law.

1. The *direction* of the acceleration is towards the Sun.
2. The *magnitude* of the acceleration is inversely proportional to the square of the planet's distance from the Sun (the *inverse square law*).

This implies that the Sun may be the physical cause of the acceleration of planets.

Newton defined the force acting on a planet to be the product of its mass and the acceleration. So:

1. Every planet is attracted towards the Sun.
2. The force acting on a planet is in direct proportion to the mass of the planet and in inverse proportion to the square of its distance from the Sun.

The Sun plays an unsymmetrical part, which is unjustified. So he assumed, in Newton's law of universal gravitation:

1. All bodies in the solar system attract one another.
2. The force between two bodies is in direct proportion to the product of their masses and in inverse proportion to the square of the distance between them.

As the planets have small masses compared to the Sun, the orbits conform approximately to Kepler's laws. Newton's model fits actual observations more accurately.

Science ⟷ Data Science?

- **Data**: Brahe provided an accurate (for the time) data base with data on the positions of different celestial bodies as a function of time.

- **Phenomenology**: Kepler was a data miner, a data scientist. He mined Brahe's data and **inferred** regularities and constructed phenomenological models (his three laws) that embodied these regularities.

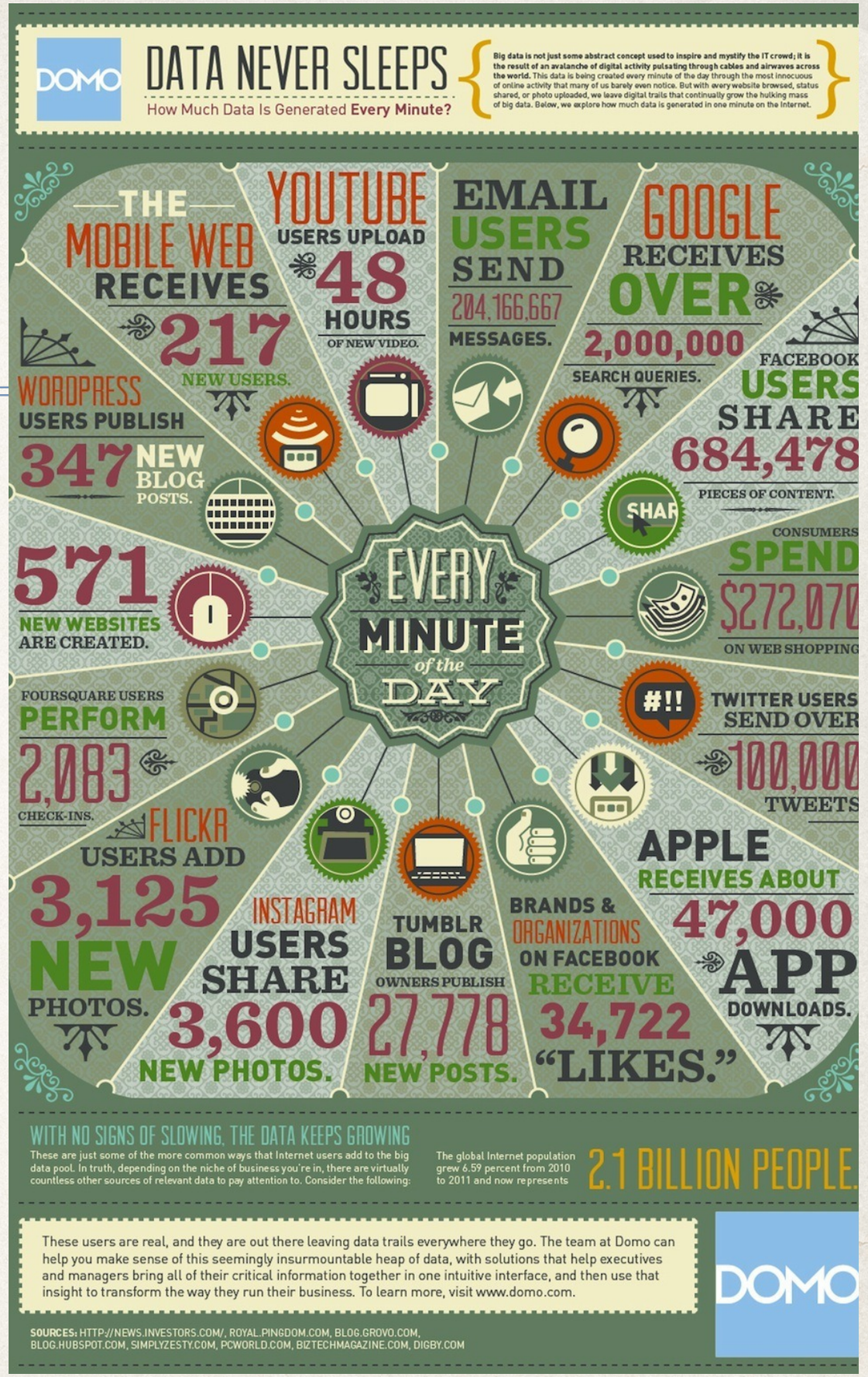
- **Theory**: Newton used Kepler's laws to construct a theoretical, "universal" model for the gravitational interaction. He **inferred** the existence and nature of an interaction between objects.

- **Where things are as a function of space and/or time allows us to infer the nature of their interactions.**
  - **By observing the spatio-temporal behaviour of different types of inanimate "things" we have deduced that in the physical world there are 4 interaction types and they are important at quite different scales.**
    - **There are only very few properties/labels of "things" that are associated with the different interactions: mass, electric charge, weak isospin, colour**
    - **These interactions DO NOT change!**

Data then versus data now

There's been a data revolution...

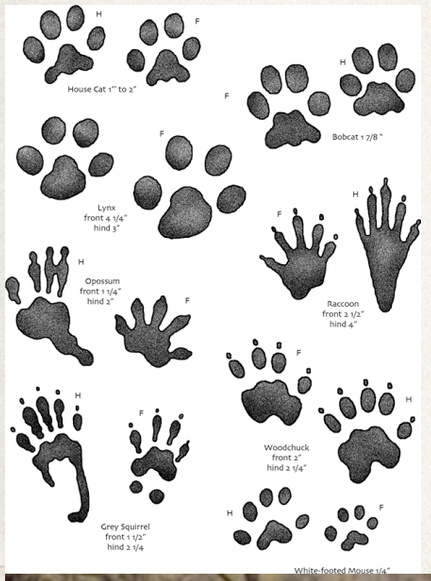
But just what's revolutionary?



Camille Kubie, Estuary, 2014

What's revolutionary?

Data types? No.



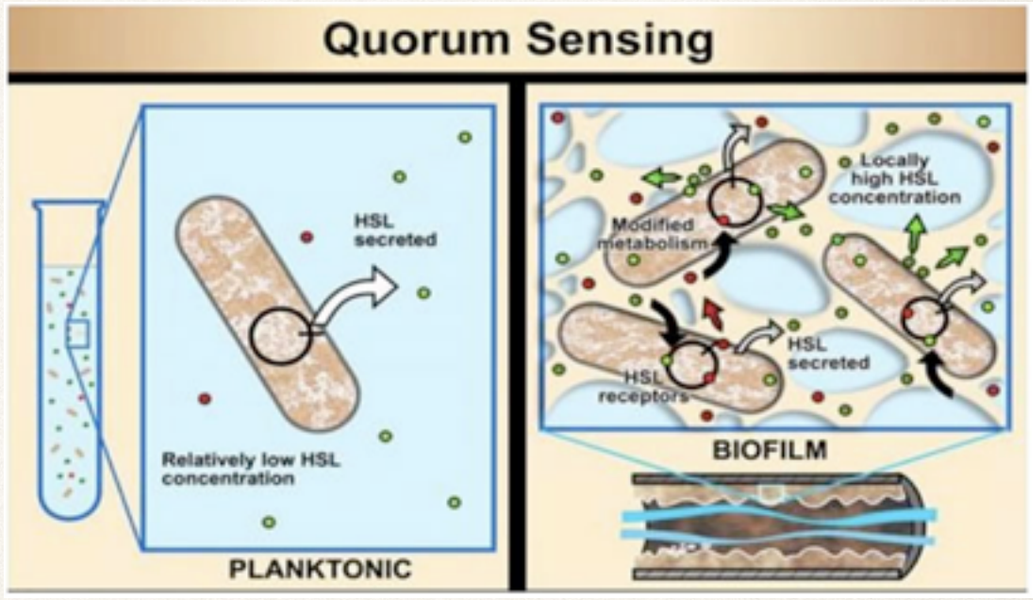
Raw data:

Chemical

Electromagnetic

Acoustic...

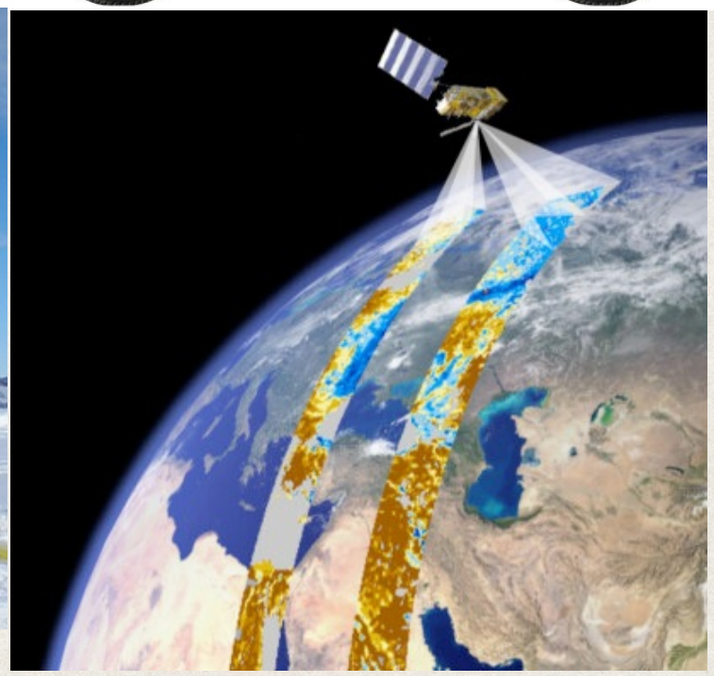
as functions of space and time tell us what is going on in the world.



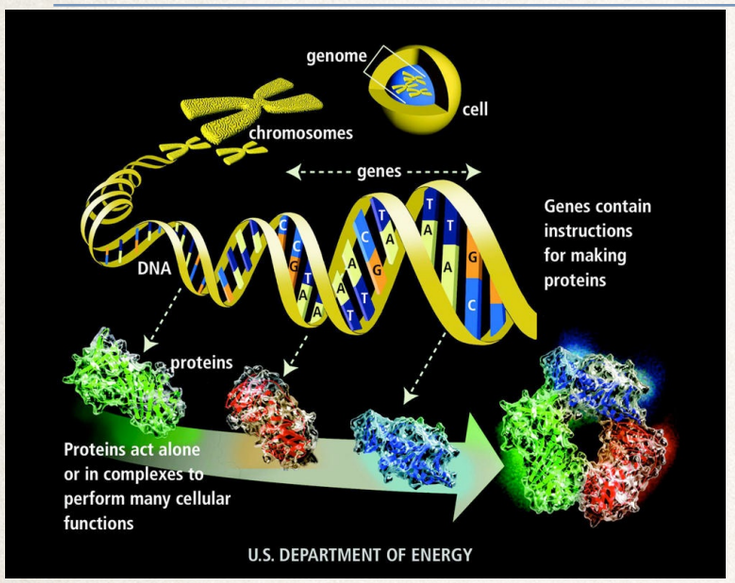
We use data about ***events***

to take ***decisions.***

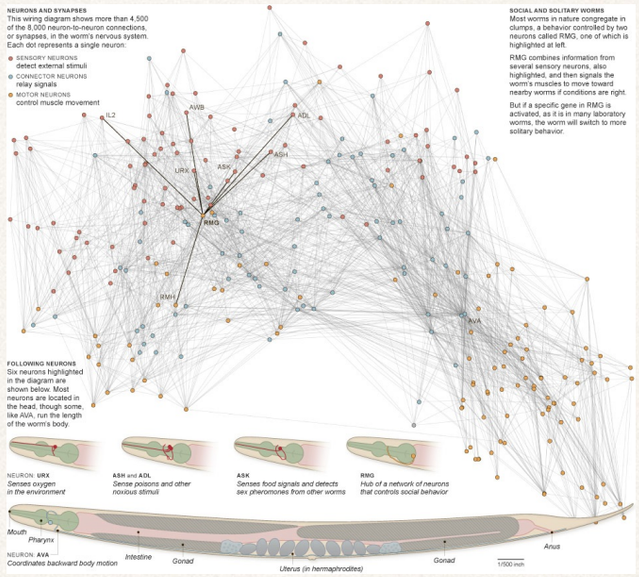
Data sensing? Yes.



Data storage and processing? Yes.



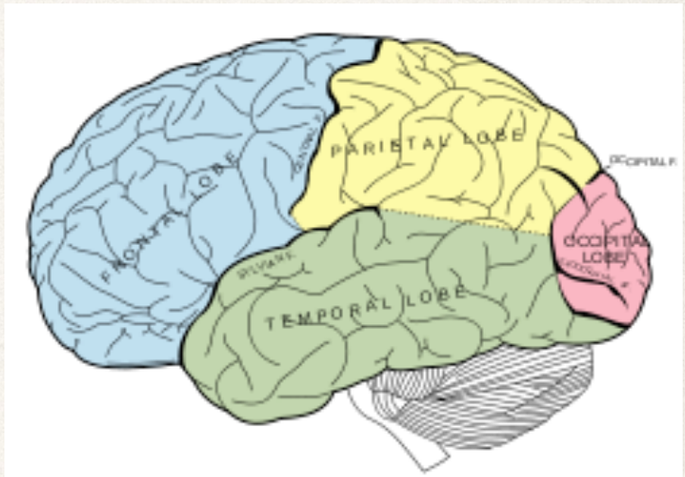
Genomes 1kB - 1.5 GB



Worm neural network 0.3MB

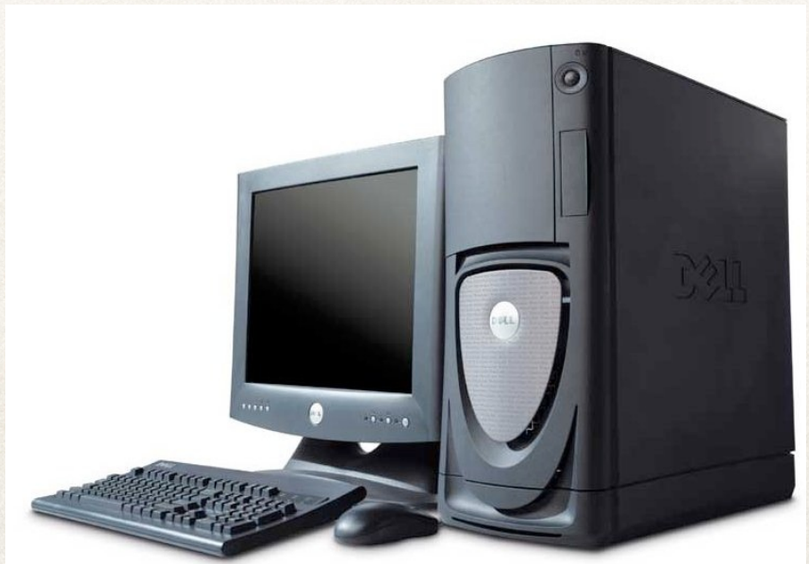
Human brain

10-100 Terrabytes



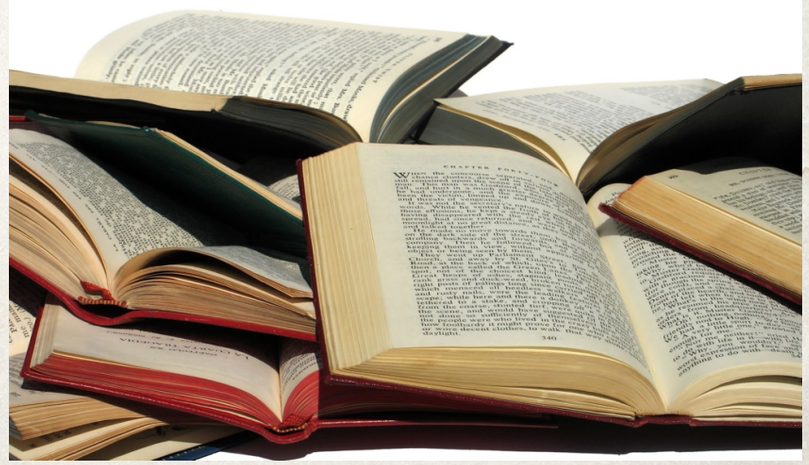
In electronic form 1 zettabyte

Raw data is processed and stored



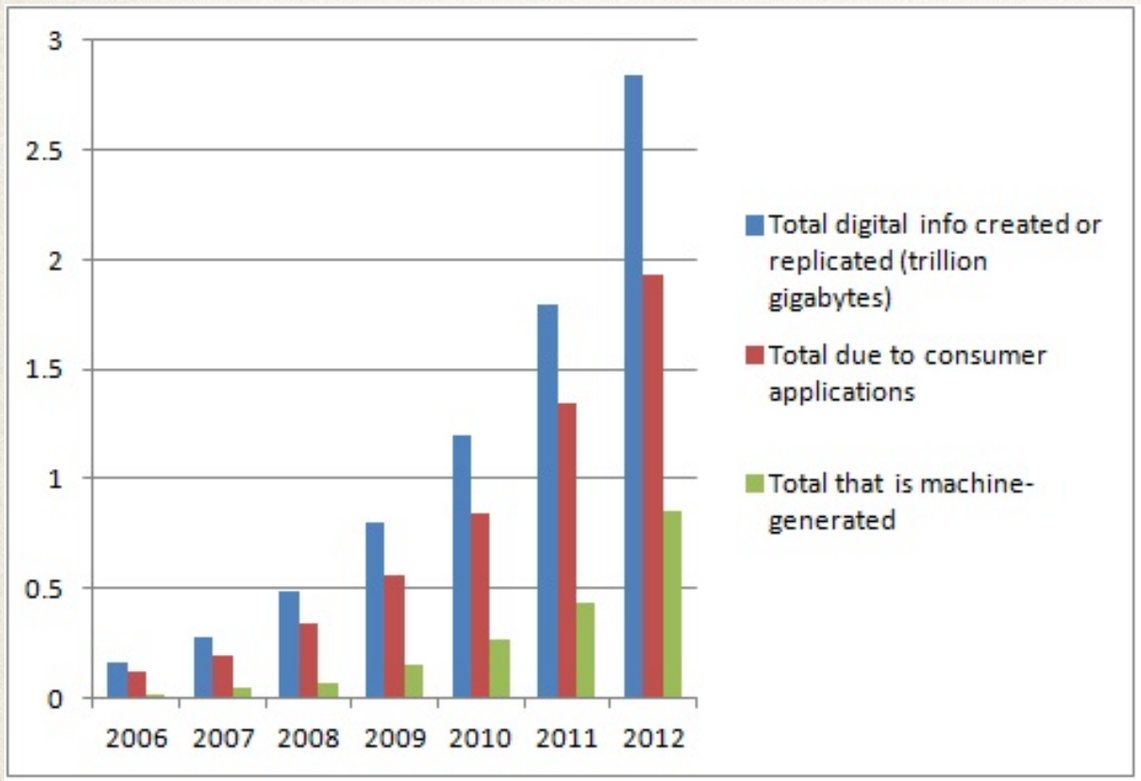
All the books in the world

30-50 Terrabytes



What's revolutionary?

Data storage and processing? Yes.



Source: IDC

Data growth by type

28 million books
> 20 terrabytes text data

Some of the world's biggest databases

amazon.com
> 42 terrabytes

AT&T
323 terrabytes
1.9 trillion call records

WORLD METEOROLOGICAL ORGANIZATION
ORGANIZACIÓN METEOROLÓGICA MUNDIAL
ORGANISATION MÉTÉOROLOGIQUE MONDIALE
ВСЕМИРНАЯ МЕТЕОРОЛОГИЧЕСКАЯ ОРГАНИЗАЦИЯ
GLOBAL ATMOSPHERE WATCH
CO_2 O_3 SO_4

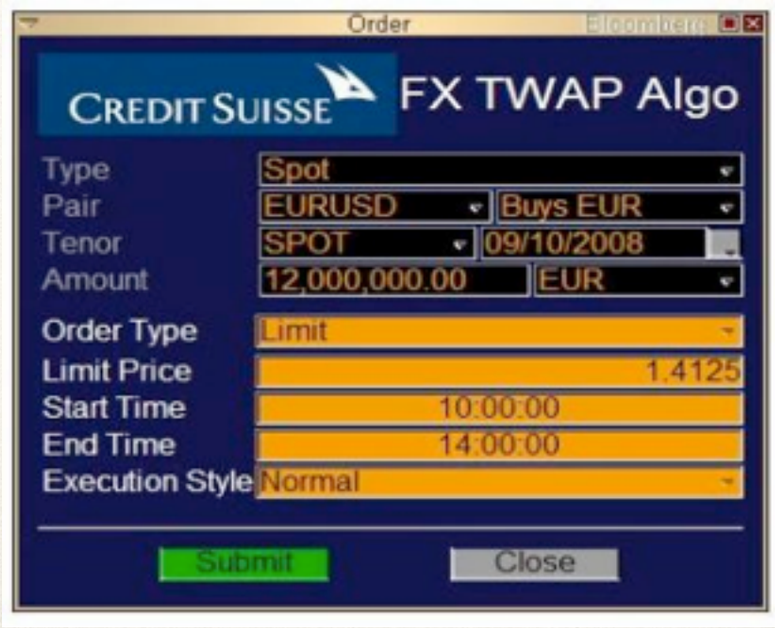
220 terabytes of web data
6 petabytes of additional data

What's revolutionary?

Data storage and processing? Yes.

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

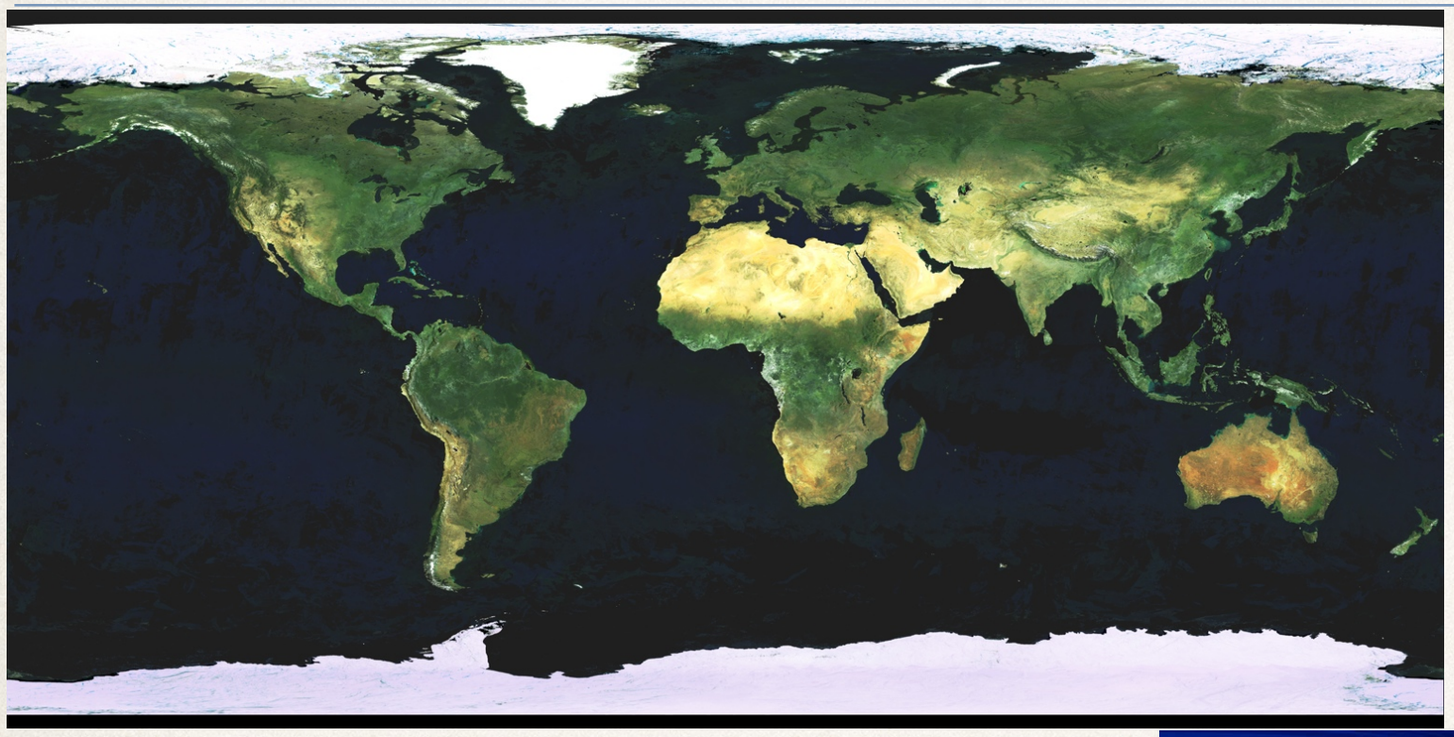
For example, a financial market where...



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

What's revolutionary?

Data connectivity? Yes.

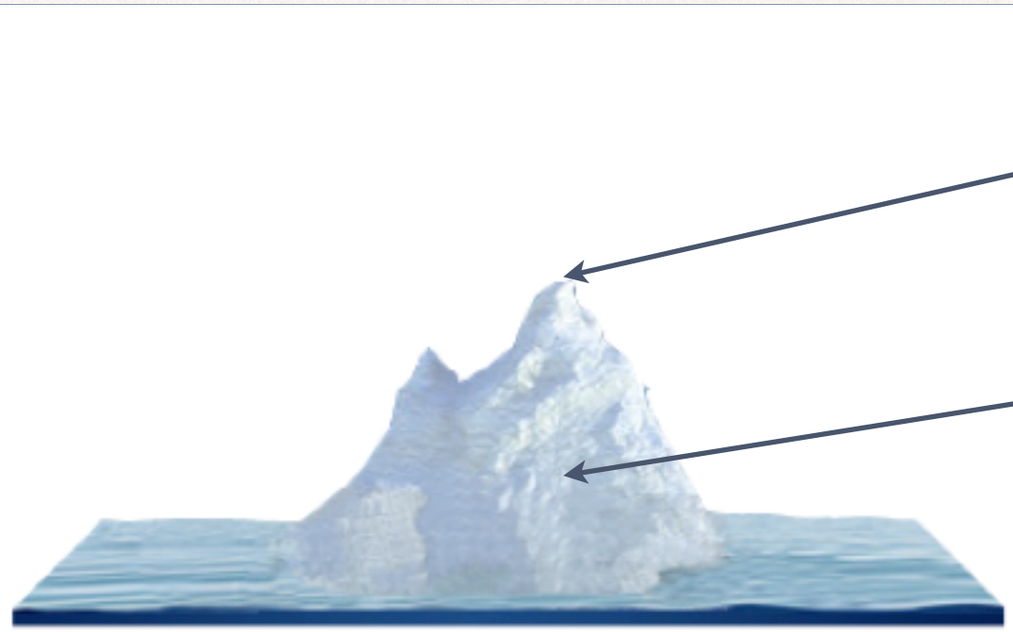


Then

Real space --> cyberspace

Now

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

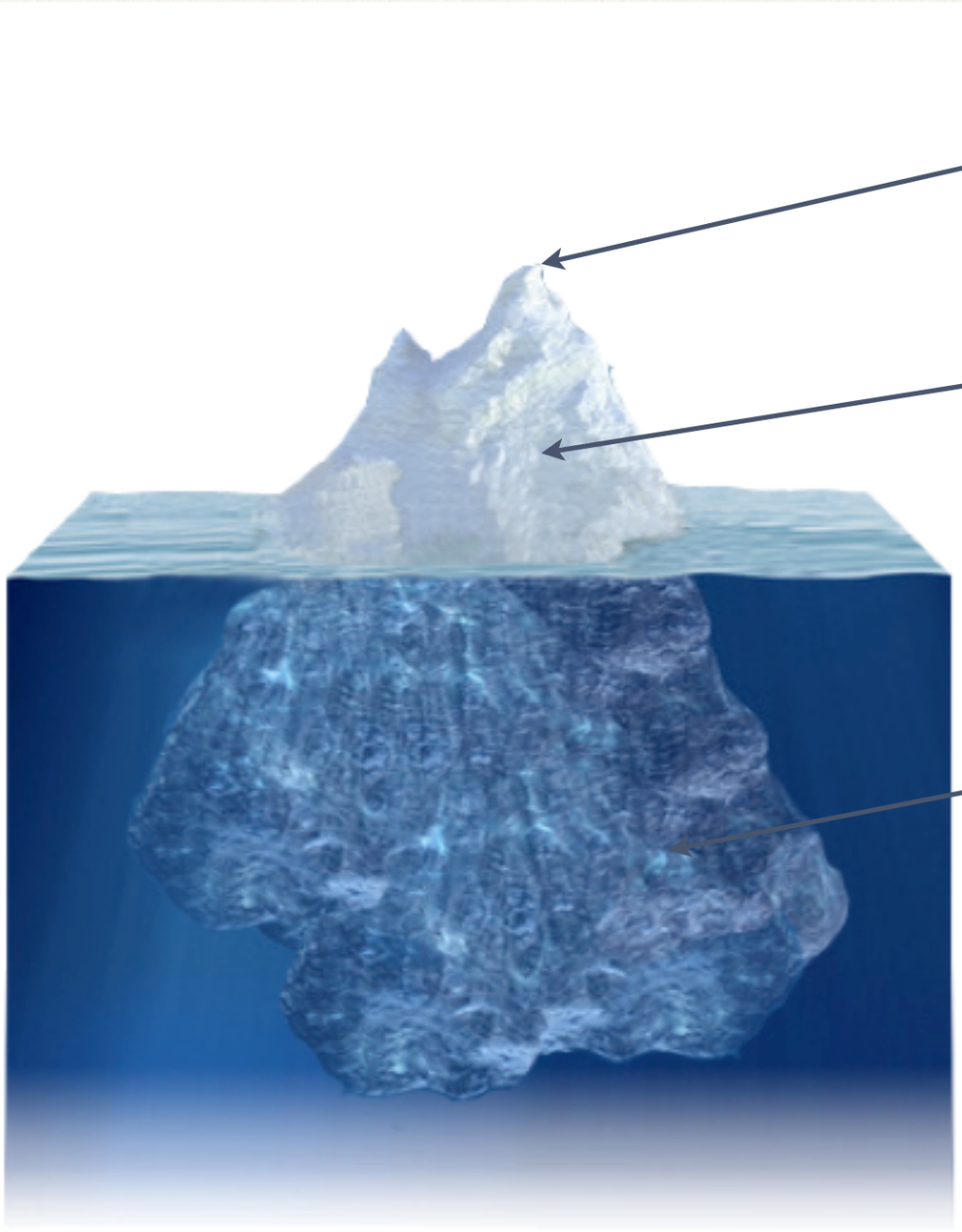


My data:
a snowflake

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 **...**

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 **...**

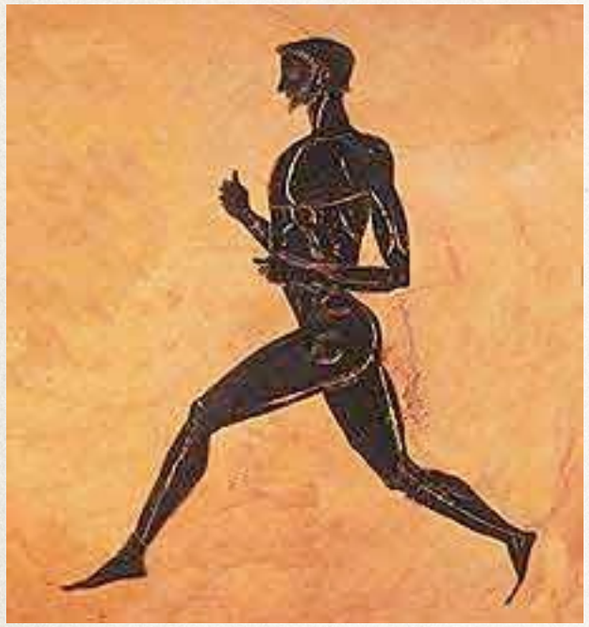
Victorian education reform: Education Act 1870
www.architecture.com/.../**Education**InAModernWor... - United Kingdom
Victorian education reform: Education Act 1870. Perspective view of Harper Street School, New Kent Road, London, 1885. Print Designer: Robert W Edis **...**

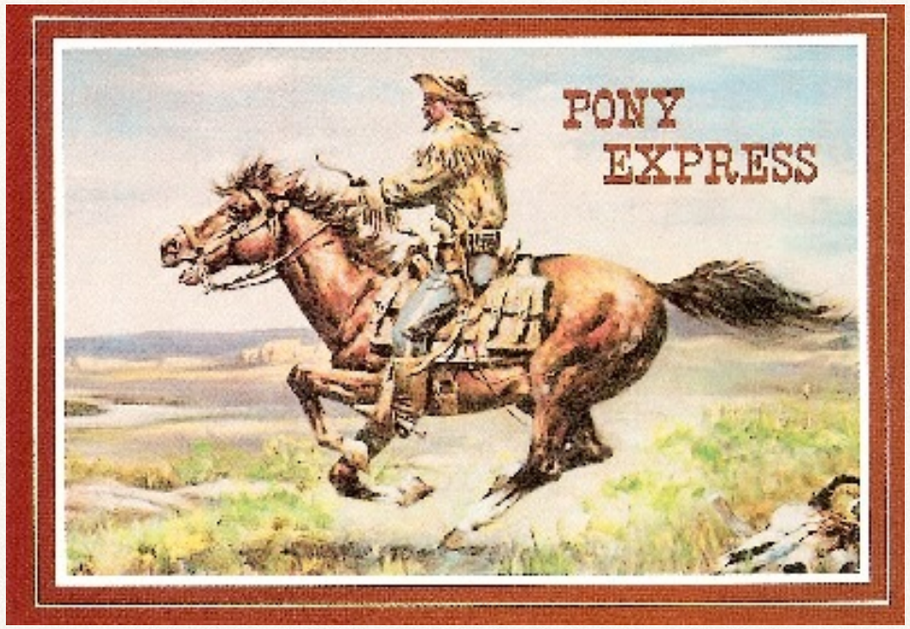
It's fast but not clever!

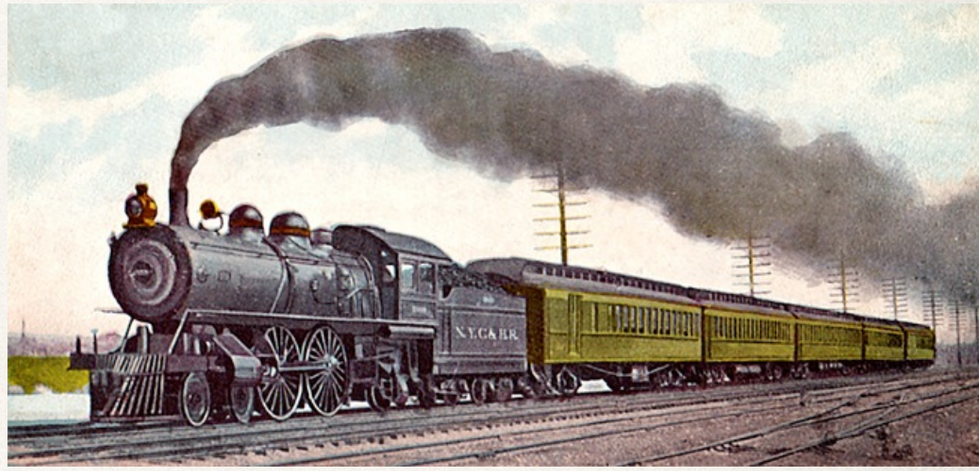
Humans are wonderful at semantics, machines aren't

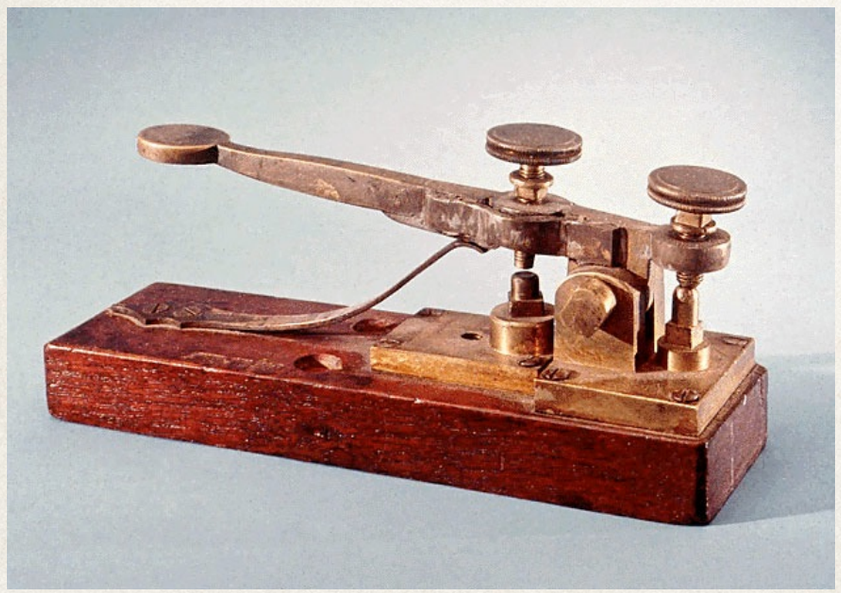
What's revolutionary?

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









What's revolutionary?

Data purpose? No.

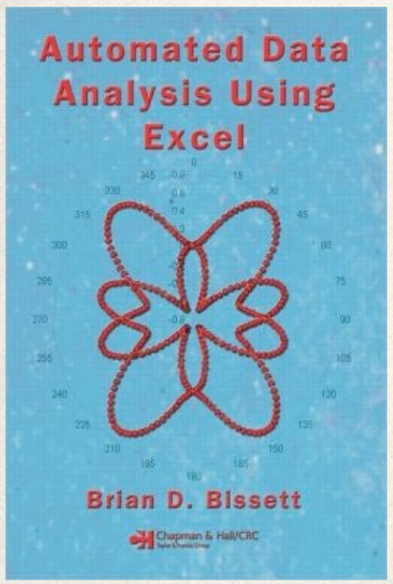
- We have always used data to take decisions

- 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.

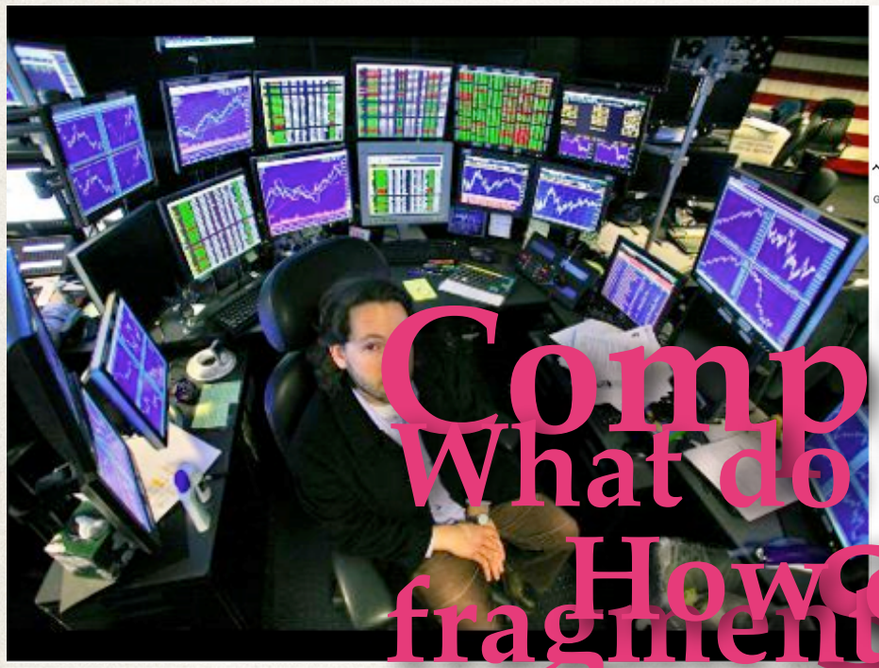
Instinct: Hard-wired data analysis

Learning: adaptive data analysis

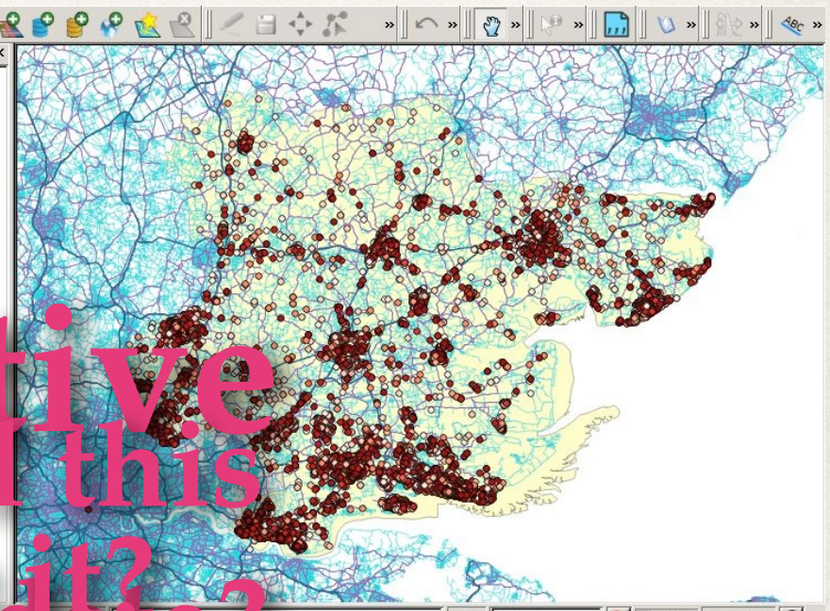


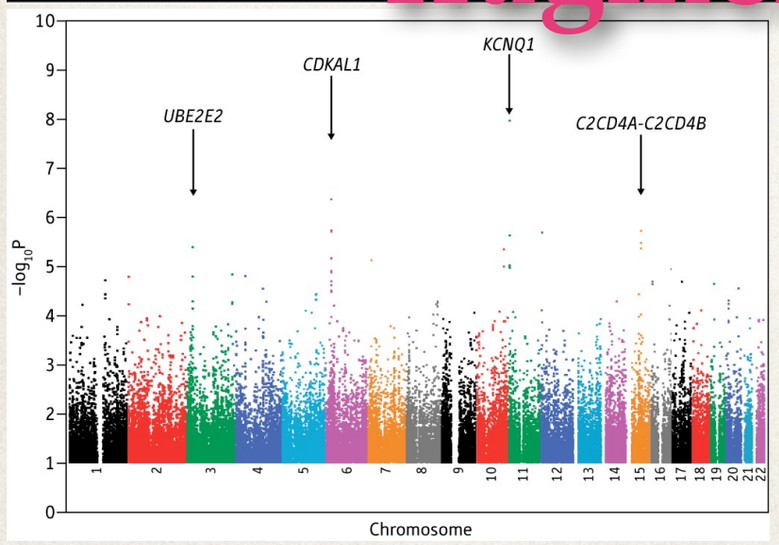
Automated data analysis - data mining.
Can consider large volumes of data with large numbers of variables

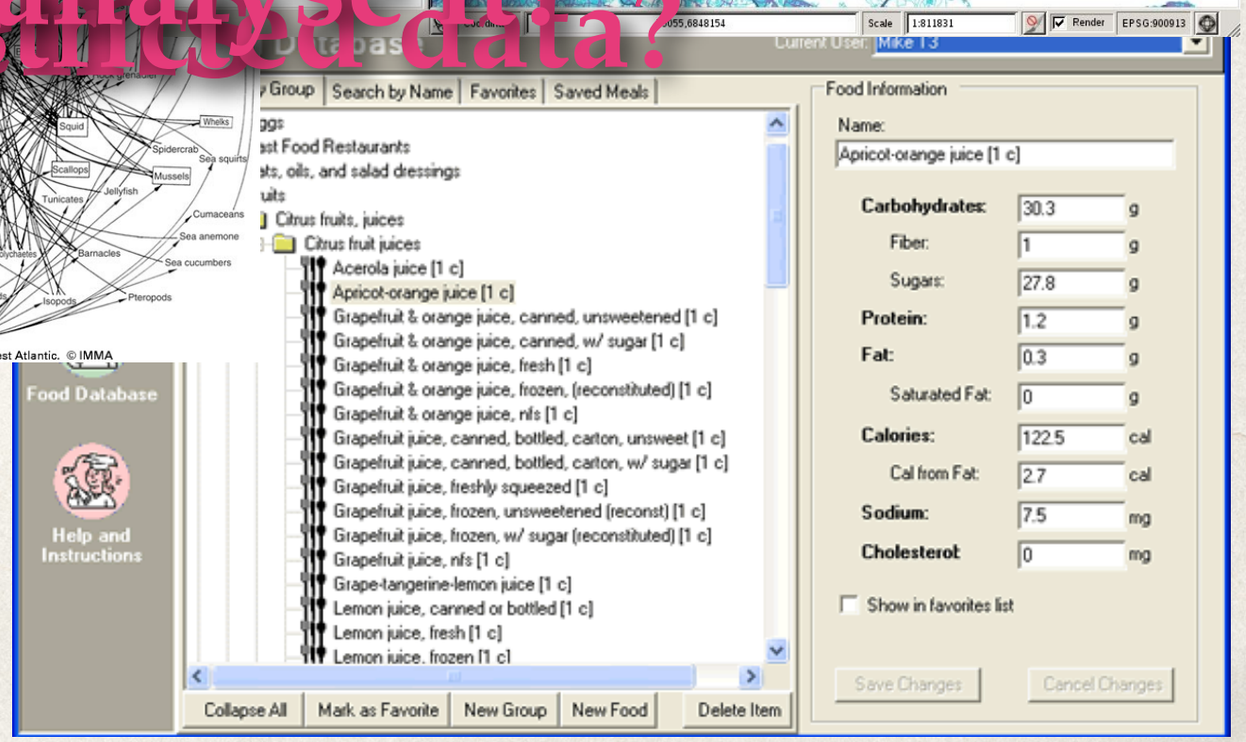
What does all the data from the data revolution represent?



A simplified food web for the Northwest Atlantic. © IMMA







Complex Adaptive What do we do with all this fragmented, restricted data?

How do we analyse it?

In Complex Adaptive Systems we need a lot of data because...

t you

Imagine what you can say about a city

versus

a crystal as big as a city!

The difference between BIG data and DEEP data

You can say a lot about a Complex Adaptive System and each thing you say depends on a lot of other things

---

Any observable of a complex system depends on a whole host of other factors

P(A,B,C,D,...; t | a,b,c,d,...; t′)

Diabetes Renal failure Obesity Father had diabetes

Angina Leukaemia SNP Rs7903146 45 mins exercise per week

Many effects Many causes

From the “micro” to the “macro”

Many disciplines

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

What distinguishes complex from non-complex phenomena?

Structural properties

- A “hierarchy” of many different scales
- Effective degrees of freedom (“collectivity”) that are qualitatively different at different scales
- Hierarchies of **building blocks** (modularity)
- Interactions that are stronger “intra-block” than “inter-block”
- The micro and macro and linked through feedback (fitness, meaning,…)

Functional properties

- Systems that are **adaptive**
- A dynamics that depends on many different **rules/strategies**
- Systems that “**learn**” (feedback from the environment to the system that is used to update the rules)
- The micro and macro and linked through feedback (fitness, meaning,…)
- More complex behaviour (the “phenotype”)
- Better described by what they **DO** than what they **ARE**

Is complexity a scientific concept?

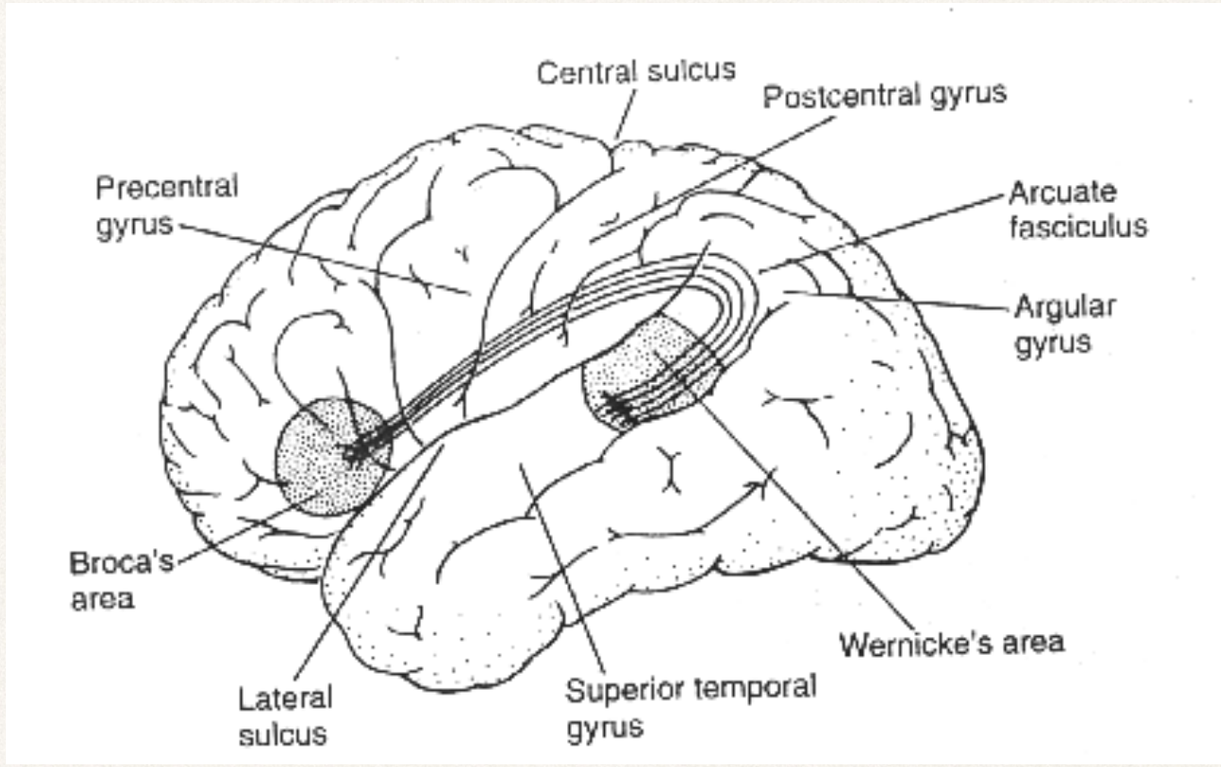
If it is, then...

How do we measure it?

What is a good measuring apparatus…?

For Symbolic Complexity

To be or not to be that is the question.

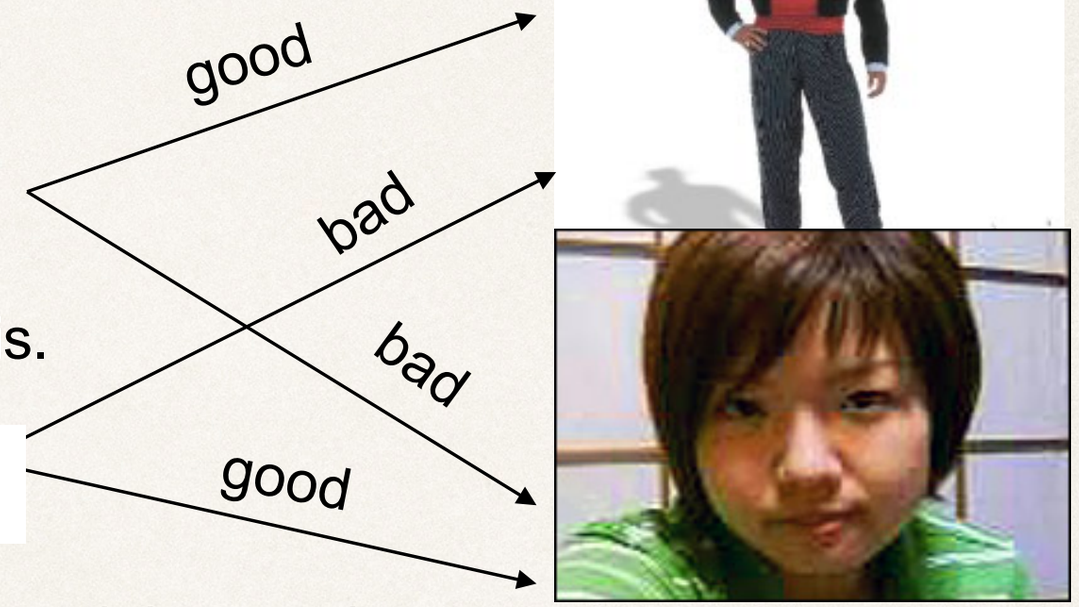


This apparatus is surely capable of measuring complexity. Or maybe not…?

How good is your apparatus?

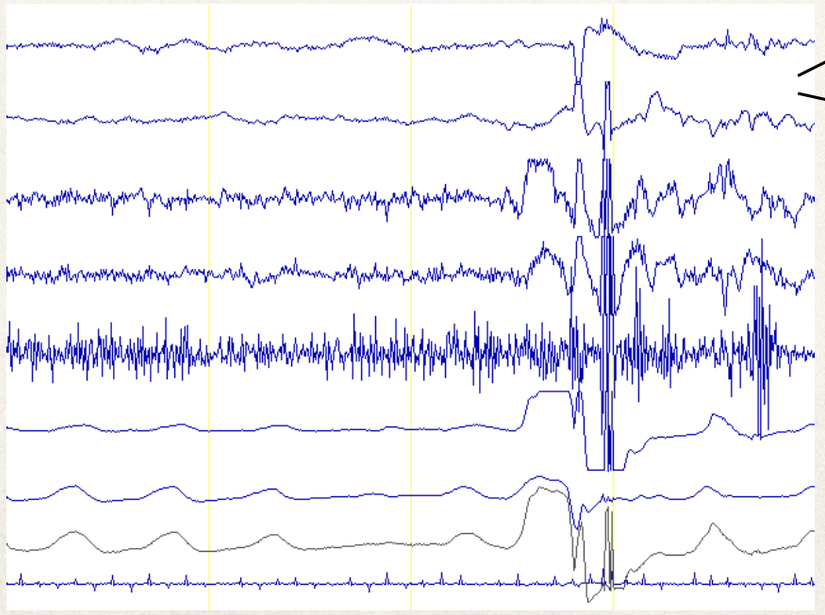
- To be or not to be that is the question.
- Para ser o no ser que es la pregunta.
- Om te zijn of te zijn niet dat de vraag is.

あるためまたはないため質問である



- Because of a certain or because it is not, it is question.
- Because or it is not for the sake of, that having asked and being convinced.
- Being not to be for the sake of, or that that, you ask, are convinced.
- It is that without having for the sake of, or, you ask, are convinced.

Is this different to physics?

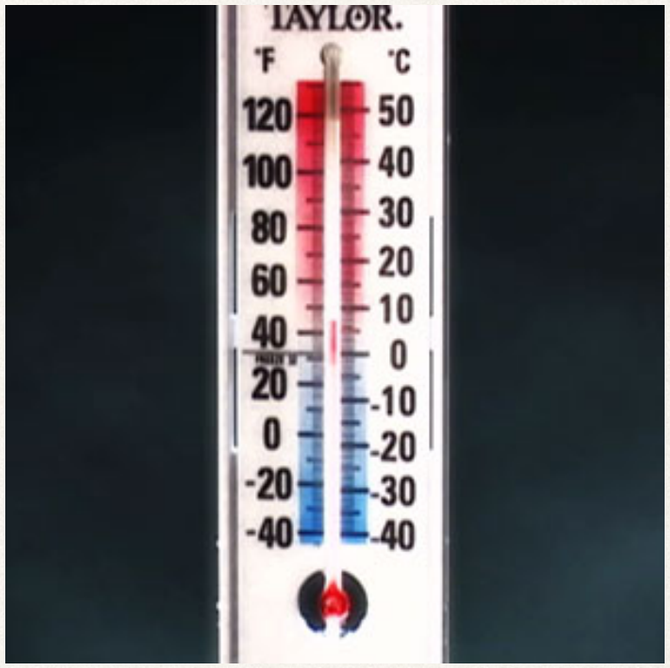


good

bad

bad

good



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

Does this represent a Complex System?

$$\mathbf{d}_i(t+\Delta t) = \sum_{j \neq i} \frac{\mathbf{c}_j(t) - \mathbf{c}_i(t)}{|(\mathbf{c}_j(t) - \mathbf{c}_i(t))|} + \sum_{j=1} \frac{\mathbf{v}_j(t)}{|\mathbf{v}_j(t)|}$$

Competition between an effective repulsion and attraction between "particles"

$$\hat{\mathbf{d}}_i(t+\Delta t) = \mathbf{d}_i(t+\Delta t) / |\mathbf{d}_i(t+\Delta t)|$$

$c_i(t)$, $v_i(t)$ – position/direction vectors of a "particle"

$$\mathbf{d}_i{}'(t+\Delta t) = \frac{\hat{\mathbf{d}}_i(t+\Delta t) + \omega \mathbf{g}_i}{|\hat{\mathbf{d}}_i(t+\Delta t) + \omega \mathbf{g}_i|}$$

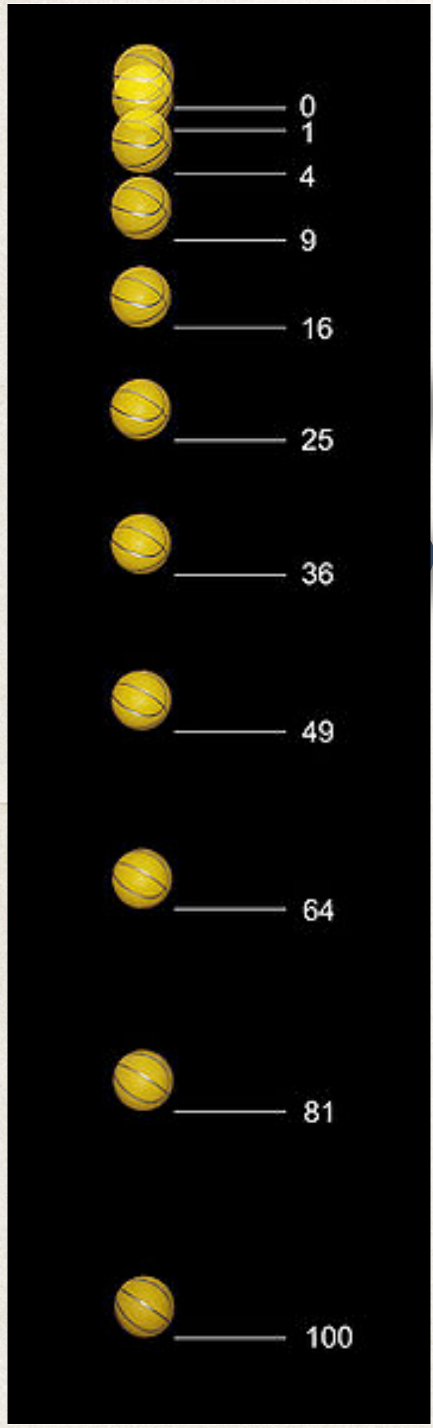
Equation for "charged" particles in an external field g_i

Couzin, I.D., Krause, J., Franks, N.R. & Levin, S.A. (2005) *Nature,* **433**, 513-516.

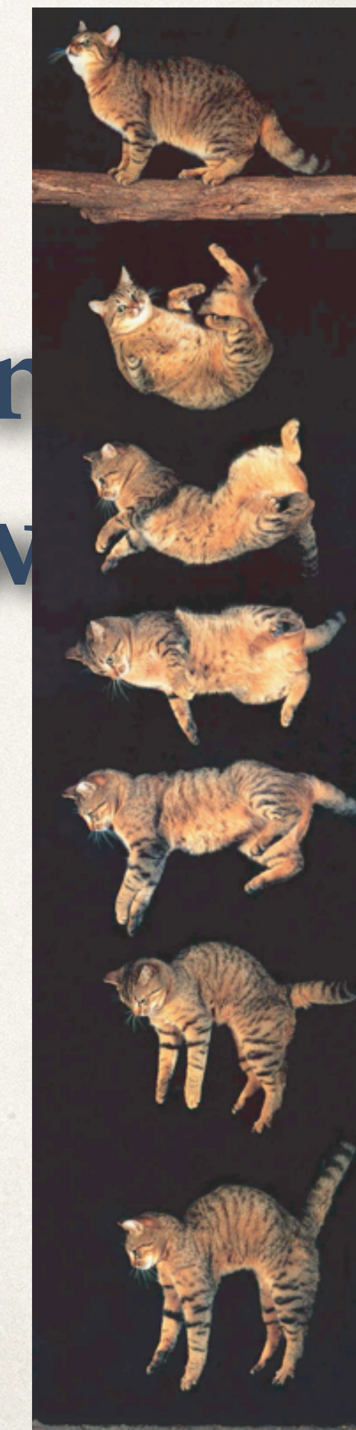
Does this represent a "complex" system?

Moral: It's important to distinguish between a description of complexity and a non-complex description of a phenomenon or behaviour associated with a complex system.

Predictability in physical systems versus Complex Adaptive Systems



Mechanistic **Adaptive**



The *evolution* of function is the revolution that allowed systems to escape the tyranny of physics alone.

...e difference between complex an... ...ple systems is the difference betw... "being" and "doing".

The cat obeys the same laws of physics as the ball. But it's not a "slave" to them.

Complexity is a consequence of that revolution.

The Difference Between "Being" and "Doing"

In biological, economic and social systems, i.e., complex adaptive systems, organisms exhibit a great diversity of **STRATEGIES (rules/models)** that lead to decisions. A strategy can be viewed as a sequence of decisions.

The dynamical state of an individual at t+1 depends not only on the state of the individual at other times t but also on the strategy (update rule) selected at time t, that in turn depends on the rules of others at t. Thus, it is necessary to work in a space of states AND strategies/rules/models – sounds like game theory but …

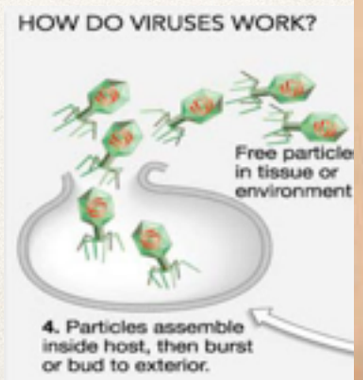
We don't know what this space is!

Besides, the payoff for a strategy is RELATIVE not absolute. Payoff (fitness) should be an emergent property. Imagine at the beginning of life trying to specify a priori the fitness of a lion or an ant!

Evolved Virtual Creatures

Examples from work in progress

m … s" … both … dual … d collective level



There are good decisions
and there are bad decisions

Predictibilidad en los Sistemas Simples versus los Sistemas Complejos

1) Suelto un objeto de mi mano. ¿Qué pasará?

Cae al suelo 100% Se queda colgado en el aire 0%

2) Dejo dos objetos de distintas masas caen de mis manos.

¿Cuál tocará piso primero?

Lo mas pesado 0% Lo menos pesado 0% Ambos al mismo tiempo 100%

3) Empujo este objeto con mi mano. ¿Qué pasa?

Se mueve 100% Se queda sin mover 0%

Fenomenología: experiencia cotidiana Las leyes de Newton

¿Qué es predecible?
¿Los seres humanos?

1) No han tomado agua (ningún liquido) en tres días. Alguien te ofrece un litro de agua o una caja de hojuelas. ¿Qué seleccionas?

Agua 100% Hojuelas 0%

2) Hay un incendio grave en el auditorio y suena la alarma. Yo les invito esperar hasta el final de mi platica o se huyen. ¿Qué haces?

Huyes 100% Se queda 0%

3) Tienen mucho, mucho hambre. Alguien les ofrece una comida de 1500 calorias para satisfacerles. Pueden seleccionar entre carnitas, enchiladas suizas y frijoles negros; o puro apio (7.5kg). ¿Qué seleccionas?

Carnitas etc. 100% Apio 0%

¿Qué es predecible?
¿Los seres humanos?

4) No han tomado agua (ningún liquido) en tres días. Alguien te ofrece un litro de Coca-cola o un litro de Pepsi. ¿Qué seleccionas?

Coca cola 70% Pepsi 30%

5) Hay un incendio grave en el auditorio y suena la alarma. Llegaste a la salida pero notas alguien quien no conoces atrapado. Regresas para tratar de ayudarles arriesgando tu propia vida o sigues corriendo?

Si regresas 50% No regresas 50%

6) Tienen mucho, mucho hambre. Alguien les ofrece una comida de 1500 calorias para satisfacerles. Pueden seleccionar entre carnitas, enchiladas suizas y frijoles negros; o barbacoa, chicharrón y arroz ¿Qué seleccionas?

Carnitas etc. 50% Barbacoa etc. 50%

Conclusions:

- All science is “data science” - from its beginnings to the present day. There is no science without data.

- Because of universality physical systems are relatively phenomenologically poor (**big** data but not **deep**!) and therefore need relatively little (shallow) data (CERN, LIGO, etc. notwithstanding)

- Phenomena at one scale in the physical sciences is shielded from the rest (“effective theories”) and that’s why physics has been so successful

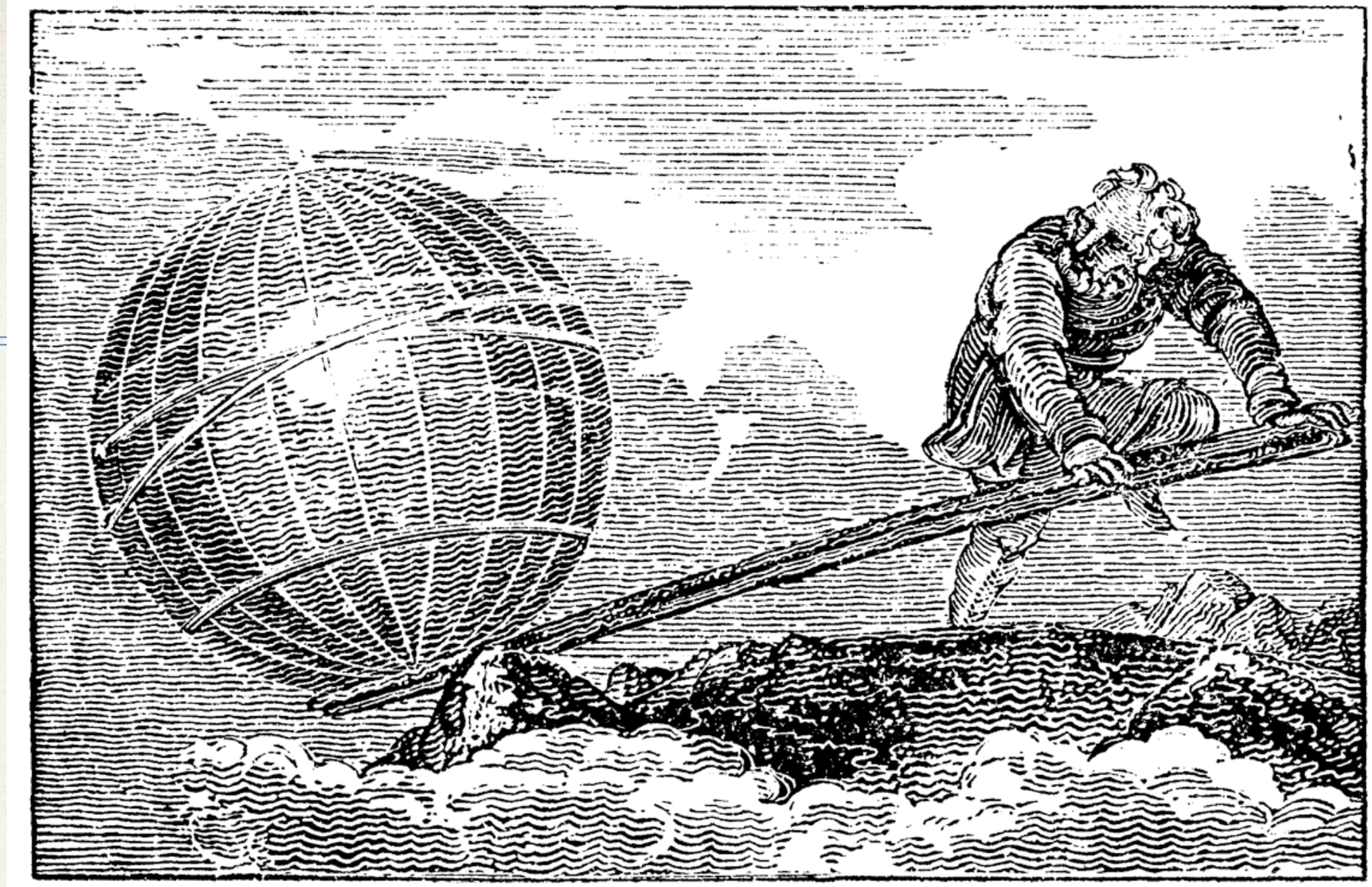
- Complex Adaptive Systems are phenomenologically deep and need a lot of data to describe them…

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

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 - we have many "Brahes" but we need more "Keplers" and less "Newtons"
- Basically all the data generated in the data revolution is "non-scientific" and is associated with complex adaptive systems
- The traditional modeling frameworks of the physical sciences are much less useful in the Complex Sciences
- Data mining/machine learning/Deep learning/… are just methodological tools to help us be better "Keplers". They offer the best way to attack this data. Its also the appropriate way to develop a better phenomenological and taxonomic understanding of complex adaptive systems



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