

# Data: Crisis and Opportunity?

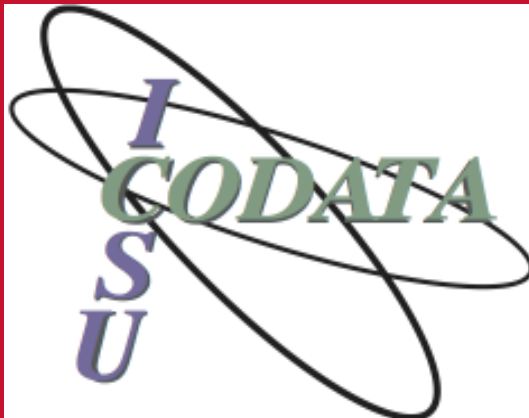
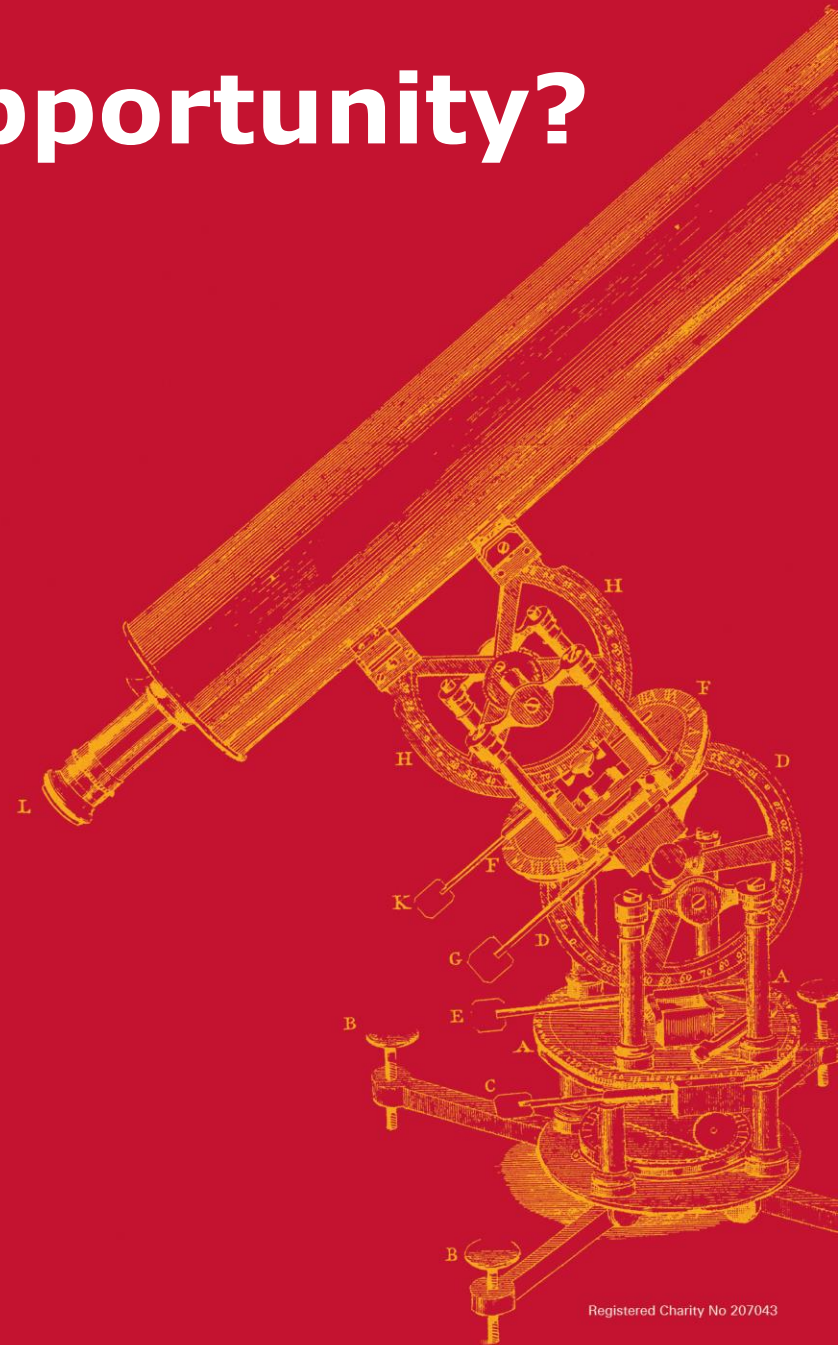
**Geoffrey Boulton**

University of Edinburgh

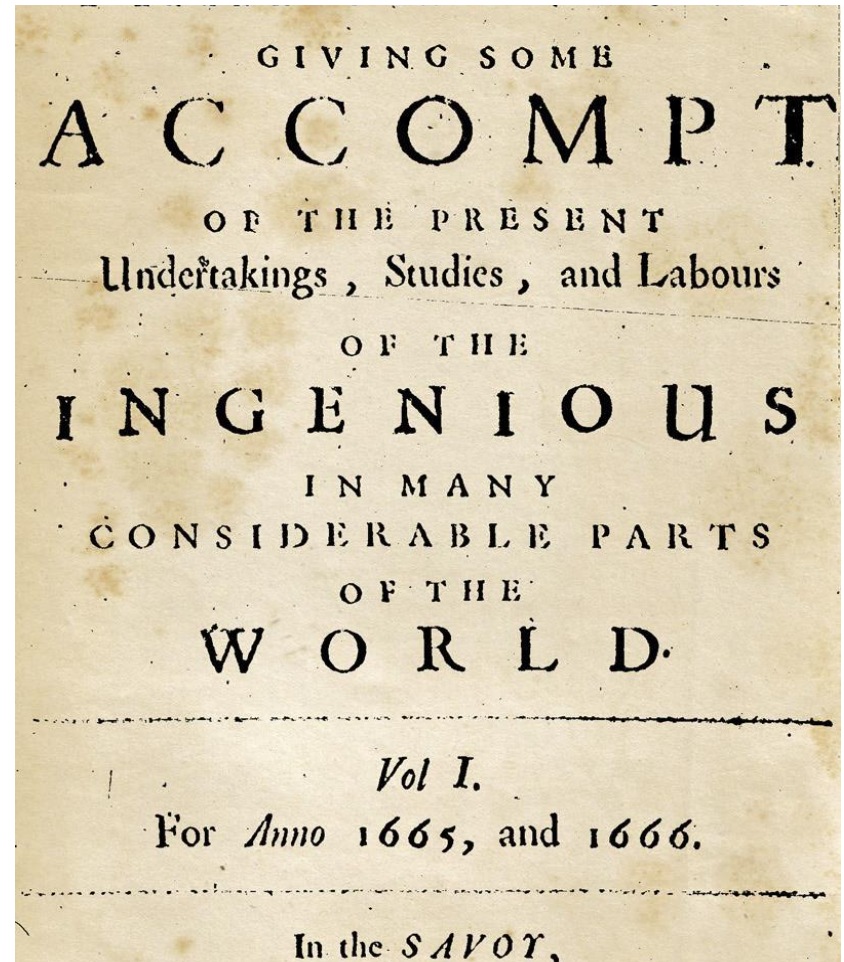
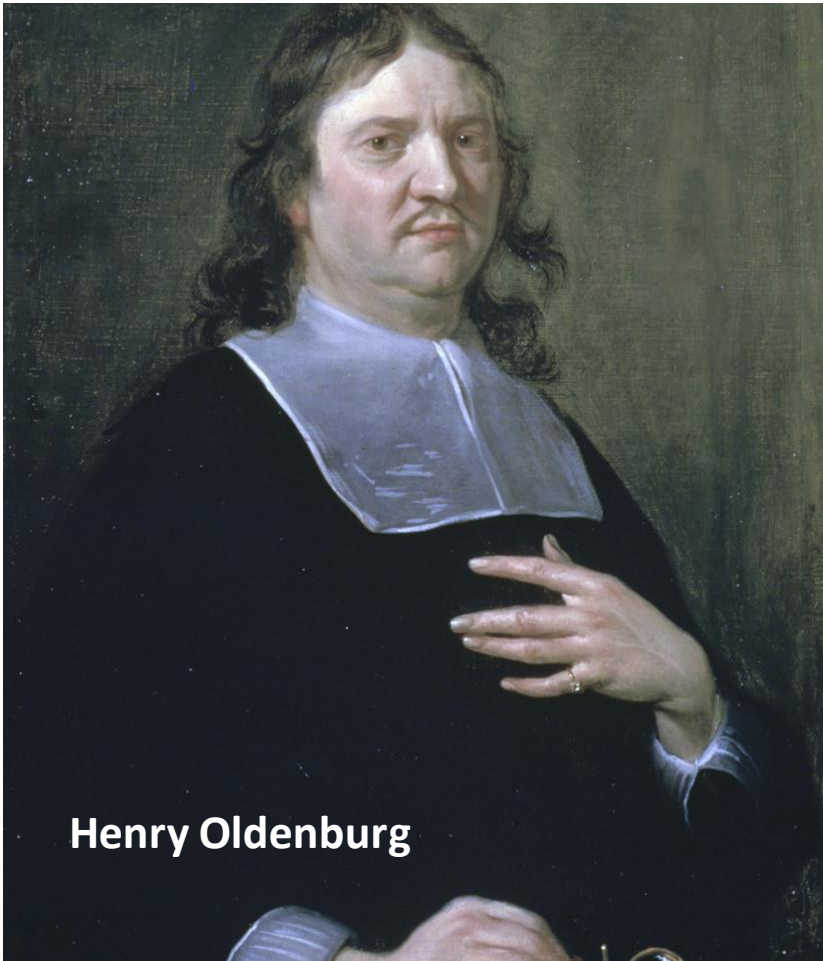
**World Science Forum**

**Budapest**

**November 2015**



# Open communication of data: the source of a scientific revolution and the basis of scientific progress



## Scientific self correction



The progress of science is strewn, like an ancient desert trail, with the bleached skeleton of discarded theories which once seemed to possess eternal life.

(Arthur Koestler)



Protein

Data

P4578

Gene

# The Data Deluge

12' 245.99

10<sup>20</sup> bytes

problems & opportunities

DATA GROWTH

IT BUDGET SHORTFALL

Available storage

IT BUDGETS (INCREASE)

COST OF STORAGE/GB (DECREASE)

2011

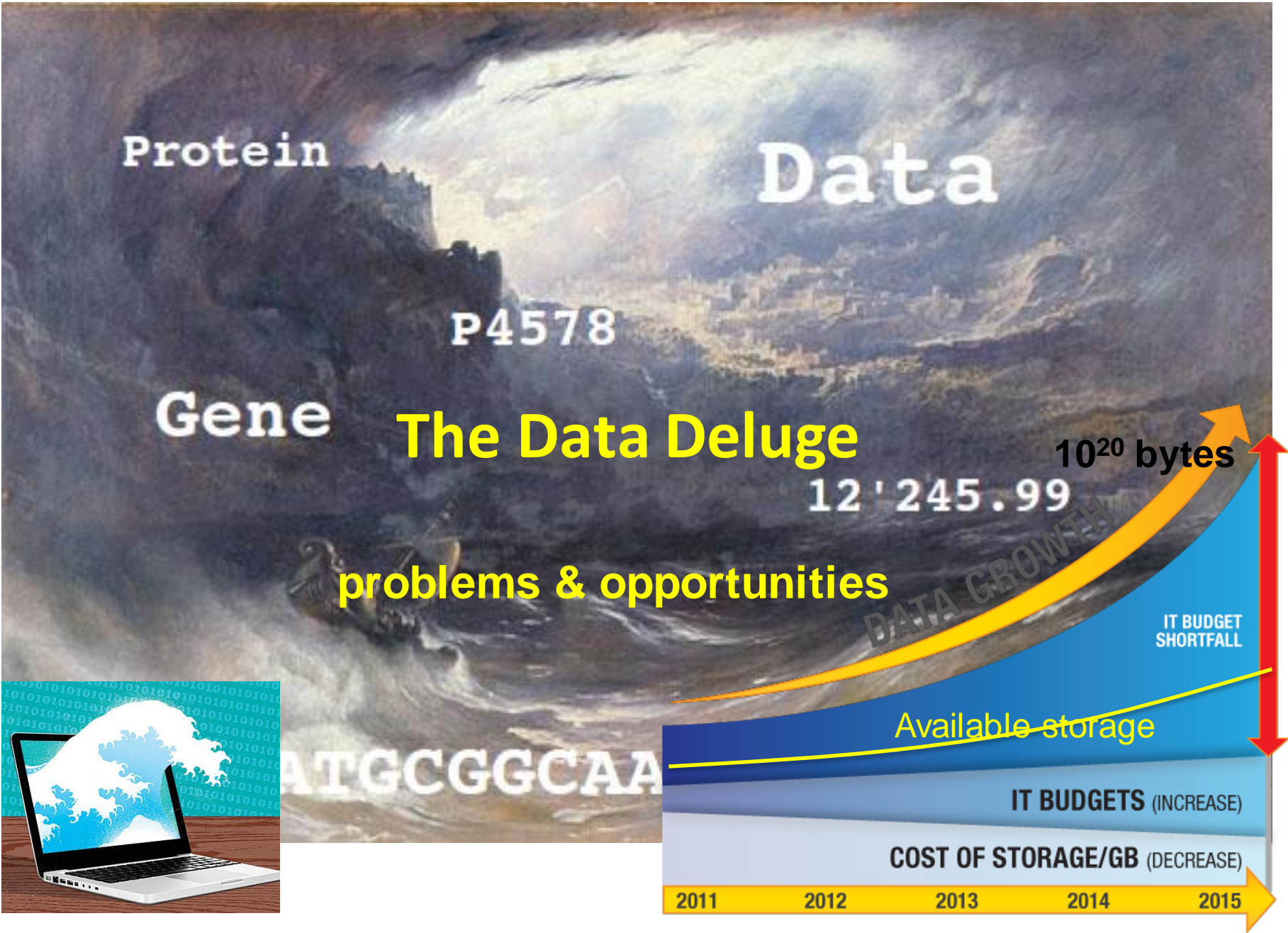
2012

2013

2014

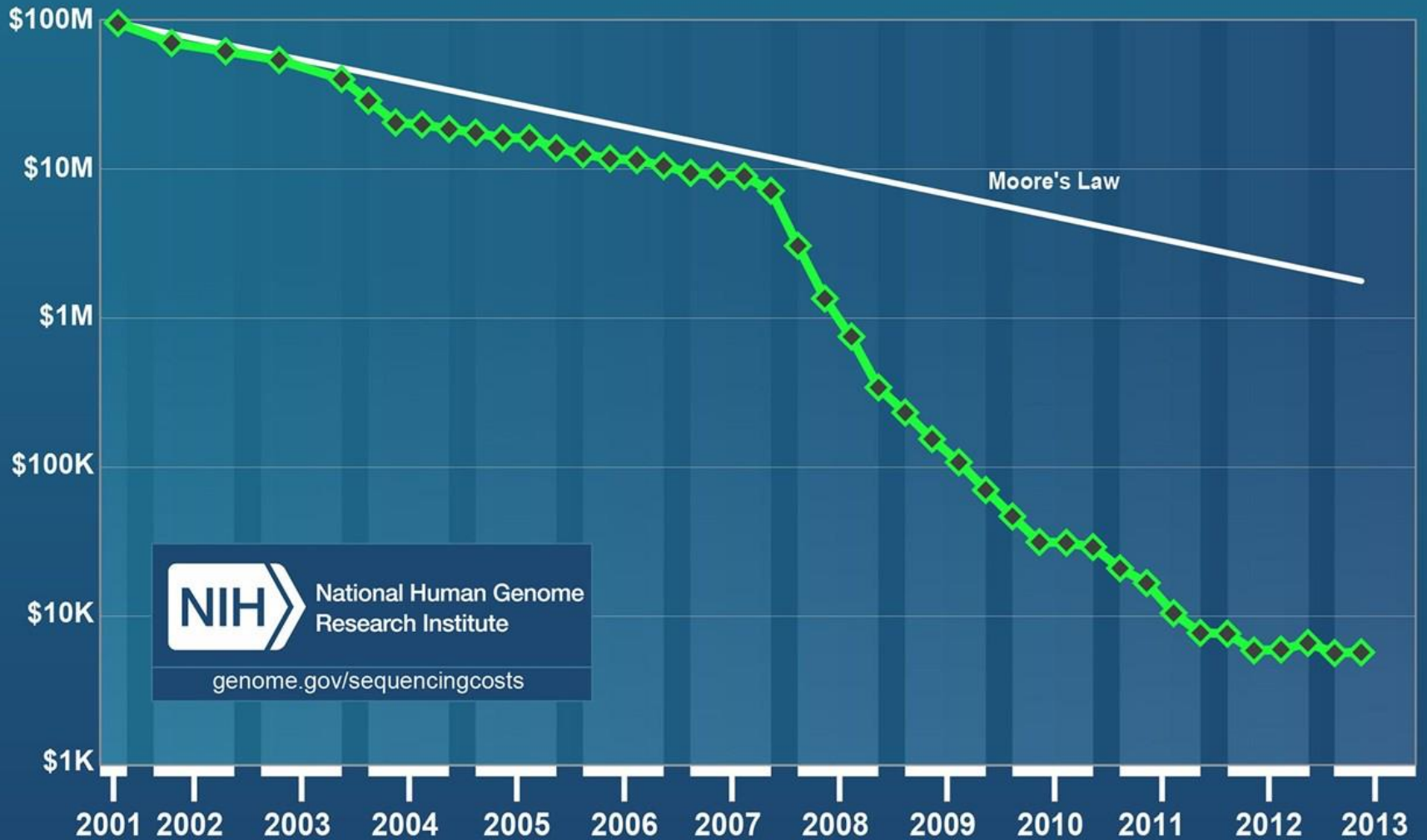
2015

ATGCGGCAA



# Data acquisition: Cost down – Flux up

## Cost per Genome



# A crisis of replicability and credibility?

**Pre-clinical oncology – 89% not reproducible**

NATURE | VOL 483 | 29 MARCH 2012

## REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.

\*Source of citations: Google Scholar, May 2011.

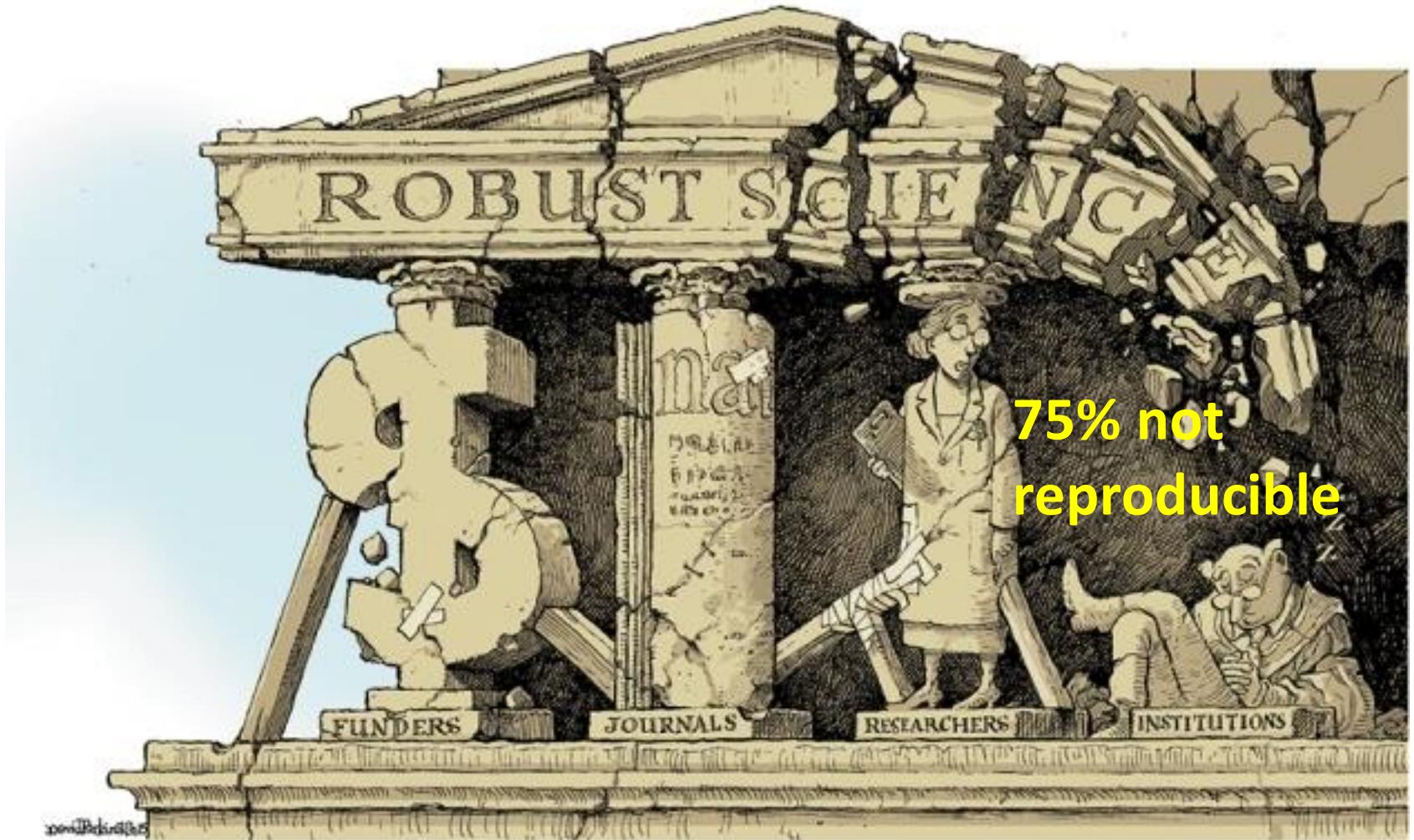
**A fundamental principle: the data providing the evidence for a published concept MUST be concurrently published, together with the metadata**

**To do otherwise should come to be regarded by all, including journals, as scientific MALPRACTICE**



## A crisis of credibility?

100 studies of social psychology in top ranking journals in 2008



**“Scientists like to think of science as self-correcting. To an alarming degree, it is not.”**





# The issue

## Non-reproducible results

### Why?

- **Fraudulent behaviour**
- **Invalid reasoning**
- **Absent or inadequate data and/or metadata**

### Cause?


- **Pressure to publish**
- **Pressure to make excessive claims**
- **Poor data science**

**The partial solution: Open data**

# Correction: intelligent openness

For effective communication, replication and re-purposing we need **intelligent openness**. Data, meta-data and, increasingly software/machine codes must be:

- **Discoverable**
- **Accessible**
- **Intelligible**
- **Assessable**
- **Re-usable**

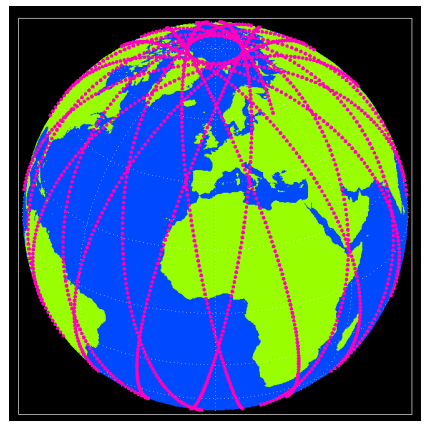
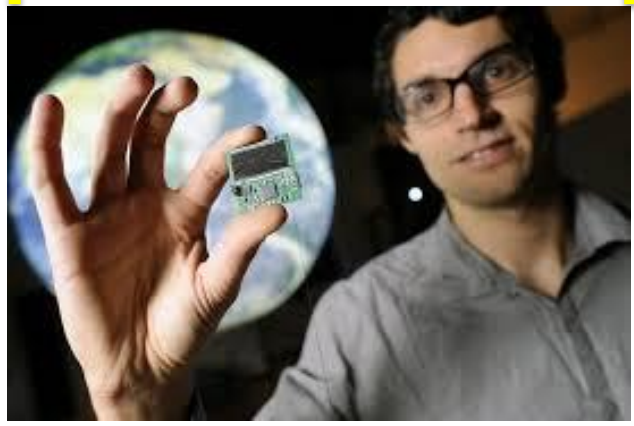
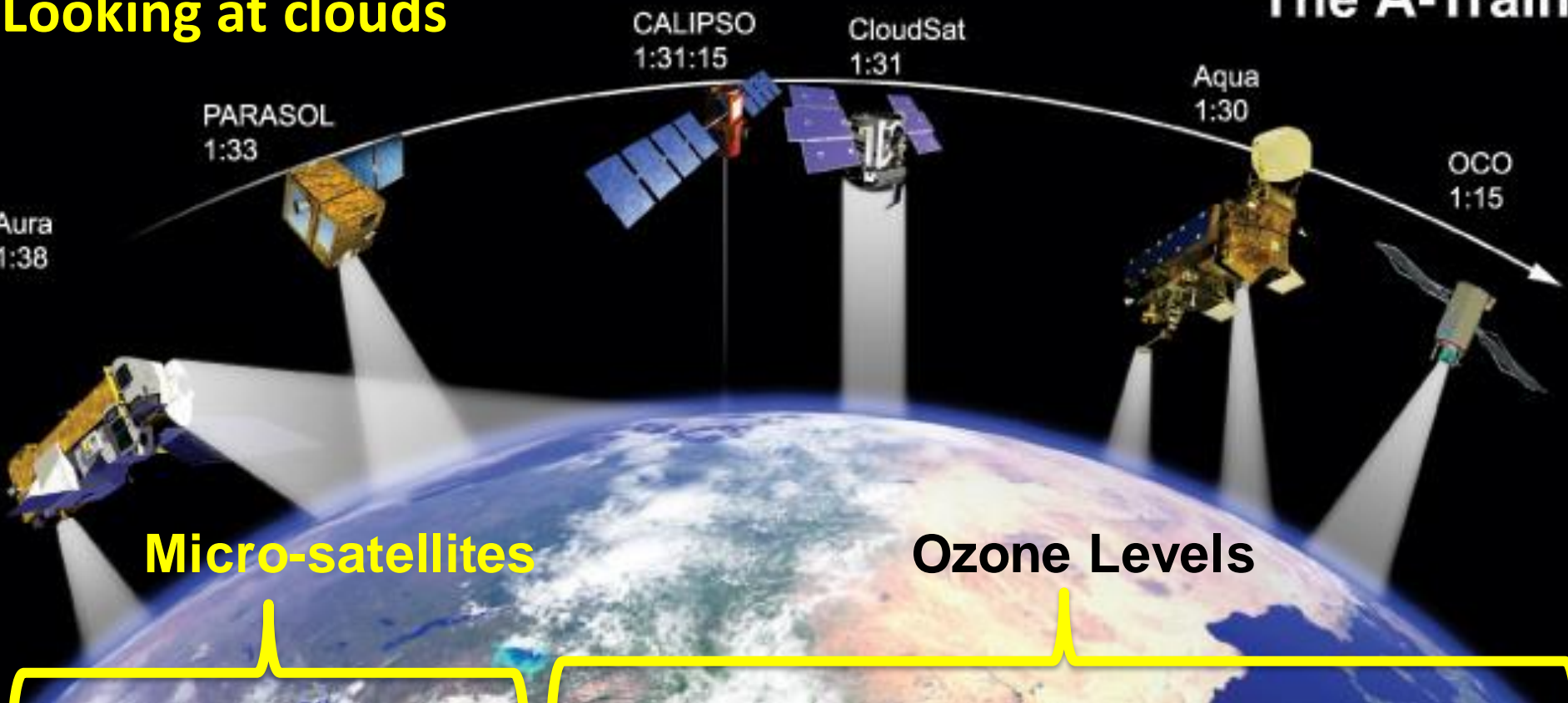


& not only the data & meta-data but also the software used to manipulate it

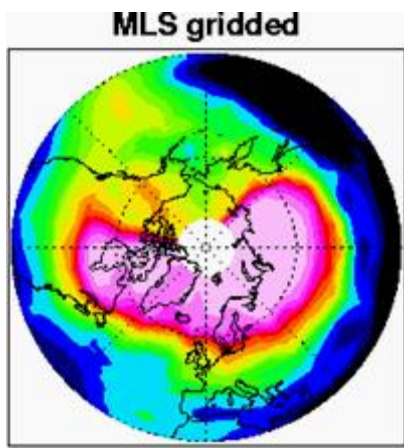
**Only when these criteria are fulfilled are data properly open.**

# Looking at clouds

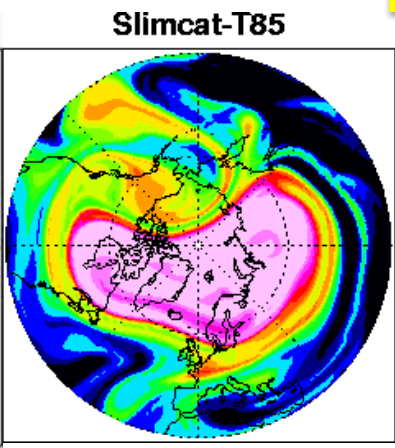
# The A-Train



Daily Trajectory



Observation



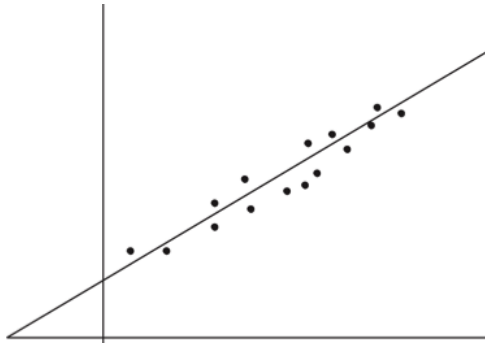
Model





# Valid reasoning from big/broad data

Simple relationships  
Classical statistics



Linear regression



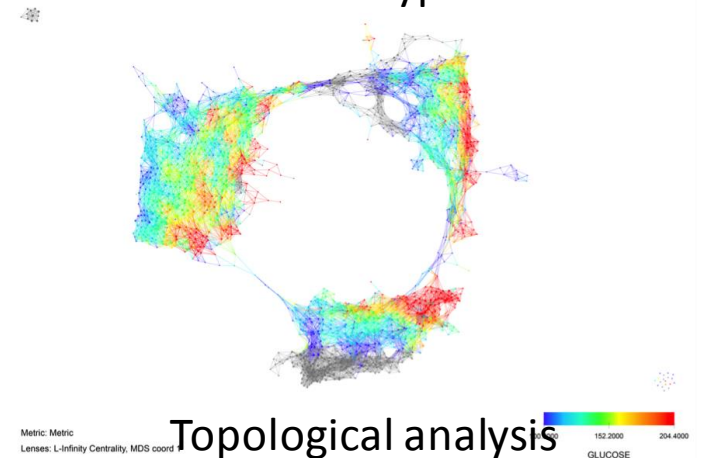
Cluster analysis

Complex systems  
No mathematical pipeline

Dynamic/complex behaviour



Glucose levels in Type II Diabetes



Topological analysis





## Self correction in science



The progress of science is strewn, like an ancient desert trail, with the bleached skeleton of discarded theories which once seemed to possess eternal life.

(Arthur Koestler)

# From “simple” science to complexity, from uncoupled to highly coupled systems

Uncoupled  
systems

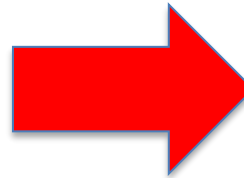
Slope



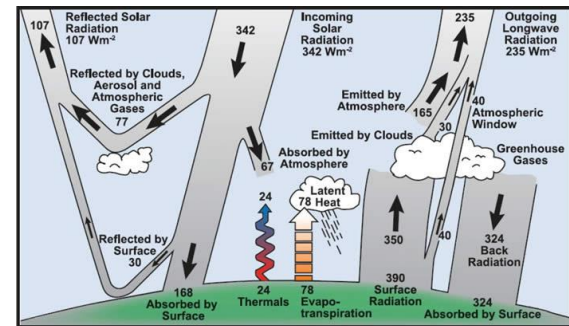
Power

Lever

Resistance

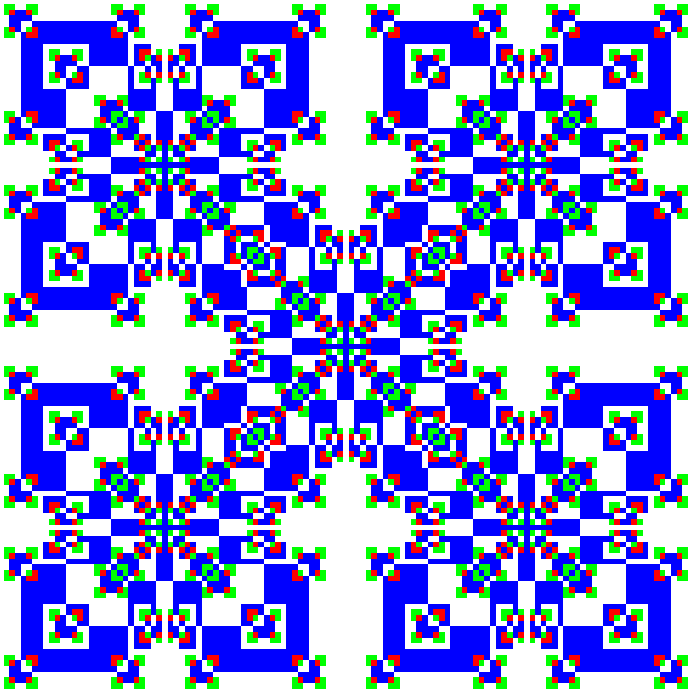


The behaviour of  
highly coupled systems



# Complex systems

## Simulating a complex system



Emergent behaviour of a specific 6-component coupled system

## Characterising a complex system

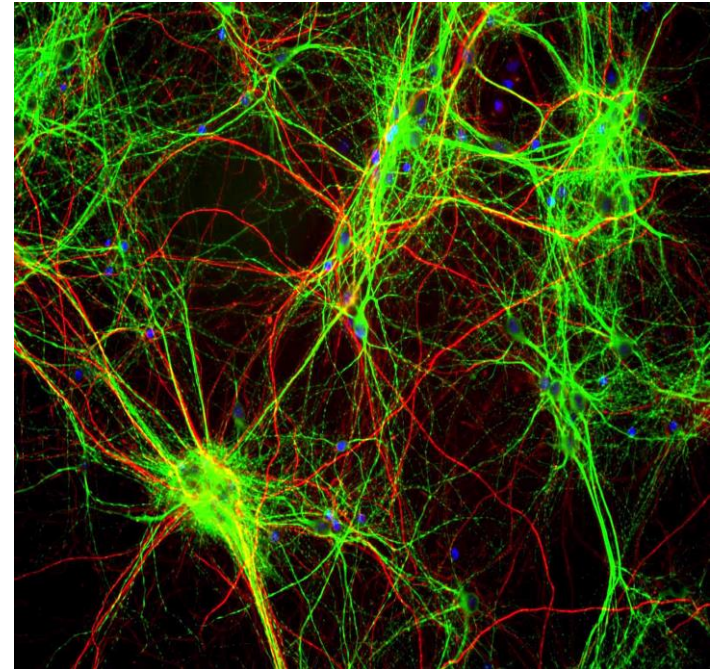
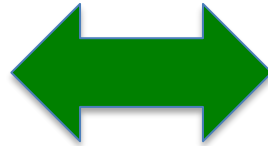


Image of brain cells in a rat







# Open data

## Why?

- To link and integrate data in identify patterns in phenomena
- To scrutinise the logic of an argument and maintain “self correction”

## Major Problems?

- Standards for reproducibility
- Machine Learning & the Human/machine interface
- Valid reasoning
- National Science Systems


## The Biggest Issue?

- Science as a public enterprise, or the privatisation of knowledge?

# Correction: intelligent openness

For effective communication, replication and re-purposing we need **intelligent openness**. Data, meta-data and, increasingly software/machine codes must be:

- **Discoverable**
- **Accessible**
- **Intelligible**
- **Assessable**
- **Re-usable**



& not only the data & meta-data but also the software used to manipulate it

**Only when these criteria are fulfilled are data properly open.**