Science as an Open Enterprise

Geoffrey Boulton
(Royal Society, University of Edinburgh)

Open Aire – Feb 2013

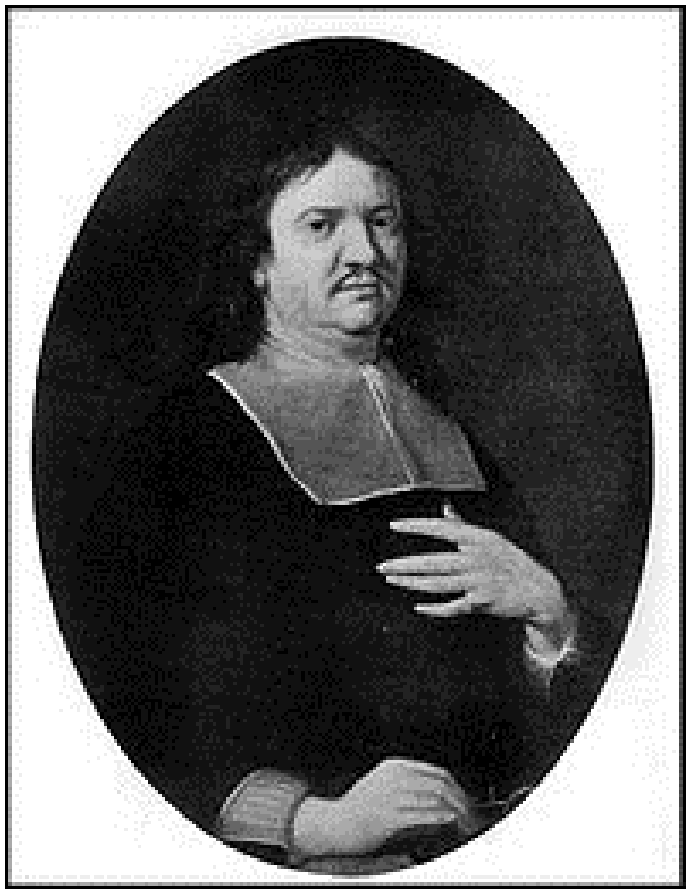
Report:

Report: www.royalsociety.org

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

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



Henry Oldenburg

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For *Anno* 1665, and 1666.

In the *SAVOY*,
Printed by *T. N.* for *John Martyn* at the Bell, a little without *Temple-Bar*, and *James Allestry* in *Duck-Lane*,
Printers to the *Royal Society*.

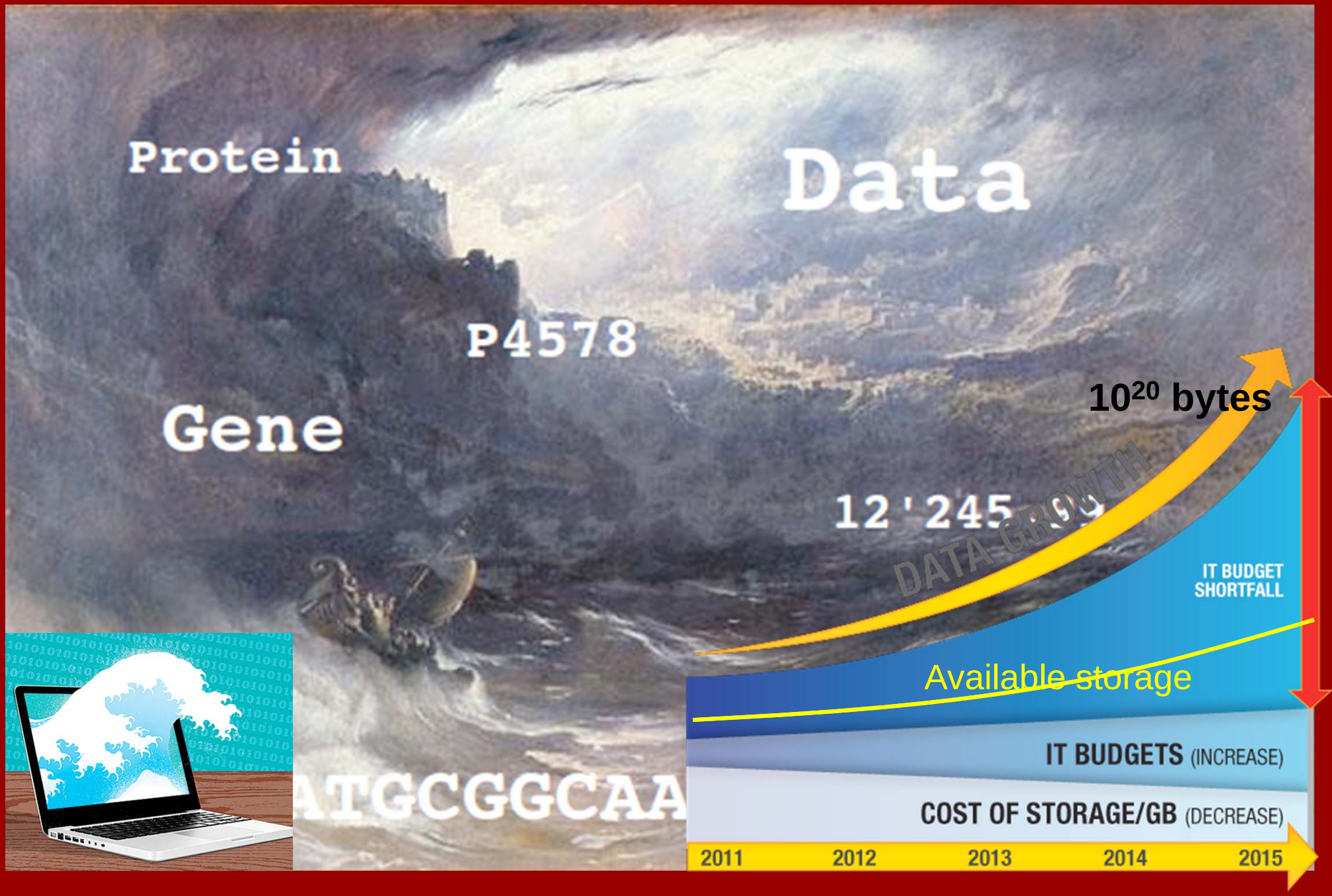
EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

"It is therefore thought fit ***to employ the [printing] press****, as the most proper way to gratify those [who] . . . delight in the advancement of Learning and profitable Discoveries [and who are] invited and encouraged to search, try,* ***and find out new things, impart their knowledge to one another, and contribute what they can to the Grand Design of improving Natural Knowledge*** *. . . for the Glory of God . . .* ***and the Universal Good of Mankind****."*

…. how do we achieve these ends in the post-Gutenberg era, when massive digital acquisition and cyber space have replaced the printing press?

Problems & opportunities in the data deluge



The challenges & opportunities?

- **Closing the concept-data gap – maintaining scientific self-correction & credibility**
- Exploiting the data deluge & computational potential
- Combating fraud
- Addressing planetary challenges
- Supporting citizen science
- Responding to citizens' demands for evidence
- Restraining the "Database State"

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

A crisis of replicability …… and of the credibility of science?

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.

The data providing the evidence for a published concept MUST be concurrently published, together with the metadata

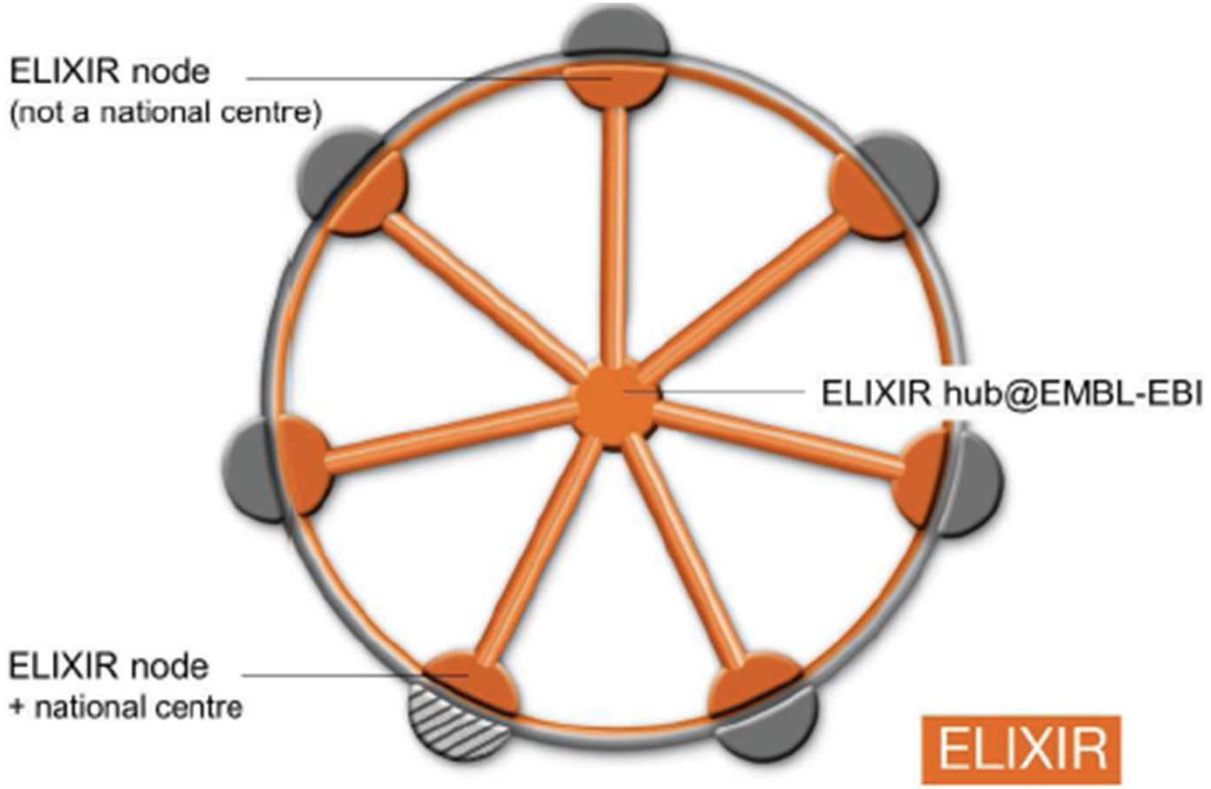
EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Challenges & opportunities?

- Closing the concept-data gap – maintaining scientific self-correction & credibility
- **Exploiting the data deluge & computational potential – data sharing**
- Combating fraud
- Addressing planetary challenges
- Supporting citizen science
- Responding to citizens' demands for evidence
- Restraining the "Database State"

Proven benefit so that data sharing becomes embedded in ethos & practice – bio-informatics



ELIXIR Hub (European Bioinformatic Institute) and ELIXIR Nodes provide infrastructure for data, computing, tools, standards and training.

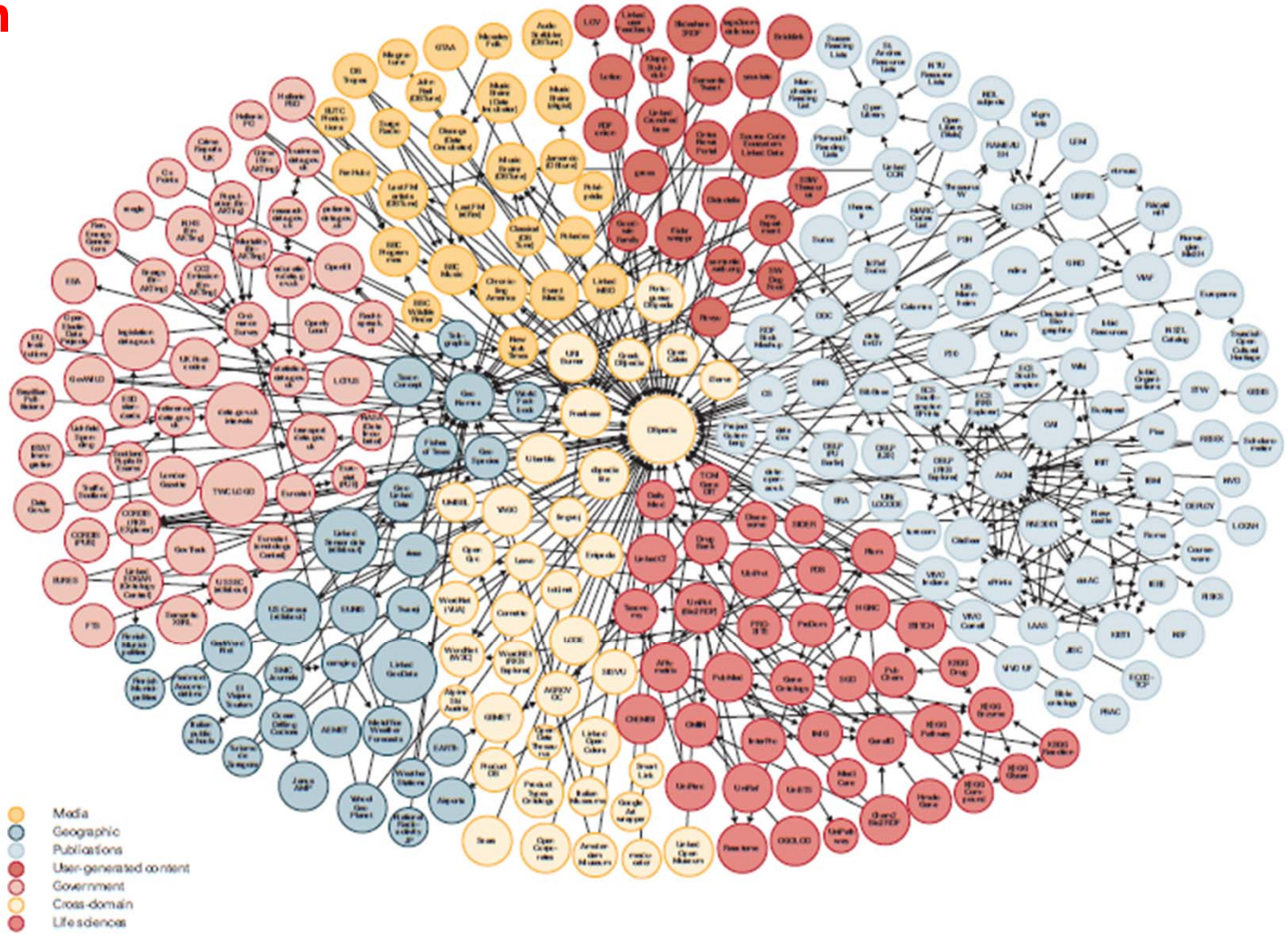
EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

New scientific knowledge from data

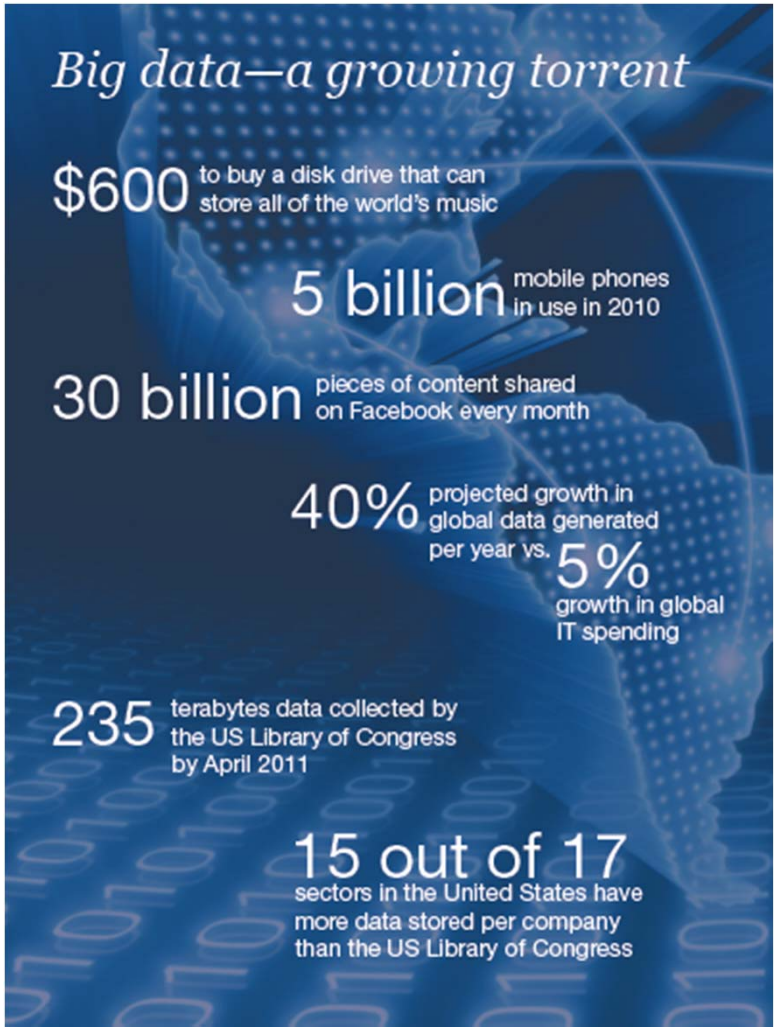
E.g. the potential of linked data

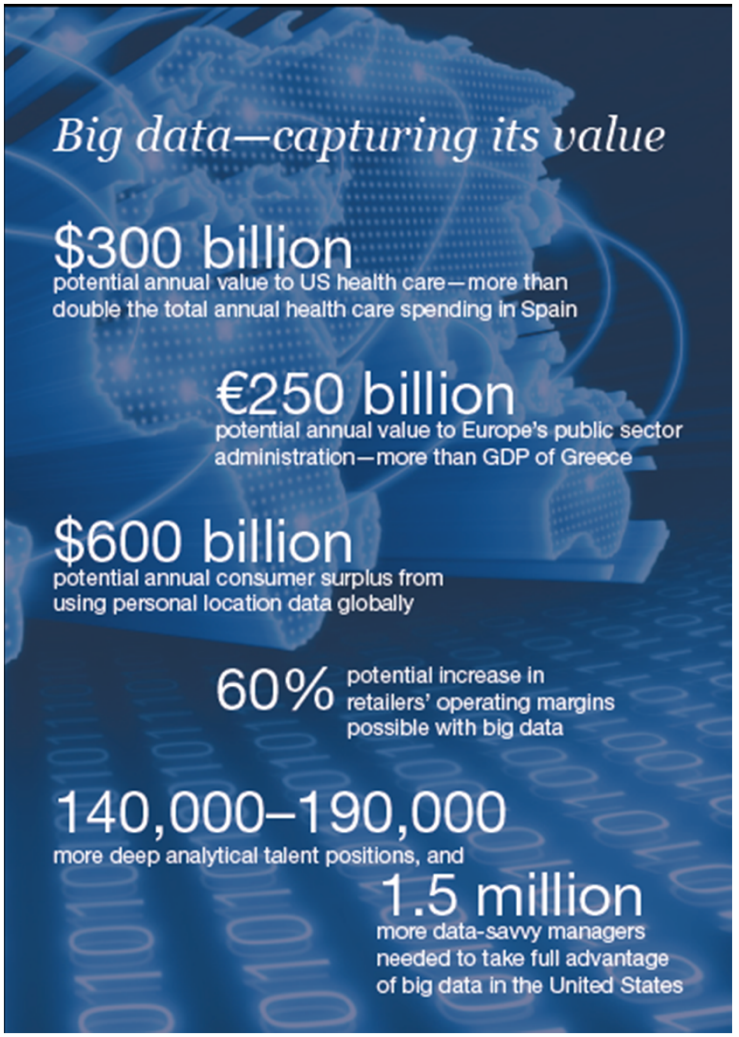
- data integration
- dynamic data



EXCELLENCE IN SCIENCE

.... and the economic implications





EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Its not just curation, retrieving and integrating data – its also what we do with it!

Jim Gray - "When you go and look at what scientists are doing, day in and day out, in terms of data analysis, it is truly dreadful. We are embarrassed by our data!"

- **Looking for inherent patterns – not just the expected / hoped for**
- **Partial reporting of data (cherry-picking) is scientific malpractice**
- **The role of Bayesian logic**

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Challenges & opportunities?

- Closing the concept-data gap – maintaining scientific self-correction & credibility
- Exploiting the data deluge & computational potential
- **Combating fraud**
- Addressing planetary challenges
- Supporting citizen science
- Responding to citizens' demands for evidence
- Restraining the "Database State"

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

theguardian NOTES&THEORIES
DISPATCHES FROM THE SCIENCE DESK

"Scientific fraud is rife: it's time to stand up for good science"

"Science is broken"

Examples:

- psychology academics making up data,
- anaesthesiologist Yoshitaka Fujii with 172 faked articles
- *Nature* - rise in biomedical retraction rates overtakes rise in published papers

Cause:

Rewards and pressures promote extreme behaviours, and normalise malpractice (e.g. selective publication of positive novel findings)

Cures:

Open data for replication

Transparent peer review

Not just personal integrity – but system integrity

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Challenges & opportunities?

- Closing the concept-data gap – maintaining scientific self-correction & credibility
- Maintaining the credibility of science
- Exploiting the data deluge & computational potential
- Combating fraud
- **Addressing planetary challenges**
- Responding to citizens' demands for evidence
- Supporting citizen science
- Restraining the "Database State"

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Why is open data an urgent issue?

- Closing the concept-data gap
- Maintaining the credibility of science
- Exploiting the data deluge & computational potential
- Combating fraud
- Addressing planetary challenges
- **Responding to citizens' demands for evidence**
- Supporting citizen science
- Restraining the "Database State"

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

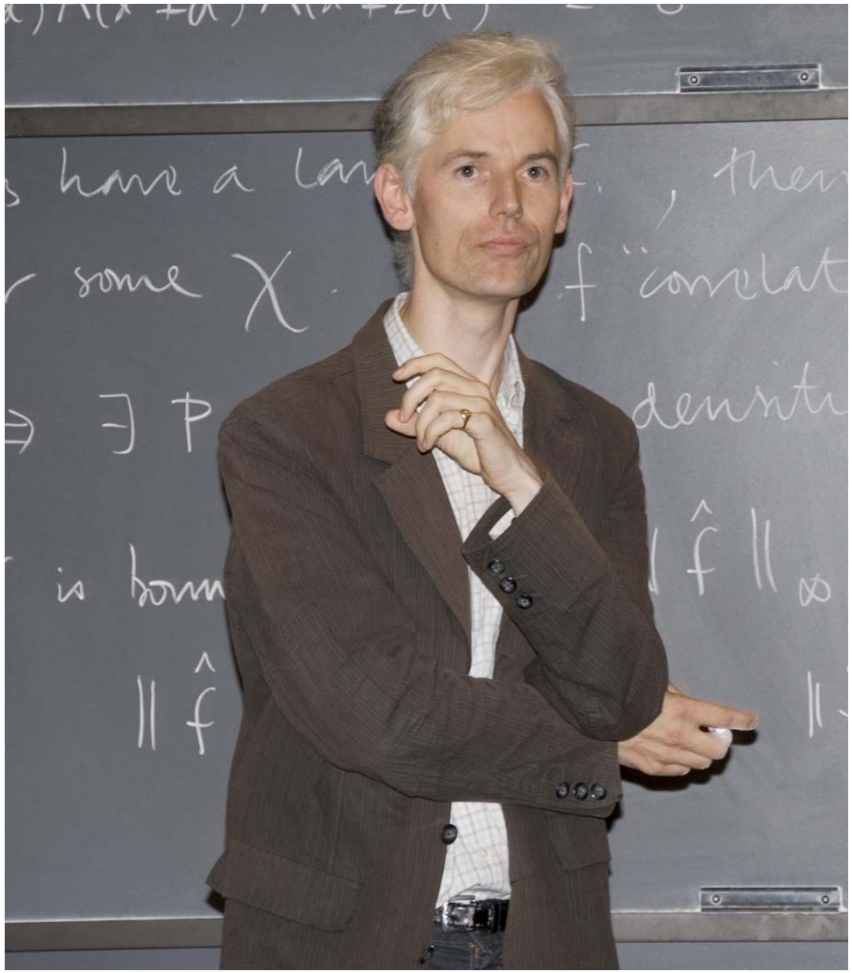
Why is open data an urgent issue?

- Closing the concept-data gap
- Maintaining the credibility of science
- Exploiting the data deluge & computational potential
- Combating fraud
- Addressing planetary challenges
- Responding to citizens' demands for evidence
- **Supporting citizen science – the 2030 question**
- Restraining the "Database State"

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Opening-up science:
e.g. crowd-sourcing



Tim Gowers
- crowd-sourced mathematics

An unsolved problem posed on his blog.

32 days – 27 people – 800 substantive contributions

Emerging contributions rapidly developed or discarded

Problem solved!

"Its like driving a car whilst normal research is like pushing it"

What inhibits such processes?
- The criteria for credit and promotion.

Why is open data an urgent issue?

- Closing the concept-data gap
- Maintaining the credibility of science
- Exploiting the data deluge & computational potential
- Combating fraud
- Addressing planetary challenges
- Supporting citizen science
- Responding to citizens' demands for evidence
- **Restraining the "Database State"**

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Openness of data *per se* has no value. Open science is more than disclosure

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

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

Only when these four criteria are fulfilled are data properly open

Metadata must be audience-sensitive

Scientific data rarely fits neatly into an EXCEL spreadsheet!

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Which publicly funded data for what purpose?

Data supporting the argument of a published paper?

- simultaneous deposition of citable data

Why should other data be open?

- greater benefit to science
- its not “our” data

Who should it be intelligently open to?

- other scientists
- citizen scientists
- the wider public

The dilemma of choice

Contradictory injunctions

Pressure to:

- commercialise, or
- share, collaborate., disseminate

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Boundaries of openness?

Openness should be the default position, with proportional exceptions for:

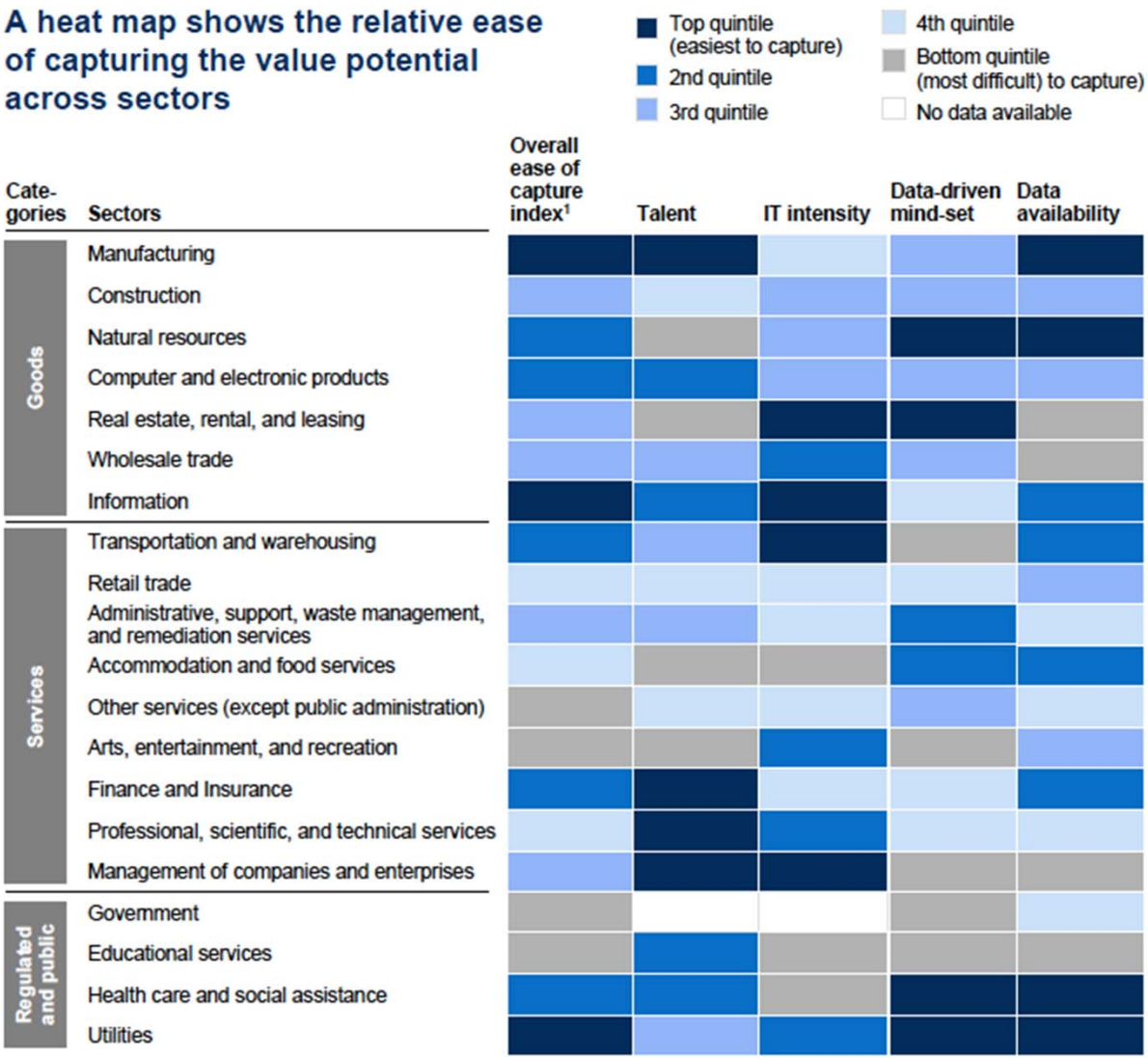
- **Legitimate commercial interests** (sectoral variation)
- **Privacy** (completely anonymised data is impossible)
- **Safety & security** (impacts contentious)

All these boundaries are fuzzy

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Commercial interests: potential by sector



1 See appendix for detailed definitions and metrics used for each of the criteria.

SOURCE: McKinsey Global Institute analysis

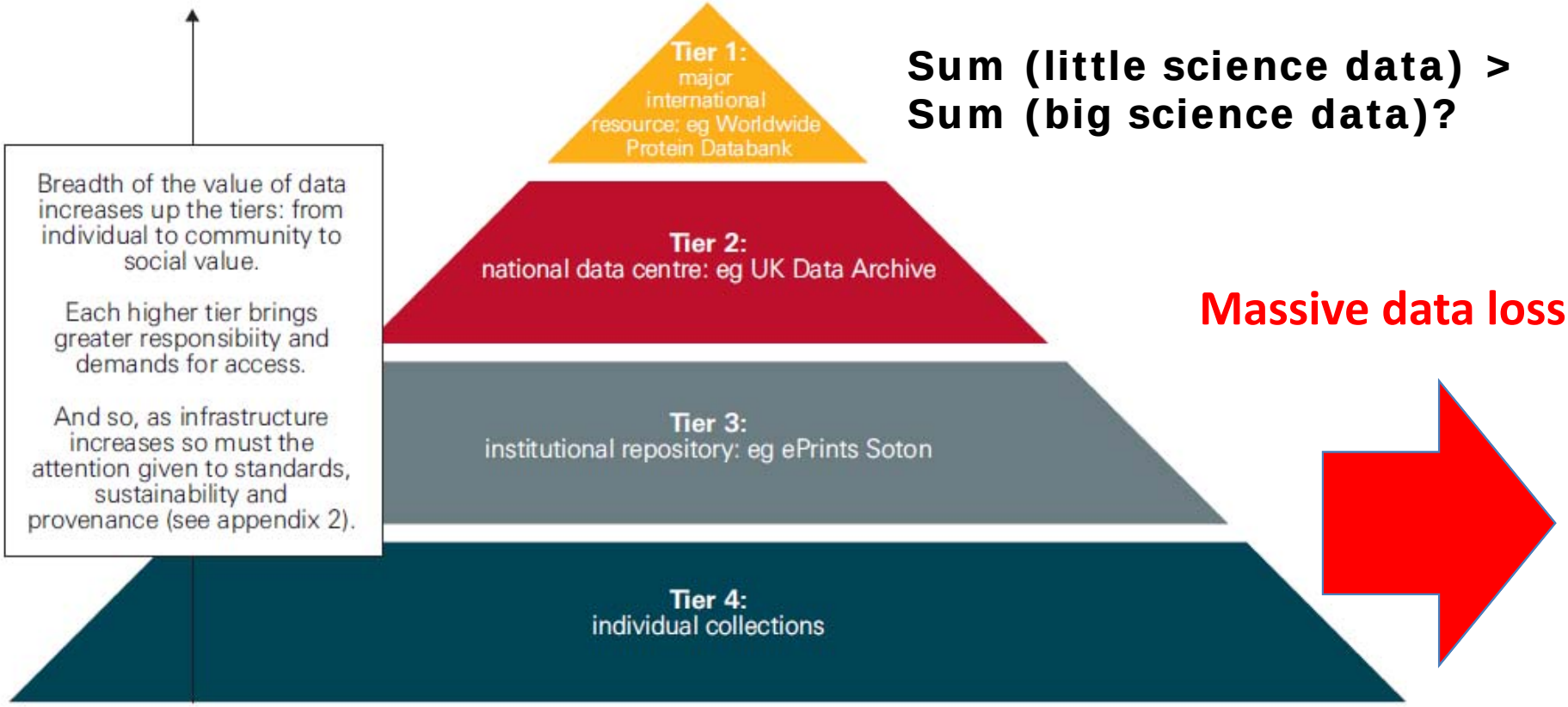
EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

A data management ecology?

The role of the top-down and the bottom-up?

Sum (little science data) > Sum (big science data)?



Massive data loss

Views of young scientists

- **The generation gap: younger researchers typically produce more data; recognise data sharing as maximising value; have most potential to develop data sharing tools; and they are the future. We should listen to them!**

1. **a shift away from a research culture where data is viewed as a private preserve**
2. **the data evidence for a published argument MUST be intelligently open at the time of publication**
3. **data management should be embedded in the community producing and using the data**
4. **science data should be as easy to "remix" as music is to a DJ**
5. **replication is by far the best guarantee of preservation (e.g. LOCKSS)**
6. **give credit for useful data communication and novel ways of collaborating**
7. **common standards for communicating data (correct?)**
8. **the cost of intelligent openness is an integral part of the cost of doing science**
9. **Training and support**

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Essential enabling tools & processes:

key issues for research & implementation

- data integration
- supporting dynamic data
- providing provenance
- annotation
- metadata generation
- citation
- access to data scientists
- changing the library

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Scripts for the actors in open science

Scientists – changing cultural assumptions Employers (universities/institutes) – data responsibilities; crediting researchers; the role of libraries Funders of research - the cost of curation is a cost of research Learned societies – influencing their communities	Publishers of research – mandating open data; open up to data mining; be careful not to be obstacles to the progress of science Business – exploiting the opportunity; awareness & skills Government – efficiency of the science base; exploiting its data Governance processes for privacy, safety, security - proportionality

Challenges for universities

- Will they rise to the scientific challenge, or leave things to the information business?
- Will they be responsible for the knowledge they create?
- The university library; doing the wrong things through the wrong people?
- Adapting scientific education?
- Training data scientists?
- Supporting the data manipulation needs of their researchers?
- Supporting intelligent openness
- Open data and commercial imperatives

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

The levels of influence

National

E.g. UK: Government "Transparency Boards" (Research, Business, Govt data) – chaired by Minister for Science

European

DGs Connect & Research

International

ICSU (International Scientific Unions)
CODATA

UK-US-Chinese-Indian science academies

BUT: the science community is the driver of creative, workable, flexible solutions – the roles of the above bodies are:

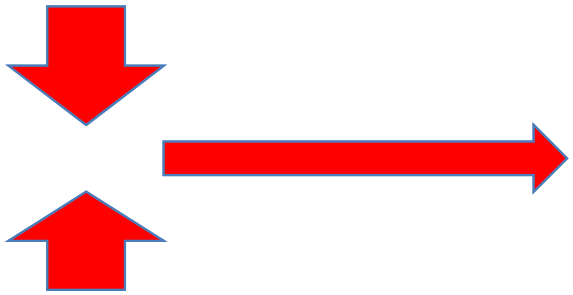
1. Remove barriers
2. Intelligent facilitation

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Challenge for the Commission as a funder of science

Top-down (present understanding)



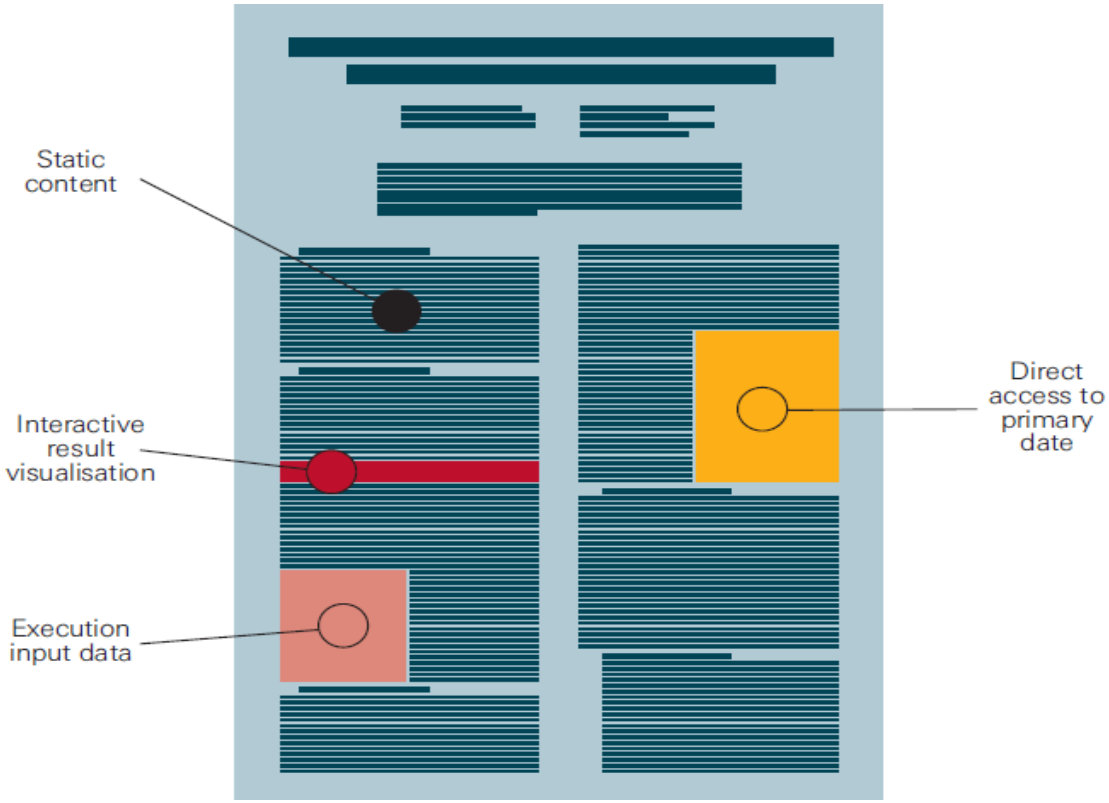
Bottom-up (new knowledge and experiment)

Digital European Research Area

1. Developing an open, interoperable e-infrastructure
2. Organising the European data space, through an open science policy
3. Opening communities, engaging individuals

… and remember - science is international!

A realiseable aspiration: all scientific literature online, all data online, and for them to interoperate



… and don't forget, this is a process, not an event!

EXCELLENCE IN SCIENCE

THE ROYAL SOCIETY

Report: www.royalsociety.org