SKA-Link kickoff

Combining knowledge to pioneer Big-Data solutions for SKA Data Centres

Instituto de Astrofísica de Andalucía (CSIC)

3rd and 4th April 2017

Overview of SKA-Link project

Aims of this meeting

Lourdes Verdes-Montenegro & AMIGA team

SKA-Link. IAA 3rd April 2017

Overview of SKA-Link project

Aims of this meeting

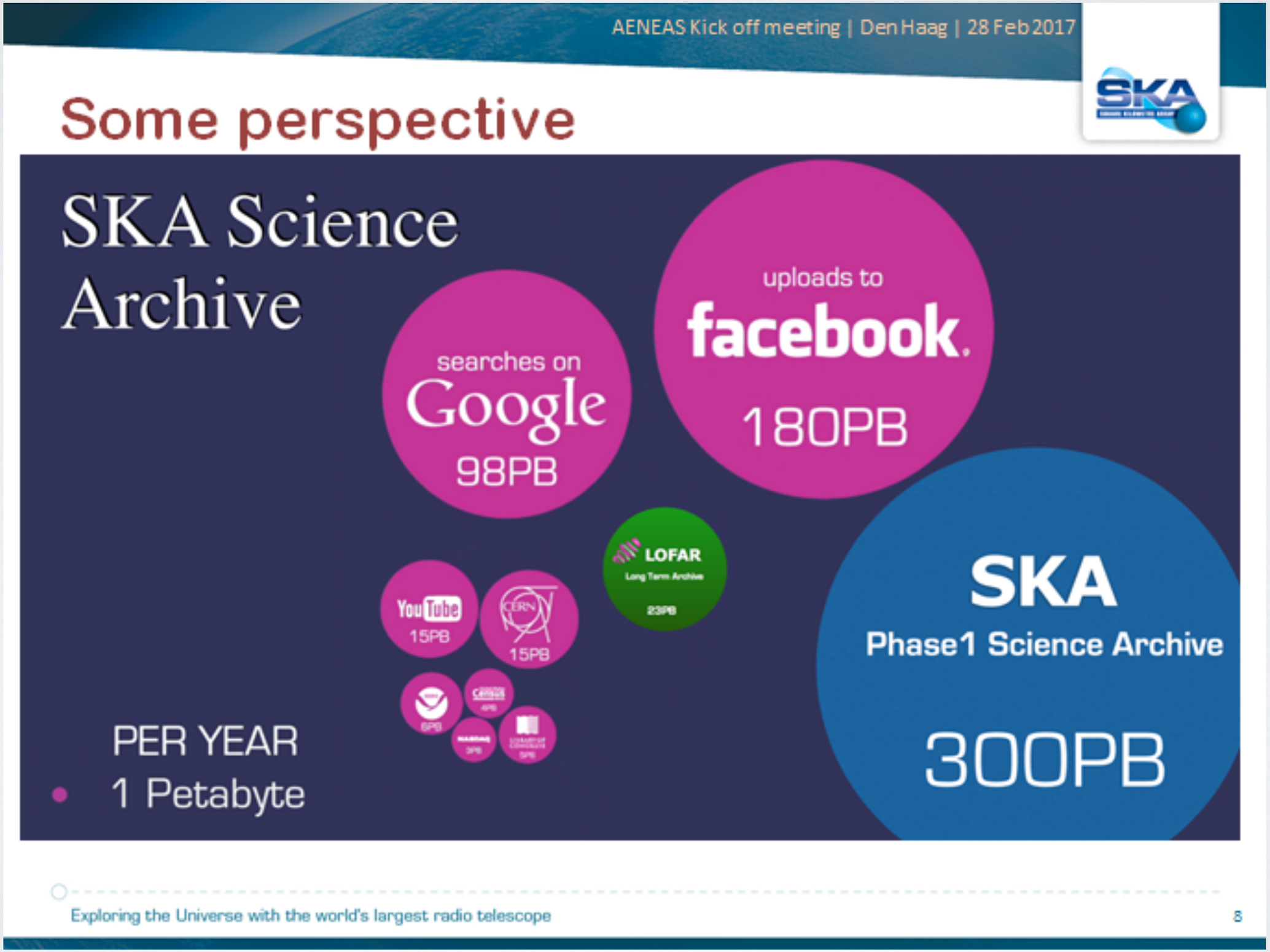
Lourdes Verdes-Montenegro & AMIGA team

- SKA-Link in a nutshell
- Why a special emphasis on the Scientific Method?
  - Reproducibility and metrics
- Some questions to address during the kick-off
- How to approach them? Agenda
- An opportunity for the SKA?
- In summary...

In Summary

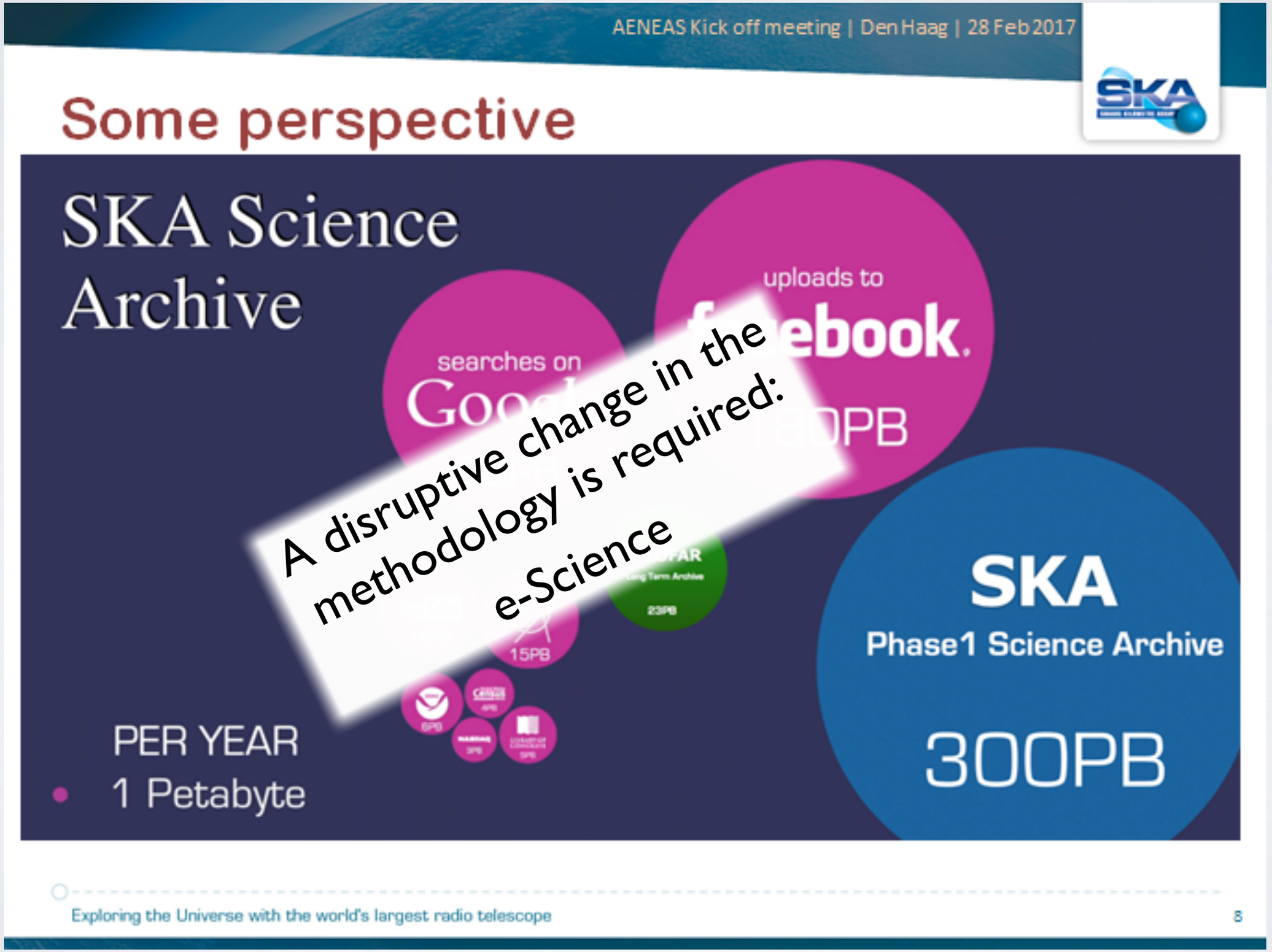
What Research does SKA want to do Tomorrow?

SKA: A BIG DATA INSTRUMENT



Slide from A. Chrysostomou's talk at AENEAS kickoff meeting

SKA: A BIG DATA INSTRUMENT



Slide from A. Chrysostomou's talk at AENEAS kickoff meeting

SKA-LINK IN A NUTSHELL

Extract scientific knowledge from such data deluge:

"*If there is a data deluge then there is also a deluge in the methods used to process it*"

De Roure & Goble 2010, Anchors in Shifting Sand: the Primacy of Method in the Web of Data

Computing / storage / network / human resources will be needed

- Data-intensive technologies for an efficient exploitation of DCIs
- Large international alliances of scientists to analyse such data deluge
  - Tools to enhance scientific collaboration
  - Platforms to share data, methods and knowledge

SKA-Link:

"Combining knowledge to pioneer Big-Data solutions for SKA Data Centres"

SKA-LINK IN A NUTSHELL

- General Aims
  - Technical strategies for succesfully exploiting the science-ready SKA data deluge
  - A set of Best Practices to be considered in the design of the SKA Regional Centres

SKA-LINK IN A NUTSHELL

- **General Aims**
  - Technical strategies for succesfully exploiting the science-ready SKA data deluge
  - A set of Best Practices to be considered in the design of the SKA Regional Centres

- **How: collaboration among**
  - Members of the Science Data Processor (SDP) consortium
  - Experts involved in the design of the SKA Regional Centres
  - Specialists on e-Science technologies for the scientific exploitation of Distributed Computing Infrastructures (DCIs)

SKA-LINK IN A NUTSHELL

- **General Aims**

  - Technical strategies for succesfully exploiting the science-ready SKA data deluge

  - A set of Best Practices to be considered in the design of the SKA Regional Centres

- **How: collaboration among**

  - Members of the Science Data Processor (SDP) consortium

  - Experts involved in the design of the SKA Regional Centres

  - Specialists on e-Science technologies for the scientific exploitation of Distributed Computing Infrastructures (DCIs)

- **What to deliver**

  - Inventory of data-intensive technologies and assessment on combinations of technologies supporting advances in the scientific methods.

  - Set of Best practices for the SKA to be considered a **reference (Metrics)** not only in science and technology, but **in scientific methodology**

ANY PROBLEM WITH THE SCIENTIFIC METHOD??

- Reproducibility is a principle of the Scientific Method (1660s)

- "*Although it was once thought that computers would improve reproducibility [...], most software tools do not provide mechanisms to package a computational analysis such that it can be easily shared and reproduced*" Dudley & Butte 2010, Reproducible in silico research in the era of cloud computing

- "*As much as 50% of published studies, even those in top-tier academic journals, cannot be repeated with the same conclusions by an industrial lab*" [L. Osherovich, "Hedging against academic risk," Science-Business eXchange, vol. 4, no. 15, 2011]

ANY PROBLEM WITH THE SCIENTIFIC METHOD??

25 May 2016

- Questionnaire on reproducibility filled by 1500 scientists
- > 70% of researchers have tried and failed to reproduce another scientist's experiments
- > 50% have failed to reproduce their own experiments
  - Chemistry: 90% (60%)
  - Biology: 80% (60%)
  - Physics and engineering: 70% (50%)
  - Medicine: 70% (60%)
  - Earth and environment science: 60% (40%)

Any problem with the Scientific Method??

25 May 2016

- Questionnaire on reproducibility filled by 1500 scientists
- > 70% of researchers have tried and failed to reproduce another scientist's experiments
- > 50% have failed to reproduce their own experiments
  - Chemistry: 90% (60%)
  - Biology: 80% (60%)
  - Physics and engineering: 70% (50%)
  - Medicine: 70% (60%)
  - Earth and environment science: 60% (40%)

Ah! So you don't empathise?

ANY PROBLEM WITH THE SCIENTIFIC METHOD??

25 May 2016

- Questionnaire on reproducibility filled by 1500 scientists
- > 70% of researchers have tried and failed to reproduce another scientist's experiments
- > 50% have failed to reproduce their own experiments
  - Chemistry: 90% (60%)
  - Biology: 80% (60%)
  - Physics and engineering: 70% (50%)
  - Medicine: 70% (60%)
  - Earth and environment science: 60% (40%)

Overly Honest Method
@OverlyHonestly

Maybe with this?

You can download our code from the URL supplied. Good luck downloading the only postdoc that can get it to run, though #OverlyHonestMethods

ANY PROBLEM WITH THE SCIENTIFIC METHOD??

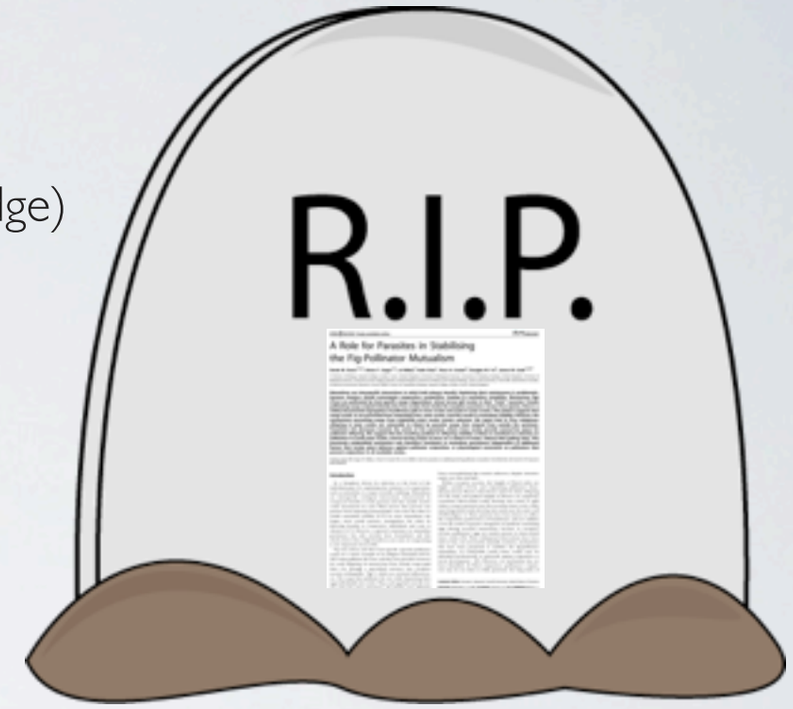
- **Knowledge Burying in paper publication**

(S. Bechhofer 2011, Research Objects: Towards Exchange and Reuse of Digital Knowledge)

- Publishing/mining cycle results in loss of knowledge

>= 40% of information lost

- **RIP**: **R**est **I**n **P**aper



http://www.clipartkid.com/rip-cliparts/

- *"The academic paper is now obsolescent...*

... as the fundamental sharable description of a piece of research. In the future we will be sharing **some other form of scholarly artefact,** something which is digital and designed for **reuse and to drop easily into the tooling of e-Research**, [...]

These could be called Knowledge Objects or Publication Objects or whatever: I shall refer to them as **Research Objects, because they capture research** "

(De Roure 2009, Director of Oxford's e-Research Centre)

ANY PROBLEM WITH THE SCIENTIFIC METHOD??

- **Knowledge Burying in paper publication**

(S. Bechhofer 2011, Research Objects: Towards Exchange and Reuse of Digital Knowledge)

- Publishing/mining cycle results in loss of knowledge

>= 40% of information lost

- **RIP**: **R**est **I**n **P**aper

http://www.clipartkid.com/rip-cliparts/

- *"The academic paper is now obsol*

Moving from narratives (last 300 yrs) to the actual output of research

... as the fundamental [illegible] of a piece of research. In the future we will be sha[illegible] **of scholarly artefact,** something which is digital and designed for reuse **and to drop easily into the tooling of e-Research**, [...]

These could be called Knowledge Objects or Publication Objects or whatever: I shall refer to them as **Research Objects, because they capture research** "

(De Roure 2009, Director of Oxford's e-Research Centre)

METRICS

nature International weekly journal of science

SPECIALS

See all spec

How to improve the use of metrics

Nature **465**, 870–872 (17 June 2010) | doi:10.1038/465870a

... "Science is being killed by numerical ranking,"[...] Ranking systems lures scientists into pursuing high rankings first and good science second.

SCIENCE METRICS

The value of scientific output is often measured, to rank one nation against another, allocate funds between universities, or even grant or deny tenure. Scientometricians have devised a multitude of 'metrics' to help in these rankings. Do they work? Are they fair? Are they over-used? *Nature* investigates.

▼ Editorial ▼ Features ▼ Opinion ▼ From the archive

EDITORIAL

Assessing assessment

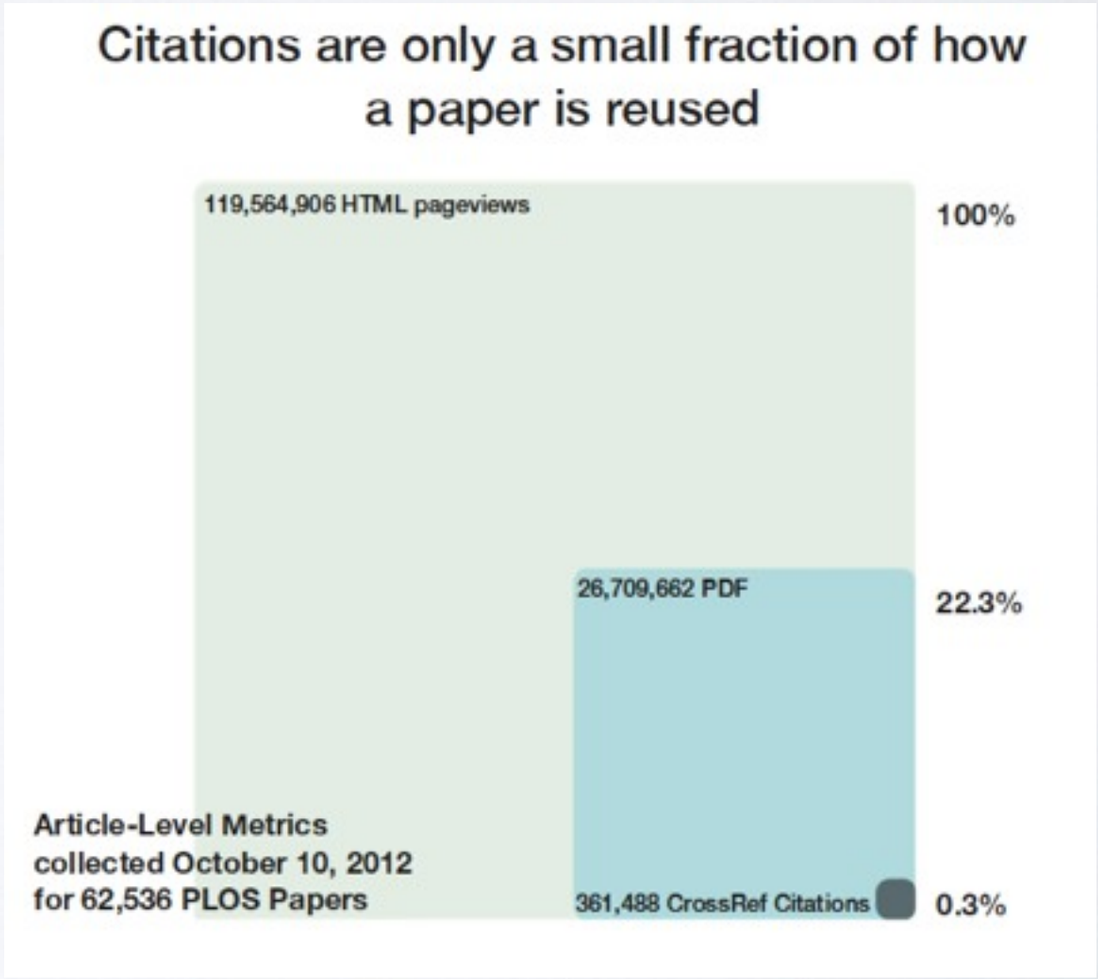
Transparency, education and communication are key to ensuring that appropriate metrics are used to measure individual scientific achievement.

METRICS

•"Within **a culture that pressures scientists to produce rather than discover, the outcome is a biased and impoverished science** in which most published results are either unconfirmed genuine discoveries or unchallenged fallacies. This observation implies **no moral judgement of scientists,** who are **as much victims of this system as they are perpetrators."**
(Chambers et al 2014, Instead of playing the game it is time to change the rules)

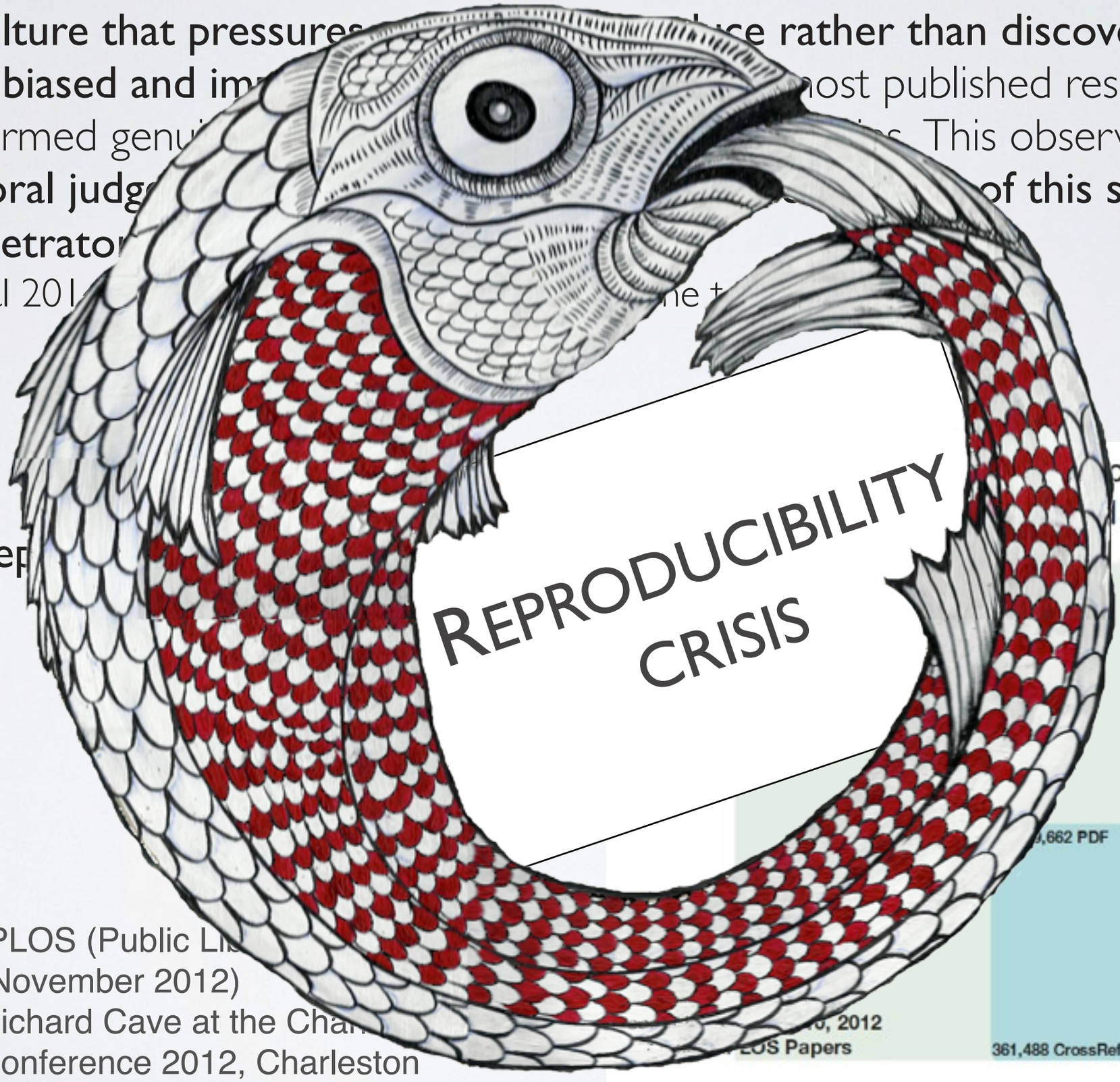
• Citations represent less than 1% of usage for an article

PLOS (Public Library of Science) (November 2012)
Richard Cave at the Charleston Conference 2012, Charleston



METRICS

•"Within **a culture that pressures** [illegible] **ce rather than discover, the outcome is a biased and im** [illegible] **most** published results are either unconfirmed genu[illegible]s. This observation implies **no moral judg**[illegible] **of this system as they are perpetrator**[illegible]
(Chambers et al 201[illegible] the t[illegible]



• Citations rep[illegible]
for an article

PLOS (Public Li[illegible]
(November 2012)
Richard Cave at the Cha[illegible]
Conference 2012, Charleston

SOME QUESTIONS TO ADDRESS DURING THE KICK-OFF

- **From the Science Data Processor (SDP) side:**
  - Can any of the tools developed by other communities be helpful for the SDP?
  - Does reproducibility requires any modification to the SDP?
  - How much effort would it require?
- **From the SKA Regional Centres (SRCs) side:**
  - What technologies can support a common interface to data and tools accross SRCs?
  - What do we understand by Methods? Is it just the software? Is it the same as a workflow?
  - How reproducibility translates into requirements for SRCs?
  - What tools are already available, and which aspects would require significant innovation?
- **From a prototype of an SRC (IDIA):**
  - What e-Science technologies are being considered?
  - Did it imply an extra effort/cost to introduce reproducibility?
- **From the non-SKA community**
  - How can I benefit and how can I contribute to an SRC?
- **As a scientist**
  - What metrics of success of SKA could better benefit Science?

SOME QUESTIONS TO ADDRESS DURING THE KICK-OFF

From Philosophy of Science to Practice: We can see SKA-Link as a

"Feasibility study about Scientific Methods and Metrics for the SKA, applying e-Science technologies"

How to approach them? Agenda

- **SKA-Link: combining knowledge to pioneer Big-Data solutions for SKA Data Centres**
  - Session 1: Project presentation and group introductions
- **To know about the regional network model for provision of SKA science data**
  - Session 2: The SKA Regional Centres
- **To make an inventory of technologies enabling scientists to exploit scientific data:**
  - Session 3: Technologies for the SRCs where those SKA-link groups experts on these technologies will describe **standards, protocols, tools** that support scientists to exploit large volume of data and that , at the same time, promote the collaboration and the knowledge sharing among the scientific community.
- **To asses combinations of those technologies supporting advances in the scientific methods:**
  - Session 4: Science Gateway examples enabling reproducible Science
- **About the Scientific Method, from theory to practice: technologies and methods**
  - Session 5: Reproducible science as a metric of SKA success

How to approach them? Agenda

- **End the kick-off with...**
  - A set of specific **questions** to be answered by SKA-Link
  - Proposed **ways** to approach them (documents, on-going projects/initiatives, etc)
  - Draft 0.0 of **outline of Best Practices document**
  - Associated **next actions** (and "who")
  - Revise/re-define proposed **collaboration** stays
  - Further funding opportunities

AN OPPORTUNITY FOR SKA?

Reproducible and sharable
Not only a nice idea: it is just happening

NSF example:
Chapter II.C.2.f(i)(c), Biographical Sketch(es), has been revised to rename the "Publications" section to "Products" and amend terminology and instructions accordingly. This change makes clear that products may include, but are not limited to, publications, data sets, software, patents, and copyrights.
To make it count, however, it needs to be both citable and accessible.

Amendment to 'guidance on submissions':
Following consultation on the draft panel criteria, the definitions at paragraphs 112-113 of 'guidance on submissions' have been amended, and are now superseded by paragraphs 43-44 as indicated below.

These changes have been made in response to concerns raised that the evolving nature of publication practices, such as online 'pre-publication', would have meant that some research outputs published near the boundary between the 2008 RAE and the 2014 REF publication periods may not in practice have been eligible for submission to either exercise.

Policies of the UK programme for assessing research quality, the Research Excellence Framework:

no grant-review sub-panel "will make any use of journal impact factors, rankings, lists or the perceived standing of publishers in assessing the quality of research outputs"

AN OPPORTUNITY FOR SKA?

Reproducible and sharable
Not only a nice idea: it is just happening

The Square Kilometre Array could

Be the first Mega-science Infrastructure taking the lead of trustable, reproducible science, going beyond numbers of papers/citations

Ignore it

AN OPPORTUNITY FOR SKA?

Reproducible and sharable
Not only a nice idea: it is just happening

The Square Kilometre Array could

Be the first Mega-science Infrastructure taking the lead of trustable, reproducible science, going beyond numbers of papers/citations

Ignore it

Difficult? yes!

But this word doesn't seem to intimidate those aiming to build the SKA

In Summary

What Research does SKA want to do Tomorrow?