

MRD: Gravitational Waves (and other big science data)

Norman Gray, Graham Woan and Tobia Carozzi
Physics and Astronomy, University of Glasgow, UK
I2S2 data challenges workshop, Leicester
2011 May 5

JISC



University
of Glasgow

- How do the GW communities manage their data now, and how representative of 'big science' is this?
- Do they have formal DM plans?
- Would they like one?
- How bad would it be to say 'read CCSDS 650.0'?
- STFC funds 'big science' – given that, what shape of DM plans should they be requiring bidders to propose?

- big money – decades, G€ / G\$
- big author lists – LIGO=0.8 kAuth; ATLAS=3 kAuth
- big admin – MOUs, councils, ...
- big careers – PhD to tenure
- big data – aLIGO \sim 1 PB/yr; ATLAS \sim 10 PB/yr (= '1 LHC')



astronomy data

- Babylonian data can be used for earth slowdown studies
- Plates are used for some astrometry
- Astronomers can (roughly) read 1627 Rudolphine tables
- ...and with help, 12C Toledan tables
- So, let's say a millennium

telescopes: large-scale facility

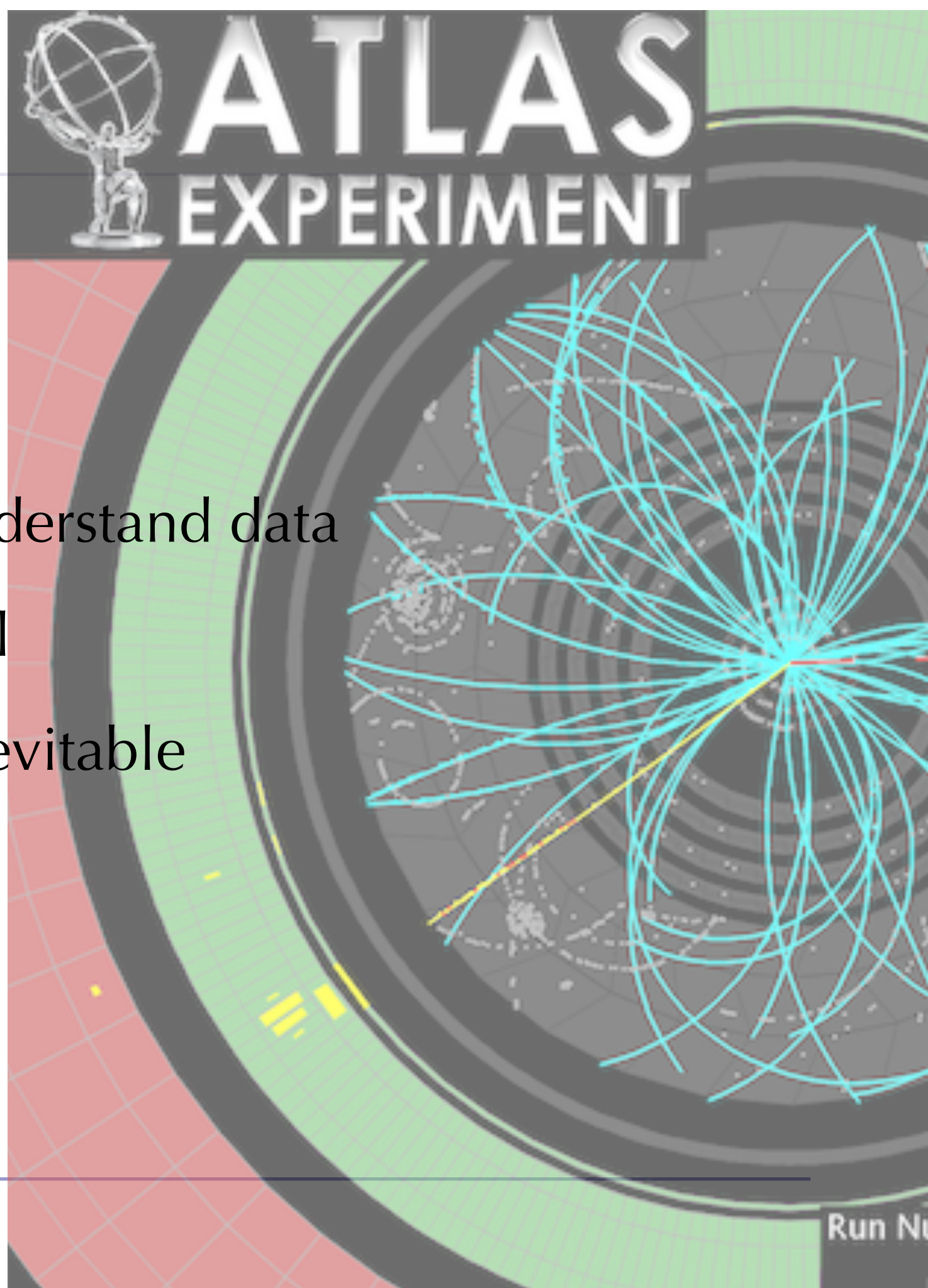
- observers bid for nights, possibly from a national allocation (similarly for satellites)
- telescopes & instruments work in 'visitor mode' or 'observer mode' (if it's cloudy, tough)
- data goes from the telescope directly into the archive, from where observers later retrieve it (ie very 'cloudy')
- proprietary period of, say, 12 months, after which it automatically becomes open

hep data

- Major challenge to understand data
- ...so software is crucial
- ...and supercession inevitable
- So, perhaps 30 years?

norman gray

Image from atlas.ch



Run N

detector: one-off experiment

- decadal commitment to designing, building and running the experiment and its software
- group/community decides on which measurements to take – no individual ownership
- data acquired directly into the archive
- analysed server-side; excerpts may be downloaded
- data is proprietary until discovery, public thereafter



gravitational waves

- Features of both astronomy and HEP
- No detection announced so far, but still \sim PB/yr
- Data reduction heavily dependent on software
- ...but the eventual data products will be intelligible

what does big science do
right?

I'm not saying 'this is *the* right way to do it', but 'these are features which work for this area, and might be helpful elsewhere'

caveat 2: big science has it easy

- One scale, few projects
- Well-resourced projects, with plenty of computing experience and lots of engineering management
- Historically large data volumes enforce data management discipline
- Shared instruments/facilities common
- Rarely (never?) commercially sensitive data

LIGO scores 3-ish on the AIDA benchmarks
(out of 5), without really trying.

(and it's clear it could move to 4s
or 5s without difficulty)

Good stuff: <http://aida.jiscinvolve.org/toolkit>

data sharing paradigms

All my data is my precioussss




VS



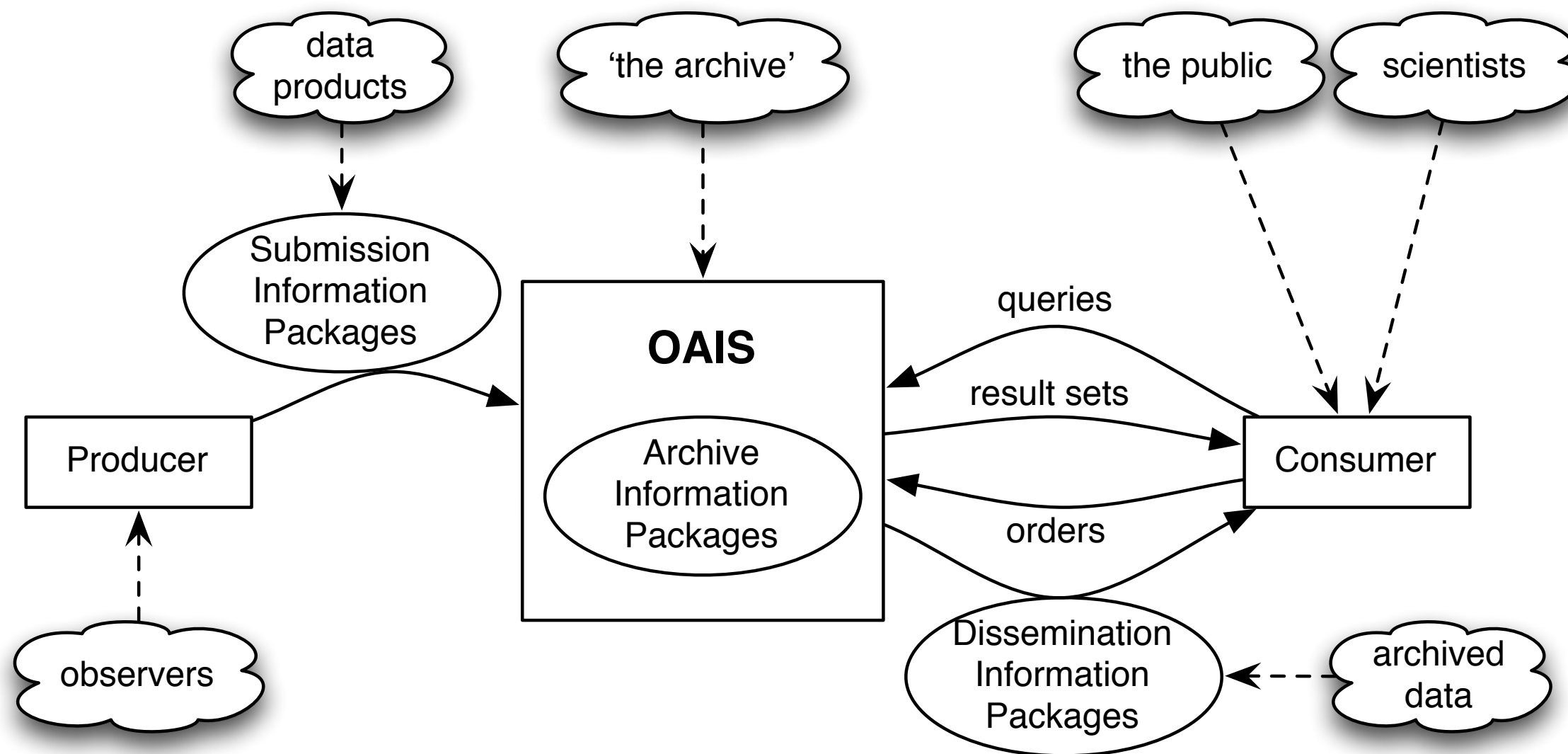
Information wants to be free!

norman gray

- 
- Predefined and documented data objects
 - Multiple levels: raw \rightarrow L1 \rightarrow ... L n
 - Obvious target of data management planning, sharing and access
 - Input to, and output from, data processing pipelines (useful for planning and design)
 - If used by data owners, as part of working archive, *preservation becomes a marginal cost*
 - 'how?' & 'whether?' \rightarrow 'what is the product?'

proprietary periods

- ‘proprietary’ = embargoed
- Payoff to originators, but ensures data ends up open
- Most big-science data has a proprietary period ranging from 6 to 36 months (usually 12–24); or other release algorithm
- Funders: ‘when?’ & ‘whether?’ → ‘how long?’
- ‘Data products’ are the natural object of such negotiations, as a function of novelty, complexity, policy, ...



so what is 'the right thing'?

Emergent:

- Formal & costed data management planning
- Identification of 'designated communities'
- Identification of data products (AIPs in OAIS-speak)
- Timescales and criteria for data release

Plus?

- Framed with OAIS conceptual model (formalised)
- ...so coupled with the OAIS validation industry

so our recommendations will be...

- Funders should simply require that a project develop a suitable profile of the OAIS specification – and then step back
- Funders should support projects in creating per-project OAIS profiles which are appropriate to the project and meet funders' strategic priorities and responsibilities
- STFC should develop a costings model for the publication and preservation of data, which is matched to the data challenges of big-science communities.

more generally...

The notions of *data products* and *proprietary periods* very naturally concretise otherwise diffuse arguments about data management and sharing, transforming them from ‘whether?’ and ‘why?’, to ‘which?’ and ‘how long?’.

<http://purl.org/nxg/projects/mrd-gw/report>

(<http://nxg.me.uk/temp/mrd-gw-final-8df5737c5802.pdf>)

Norman Gray
University of Glasgow
<http://nxg.me.uk>