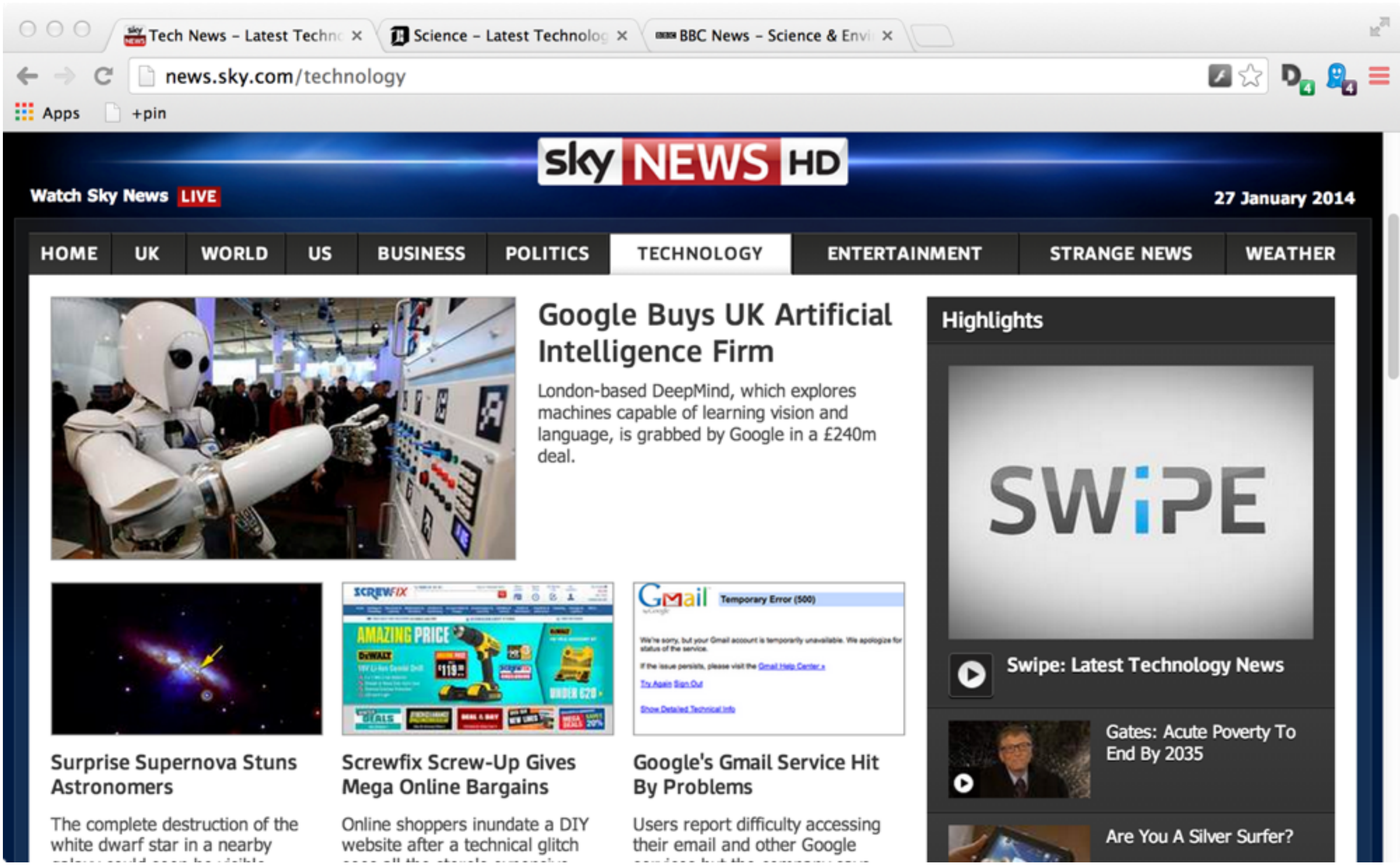
Big science; big science data

Norman Gray

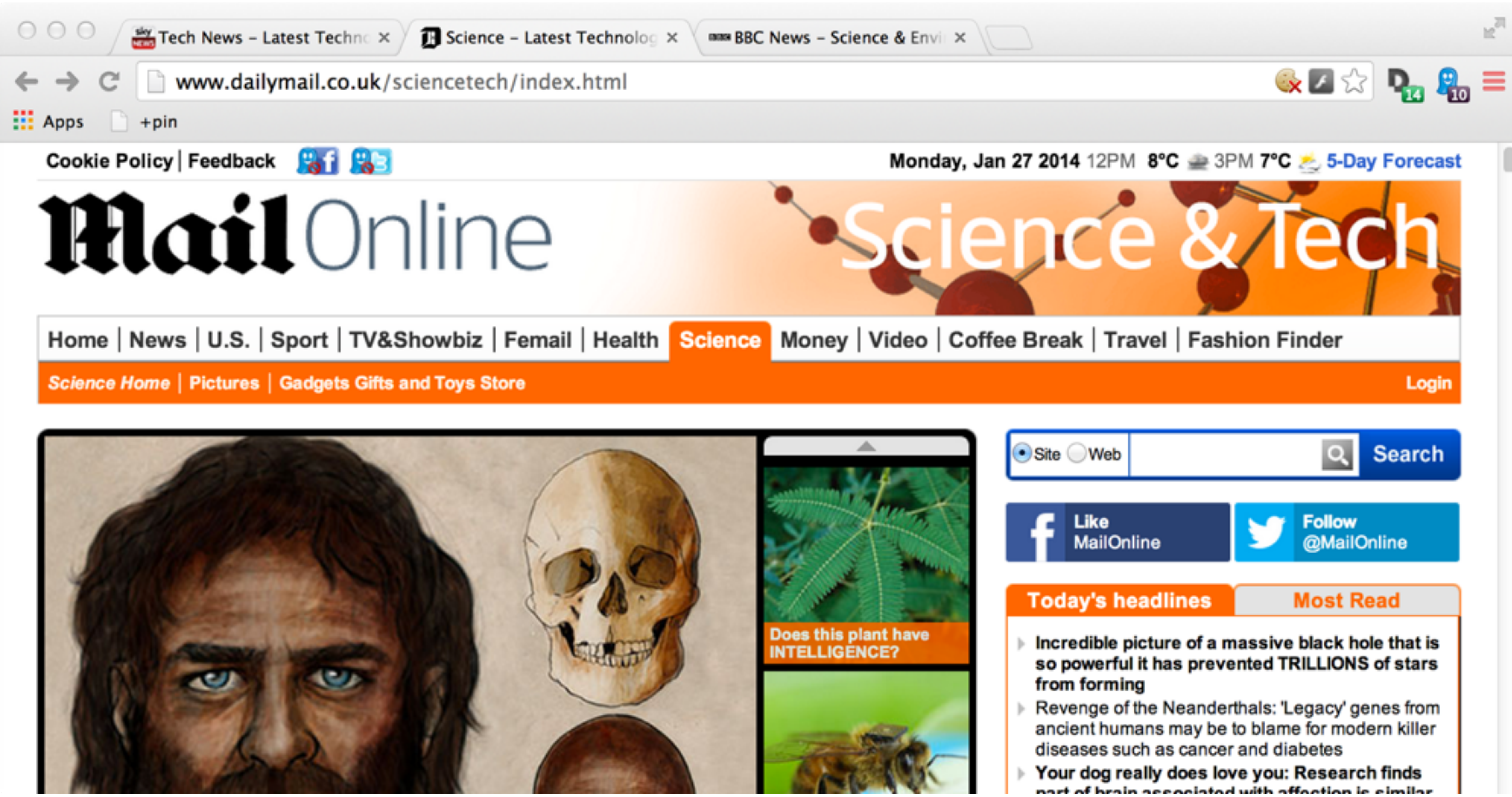
Highlights of Astronomy

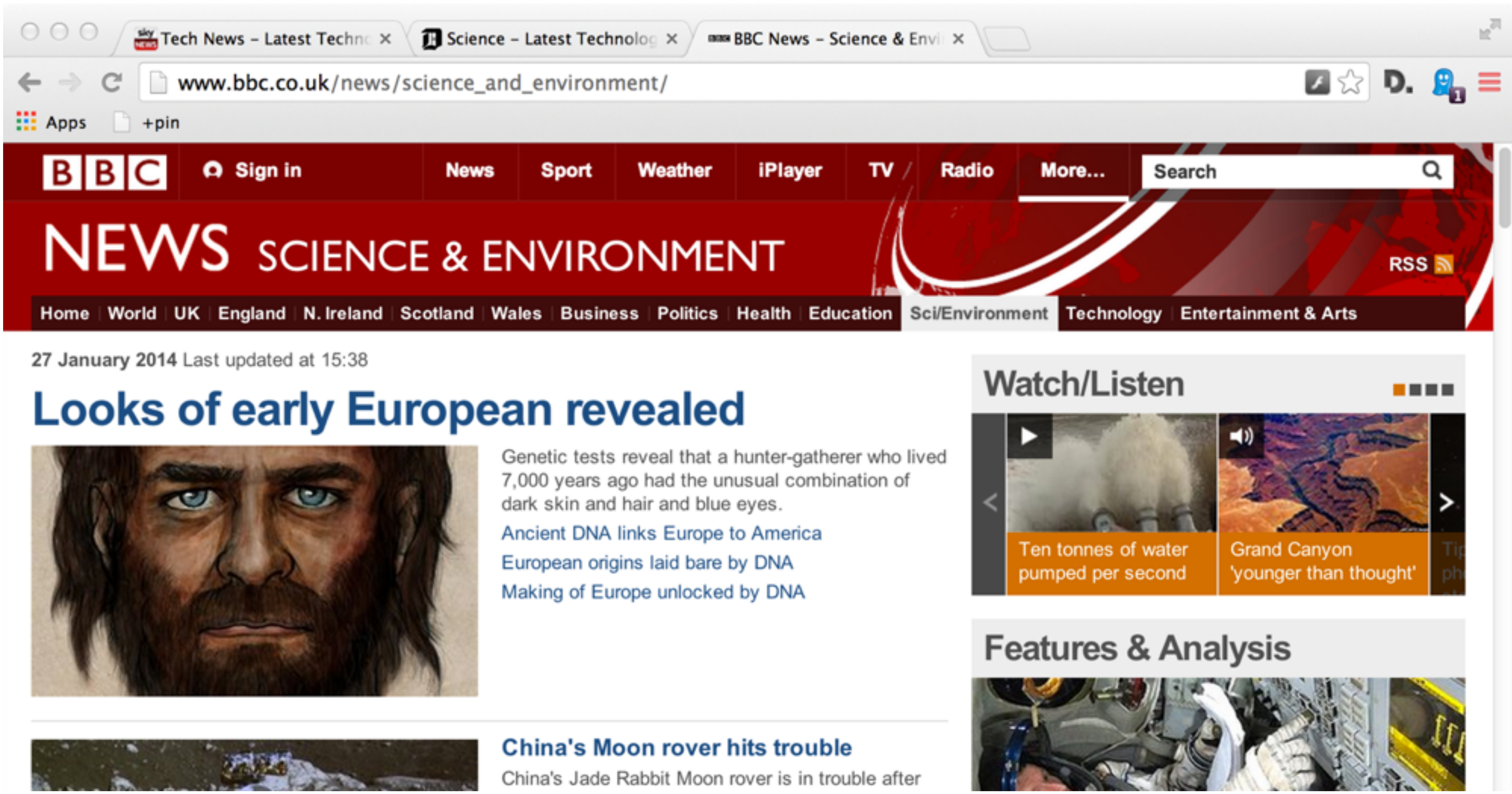
Glasgow University Centre for Open Studies

2014 January 27



What's wrong with this picture?
Science != Technology


Tech News – Latest Techno
Science – Latest Technolog
BBC News – Science & Envi
www.dailymail.co.uk/sciencetech/index.html
Apps
+pin
Cookie Policy | Feedback
Monday, Jan 27 2014 12PM 8°C 3PM 7°C 5-Day Forecast
Mail Online
Science & Tech
Home | News | U.S. | Sport | TV&Showbiz | Femail | Health | Science | Money | Video | Coffee Break | Travel | Fashion Finder
Science Home | Pictures | Gadgets Gifts and Toys Store
Login
Site Web
Search
Like MailOnline
Follow @MailOnline
Today's headlines
Most Read
Does this plant have INTELLIGENCE?
Incredible picture of a massive black hole that is so powerful it has prevented TRILLIONS of stars from forming
Revenge of the Neanderthals: 'Legacy' genes from ancient humans may be to blame for modern killer diseases such as cancer and diabetes
Your dog really does love you: Research finds



Gets it correct: science+environment
Several of the stories are actually science policy stories, or science+politics
How does science work, as a process?

How does science work?

Is ‘Big Science’ actually different?

Why is science funded?

What does ‘big data’ tell us?

1. People should know – it’s more complicated than it might seem.
2. Yes and no, but not really, deeply, I don’t think
3. Different answers for different people
4. Talking about ‘data’ is a route in to several of these questions.

Lots of potential questions here

astronomy has form here

Solar/lunar ephemeris for 104BCE March 23–101BCE April 10, generated 103BCE December 20

Information duplicated (backups, mirrors or refreshes?)
Have acquisition metadata from Babylon and Bloomsbury
Babylon data centre still working in 1st C CE, but few acquisitions due to funding cuts, in the ruins of a deserted city
Very little/compact Representation Information
Can be of some astronomical interest

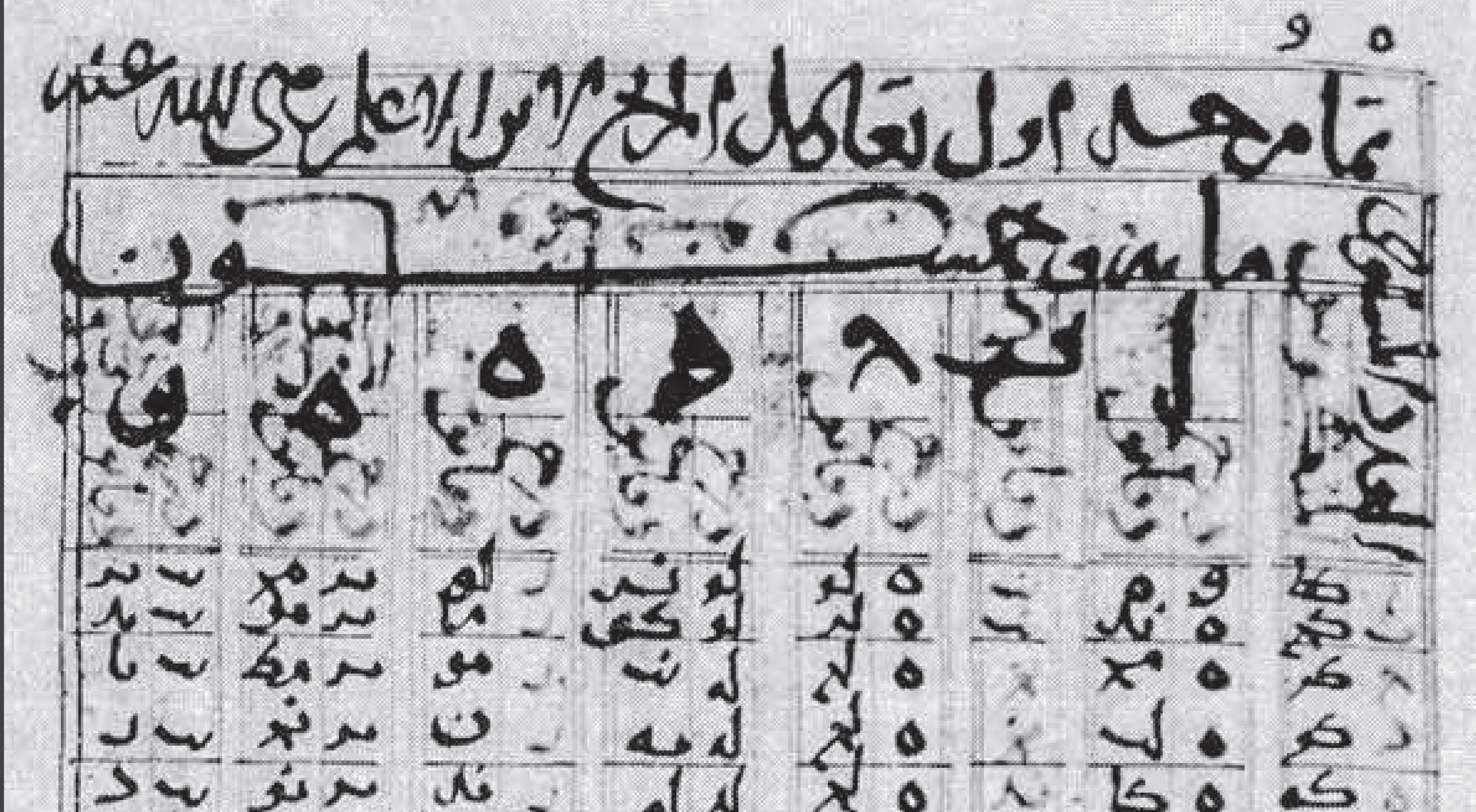


Table of equations for Mars, from al-Zij al-Mumtahan
(from Benno van Dalen, *A Second Manuscript of the Mumtahan Zîj*, Suhayl 4 (2004), pp. 9-44)

Another ephemeris, from Baghdad, this time (via Greece)
I’ll come back to this one later

64 *Tabularum Rudolphi*

Tabula Æquationum MARTIS.

Anomalia Eccentri, Cum æquatione parte physica	Intercolumnium, Cum Logarithmo.	Anomalia coæquata.	Intervallũ Cum Logarithmo	
120 4.35.50	8890 1. 5.36	115.17.11	145293 37357	25
121 4.33. 1	9190 1. 5.47	116.19.52	145080 37211	25
122 4.30. 6	9470 1. 5.59	117.22.39	144871 37067	25
123 4.27. 6	9730 1. 6.11	118.25.31	174663 36924	25
124 4.24. 2	10070 1. 6.22	119.28.29	144458 36782	25
125 4.20.53	10360 1. 6.34	120.31.33	144255 36642	24
126 4.17.40	10650 1. 6.46	121.34.42	144055 36503	24
127 4.14.22	10930 1. 6.57	122.37.56	143857 36365	24
128 4.10.59	11210 1. 7. 8	123.41.14	143661 36229	24
129 4. 7.31	11480 1. 7.19	124.44.37	143468 36095	24
130 4. 3.58	11740 1. 7.30	125.48. 6	143278 35962	24
131 4. 0.21	12000 1. 7.40	126.51.40	143091 35831	23
132 3.56.40	12260 1. 7.50	127.55.19	142906 35702	23
133 3.52.55	12510 1. 8. 1	128.59. 3	142724 35575	23
134 3.49. 6	12760 1. 8.11	130. 2.52	142545 35450	23
135 3.45.13	13000 1. 8.21	131. 6.45	142370 35327	22
136 3.41.16	13240 1. 8.30	132.10.43	142198 35206	22
137 3.37.15	13480 1. 8.40	133.14.46	142028 35086	22
138 3.33.10	13710 1. 8.49	134.18.53	141861 34968	21
139 3.29. 1	13940 1. 8.59	135.23. 4	141697 34852	21
140 3.24.48	14160 1. 9. 8	136.27.20	141537 34739	20
141 3.20.31	14380 1. 9.16	137.31.41	141381 34628	20
142 3.16.10	14600 1. 9.27	138.36. 6	141228 34520	19
143 3.11.46	14820 1. 9.35	139.40.34	141078 34414	19
144 3. 7.18	15030 1. 9.44	140.45. 7	140931 34310	18
145 3. 2.46	15240 1. 9.54	141.49.44	140788 34209	18
146 2.58.10	15450 1.10. 2	142.54.24	140649 34110	17
147 2.53.31	15650 1.10.10	143.59. 8	140513 34013	17
148 2.48.48	15850 1.10.19	145. 3.56	140381 33919	16
149 2.44. 2	16040 1.10.27	146. 8.48	140252 33827	16
150 2.39.14	16230 1.10.35	147.13.44	140127 33738	

Anomalia Eccentri, Cum æquatione parte physica	Intercolumnium, Cum Logarithmo.	Anomalia coæquata.	Intervallũ Cum Logarithmo	
150 2.39.14	16230 1.10.35	147.13.44	140127 33738	15
151 2.34.23	16410 1.10.43	148.18.42	140005 33651	15
152 2.29.29	16580 1.10.49	149.23.44	139887 33566	14
153 2.24.33	16750 1.10.56	150.28.49	139773 33484	14
154 2.19.34	16910 1.11. 3	151.33.57	139663 33406	14
155 2.14.33	17060 1.11.10	152.39. 9	139558 33331	13
156 2. 9.30	17210 1.11.16	153.44.23	139456 33259	13
157 2. 4.23	17350 1.11.22	154.49.40	139358 33187	12
158 1.59.18	17480 1.11.28	155.55. 0	139263 33119	12
159 1.54. 8	17600 1.11.33	157. 0.23	139173 33054	11
160 1.48.56	17720 1.11.38	158. 5.49	139087 32992	11
161 1.43.42	17840 1.11.43	159.11.17	139005 32931	10
162 1.38.26	17950 1.11.48	160.16.47	138927 32877	10
163 1.33. 8	18060 1.11.53	161.22.19	138852 32824	9
164 1.27.48	18160 1.11.57	162.27.53	138782 32773	9
165 1.22.27	18260 1.12. 2	163.33.29	138716 32725	8
166 1.17. 4	18350 1.12. 6	164.39. 6	138654 32681	8
167 1.11.40	18440 1.12.10	165.44.45	138597 32640	7
168 1. 6.15	18520 1.12.13	166.50.26	138544 32602	7
169 1. 0.48	18590 1.12.16	167.56. 8	138495 32566	6
170 0.55.20	18650 1.12.19	169. 1.52	138450 32533	6
171 0.49.51	18710 1.12.21	170. 7.37	138410 32504	5
172 0.44.21	18770 1.12.24	171.13.24	138374 32478	4
173 0.38.50	18820 1.12.26	172.19.12	138341 32455	4
174 0.33.18	18870 1.12.28	173.25. 0	138313 32434	3
175 0.27.46	18910 1.12.29	174.30.49	138289 32417	3
176 0.22.43	18940 1.12.31	175.36.39	138269 32403	2
177 0.16.40	18960 1.12.32	176.42.29	138254 32392	1
178 0.11. 7	18980 1.12.33	177.48.19	138244 32385	1
179 0. 5.34	18990 1.12.34	178.54.10	138237 32380	0
180 0. 0. 0	18990 1.12.34	180. 0. 0	138234 32379	

Tab.Lat.

Rudolphine Tables, Kepler, 1627

Support for calculating an ephemeris for Mars
Tycho died in 1601, and Kepler didn't finish until 1624, after long negotiations over data access with Tycho's heirs
Some need for RepInfo

$00^{h}\ 00^{m}\ 00^{s}$ - $00^{h}\ 01^{m}\ 15^{s}$ 2

1 - 100

Number		Descriptor: epoch J1991.25					Position: epoch J1991.25			Par.	Proper Motion	
HIP		RA	Dec	V			α (ICRS)	δ		π	$\mu_{\alpha*}$	μ_{δ}
		h m s	±° ′ ″	mag			deg	deg		mas	mas/yr	
1	2	3	4	5	6	7	8	9	10	11	12	13
1		00 00 00.22	+01 05 20.4	9.10		H	0.000 911 85	+01.089 013 32		3.54	−5.20	−1.88
2		00 00 00.91	−19 29 55.8	9.27		G	0.003 797 37	−19.498 837 45	+	21.90	181.21	−0.93
3		00 00 01.20	+38 51 33.4	6.61		G	0.005 007 95	+38.859 286 08		2.81	5.24	−2.91
4		00 00 02.01	−51 53 36.8	8.06		H	0.008 381 70	−51.893 546 12		7.75	62.85	0.16
5		00 00 02.39	−40 35 28.4	8.55		H	0.009 965 34	−40.591 224 40		2.87	2.53	9.07
6		00 00 04.35	+03 56 47.4	12.31		G	0.018 141 44	+03.946 488 93		18.80	226.29	−12.84
7		00 00 05.41	+20 02 11.8	9.64		G	0.022 548 91	+20.036 602 16		17.74	−208.12	−200.79
8		00 00 06.55	+25 53 11.3	9.05	3	H	0.027 291 60	+25.886 474 45		5.17	19.09	−5.66
9		00 00 08.48	+36 35 09.4	8.59		H	0.035 341 89	+36.585 937 77		4.81	−6.30	8.42
10		00 00 08.70	−50 52 01.5	8.59		H	0.036 253 09	−50.867 073 60		10.76	42.23	40.02

Hipparcos catalogue, European Space Agency, 1997

Hipparcos catalogue of high-quality positions of stars
Published as database, CD and paper/PDF, with checksums so it can be rescanned after the apocalypse

```
*******************************************************************************************************************
 Date__(UT)__HR:MN     R.A._(ICRF/J2000.0)_DEC  APmag  S-brt            delta      deldot    S-O-T /r    S-T-O
*******************************************************************************************************************
$$SOE
 2014-Jan-27 00:00     13 23 04.76 -06 04 24.1   0.36   4.64 1.10571128682739 -17.2472507 105.2740 /L  34.8354
 2014-Jan-28 00:00     13 24 15.59 -06 10 41.0   0.34   4.63 1.09591034007907 -17.2017577 105.9780 /L  34.7095
 2014-Jan-29 00:00     13 25 24.94 -06 16 48.1   0.32   4.63 1.08613636280390 -17.1524613 106.6885 /L  34.5767
 2014-Jan-30 00:00     13 26 32.77 -06 22 45.4   0.29   4.63 1.07639158121121 -17.0991836 107.4056 /L  34.4369
 2014-Jan-31 00:00     13 27 39.06 -06 28 32.5   0.27   4.63 1.06667830350106 -17.0418209 108.1296 /L  34.2900
 2014-Feb-01 00:00     13 28 43.75 -06 34 09.5   0.25   4.62 1.05699887158405 -16.9803590 108.8604 /L  34.1358
 2014-Feb-02 00:00     13 29 46.82 -06 39 36.2   0.23   4.62 1.04735561053111 -16.9148649 109.5983 /L  33.9741
 2014-Feb-03 00:00     13 30 48.23 -06 44 52.4   0.20   4.62 1.03775078956878 -16.8454566 110.3432 /L  33.8048
 2014-Feb-04 00:00     13 31 47.95 -06 49 58.1   0.18   4.61 1.02818660321171 -16.7722667 111.0954 /L  33.6277
 2014-Feb-05 00:00     13 32 45.95 -06 54 53.1   0.16   4.61 1.01866517244726 -16.6954112 111.8549 /L  33.4427
 2014-Feb-06 00:00  m  13 33 42.18 -06 59 37.4   0.13   4.61 1.00918855949709 -16.6149737 112.6219 /L  33.2497
 2014-Feb-07 00:00  m  13 34 36.61 -07 04 10.7   0.11   4.61 0.99975878814917 -16.5310016 113.3966 /L  33.0484
 2014-Feb-08 00:00  m  13 35 29.20 -07 08 32.9   0.08   4.60 0.99037786367079 -16.4435102 114.1790 /L  32.8388
 2014-Feb-09 00:00  m  13 36 19.92 -07 12 44.0   0.06   4.60 0.98104778933740 -16.3524897 114.9693 /L  32.6207
 2014-Feb-10 00:00  m  13 37 08.73 -07 16 43.7   0.04   4.60 0.97177057900355 -16.2579117 115.7677 /L  32.3938
 2014-Feb-11 00:00  m  13 37 55.58 -07 20 32.0   0.01   4.60 0.96254826648275 -16.1597333 116.5744 /L  32.1581
 2014-Feb-12 00:00  m  13 38 40.44 -07 24 08.6  -0.01   4.60 0.95338291301032 -16.0578998 117.3894 /L  31.9134
 2014-Feb-13 00:00  m  13 39 23.26 -07 27 33.5  -0.04   4.59 0.94427661407206 -15.9523444 118.2131 /L  31.6594
 2014-Feb-14 00:00  m  13 40 04.00 -07 30 46.6  -0.07   4.59 0.93523150662912 -15.8429871 119.0455 /L  31.3960
 2014-Feb-15 00:00  m  13 40 42.61 -07 33 47.5  -0.09   4.59 0.92624977739127 -15.7297322 119.8870 /L  31.1230
 2014-Feb-16 00:00  m  13 41 19.04 -07 36 36.2  -0.12   4.59 0.91733367237105 -15.6124666 120.7375 /L  30.8403
 2014-Feb-17 00:00  m  13 41 53.26 -07 39 12.6  -0.14   4.58 0.90848550757430 -15.4910576 121.5975 /L  30.5475
 2014-Feb-18 00:00  m  13 42 25.21 -07 41 36.4  -0.17   4.58 0.89970768042736 -15.3653525 122.4670 /L  30.2445
 2014-Feb-19 00:00  m  13 42 54.85 -07 43 47.6  -0.20   4.57 0.89100268146319 -15.2351783 123.3462 /L  29.9312
 2014-Feb-20 00:00  m  13 43 22.12 -07 45 45.8  -0.22   4.57 0.88237310587581 -15.1003423 124.2354 /L  29.6072
 2014-Feb-21 00:00     13 43 46.97 -07 47 31.1  -0.25   4.56 0.87382166470638 -14.9606327 125.1349 /L  29.2725
 2014-Feb-22 00:00     13 44 09.37 -07 49 03.2  -0.28   4.56 0.86535119546583 -14.8158200 126.0447 /L  28.9267
 2014-Feb-23 00:00     13 44 29.25 -07 50 21.9  -0.31   4.55 0.85696467171760 -14.6656594 126.9651 /L  28.5696
 2014-Feb-24 00:00     13 44 46.57 -07 51 27.1  -0.33   4.54 0.84866521037136 -14.5098969 127.8963 /L  28.2012
 2014-Feb-25 00:00     13 45 01.28 -07 52 18.6  -0.36   4.54 0.84045607420631 -14.3482802 128.8384 /L  27.8211
 2014-Feb-26 00:00     13 45 13.33 -07 52 56.4  -0.39   4.53 0.83234066587518 -14.1805770 129.7917 /L  27.4292
$$EOE
```

ephemeris for Mars, from Glasgow, 2014 January 27 to 2014 February 26, Jet Propulsion Laboratory, 2014

http://ssd.jpl.nasa.gov/horizons.cgi

what is big science?

- big money: ~20 year history, and millions of $/€/£ (LHC budget is €3bn + detectors, hardware and people)
- big author lists: *collaborations* of 100s of people (LIGO is 800 authors, ATLAS 3000)
- big data: petabytes per year (1LHC=10PB/yr)
- big admin: MOUs, councils, workshop series
- big careers: PhD to tenure on a single project

The real author is ‘The X Collaboration’

lots of data

- ATLAS/CMS at LHC: 10 PB/yr
- LIGO: ~1PB/yr
- SKA (by 2020): 1 TB/min or 0.5 EB/yr intercontinentally (this is 0.05% of 1 ZB/yr total worldwide 2015 IP traffic)
- Not a problem

kilo ➛ mega ➛ giga ➛ tera ➛ peta ➛ exa ➛ zetta ➛ yotta

Always at the limits of what it is feasible to store and transport (ie big-science projects are often implicitly ICT research projects)
Willing to experiment with innovative data-management solutions

$0bn problem

- Well, it is a problem, but it's not just *our* problem
- Jim Gray: "astronomy data is a zero-billion dollar problem"
- SDSS uses SQLServer, CERN uses Oracle

This slide is about: WHY is big science funded?

- Very large custom data-analysis software suites
- ...which are hard to use
- ...and require lots of tacit knowledge (ie gained from officemates, and maybe written into wikis)
- A major software preservation challenge

Particle physics data becomes unintelligible about 30 times faster than astronomy data

1000 year old astronomy data intelligible, 30 year old HEP data is _old_

things that make it easy

- Big science projects are often well-resourced, with IT experience, engineering management and clear collaboration infrastructure
- Historical experience of 'large' data volumes mean everyone knows ad hoc doesn't work
- Always shared facilities, so documented interfaces and SLAs are natural
- Confidentiality concerns are well understood (professional priority rather than family secrets)

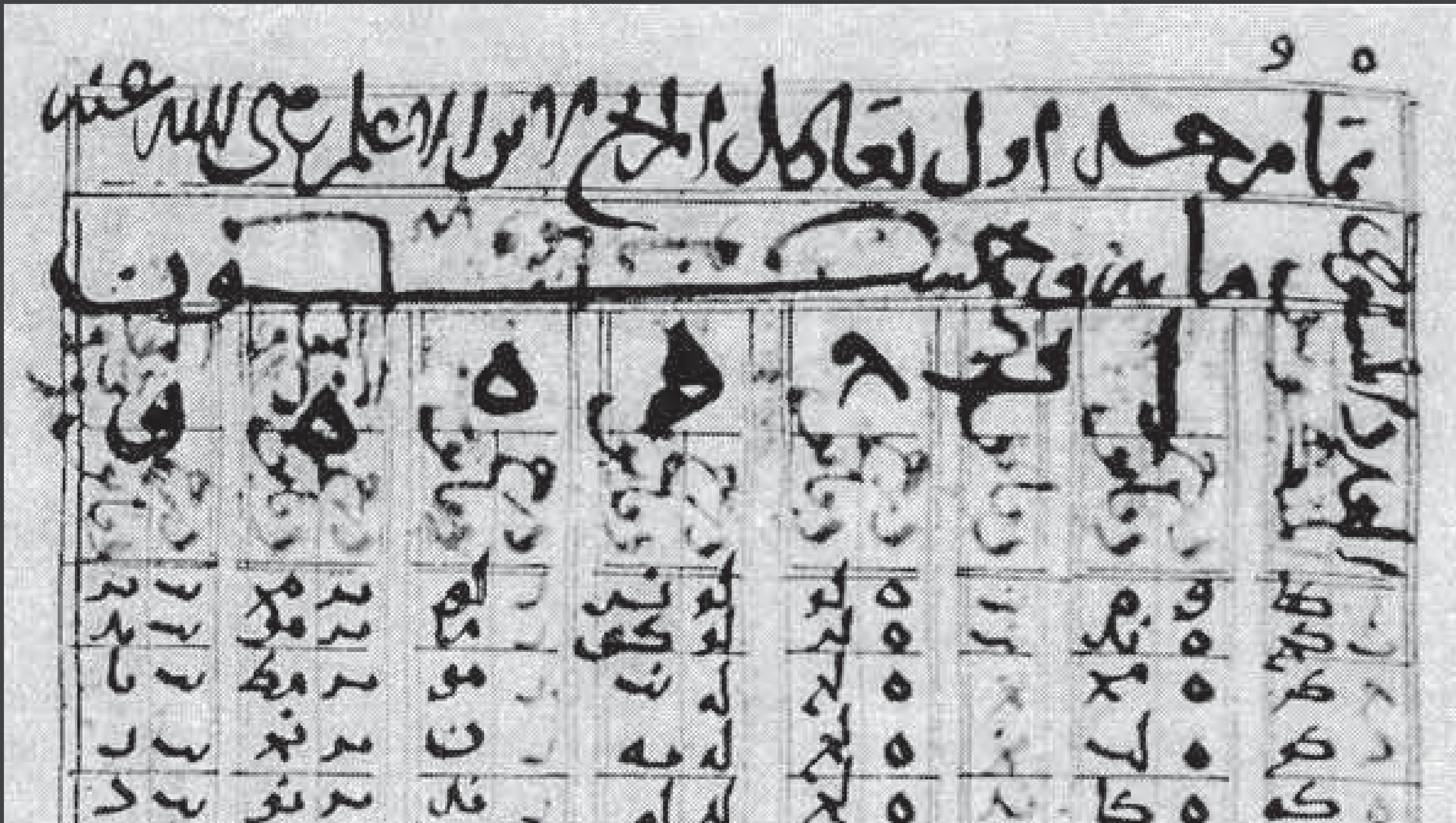


Table of equations for Mars, from al-Zij al-Mumtahan
(from Benno van Dalen, *A Second Manuscript of the Mumtahan Zîj*, Suhayl 4 (2004), pp. 9-44)

al-Zij al-Mumtahan -> ‘corrected tables’, updating Greek observations

asshab al-Mumtahan

- Shammasiyya Observatory, founded 828 CE – part of Caliph al-Ma'mun's 'House of Wisdom'
- Involved Mansur, al-Khwarizmi and many others (observers, technicians, administrators)
- Reobserving Greek data
- al-Zij al-Mumtahan published by Asshab al-Mumtahan
- Also diameter of the earth, and a new map

Asshab al-Mumtahan -> Mumtahan Collaboration
Shammasiyya may have been the first purpose-built, state-funded observatory

data products and proprietary periods

hierarchies of data

- raw data (level 0): direct output of detector, or CCD frame, or satellite telemetry – barely usable to anyone but the instrument team
- data products (possibly multiple levels): ‘reduced data’ (calibrated/interpreted), in standard/ documented formats – scientifically usable without specialised knowledge
- publications: articles and catalogues – peer-reviewed outputs

Natural object of preservation (corresponds to AIP and DIP in OAIS)

- (‘embargoed’ would be a better term)
- Data available (from the archive) only to data ‘owner’ for some period, or to consortium; thereafter public
- Periods of 12, 18, 24, 36 months
- …or until some discovery happens (eg HEP & GW)

The proprietary period is part of the ‘currency’ of negotiation
Questions of public access, to both data and papers

what does science data look like?

Summit of Mauna Kea, in Hawai`i. Left to right: Subaru, Keck I and II, and the NASA Infrared Telescope Facility

- Image data: multi-MB CCD images, plus calibration images, stored as flat files; easy to understand
- No file format problem – the whole world gets FITS
- Catalogue data: list of object properties (RDBMS)
- Non-optical data: different detail, same expectations
- Data goes from the instrument direct to the archive
- Astronomers bid for 'telescope time' on shared facilities/instruments

All telescopes are different, but relatively low-level data products will be broadly intelligible

The VO: a Vision of having all the astronomical data in the world, available to be processed meaningfully: "What does object X look like in X-rays and radio?"

Requires:

- Archive metadata (what is this image looking at?)
- Provenance (where did this image come from?)
- Semantics (what does this number mean?)

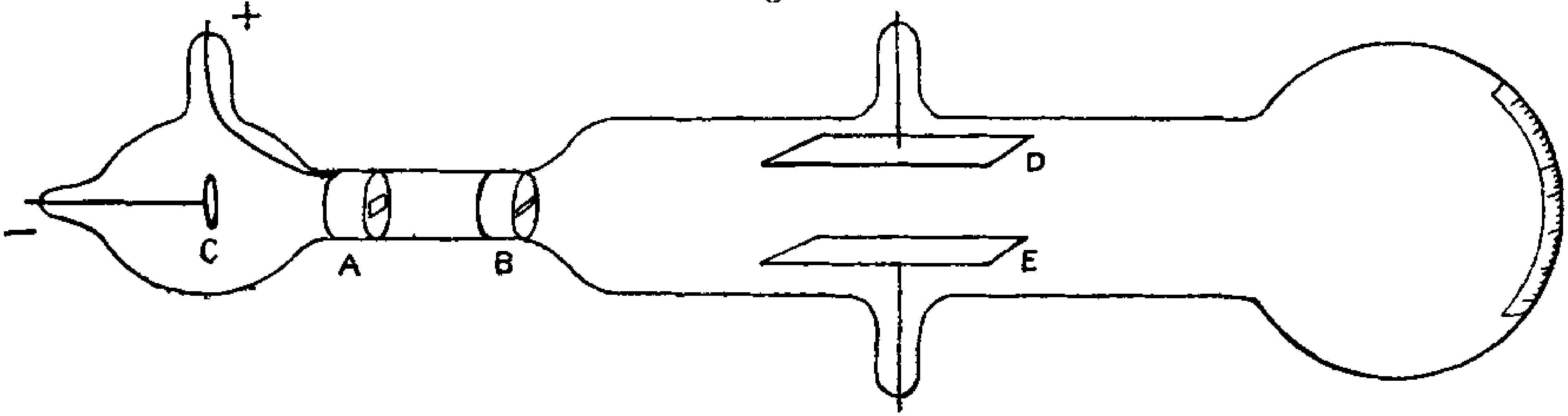
It's mostly working; `www.ivoa.net`. Only possible because (a) astronomy is an observational science, & (b) there's only one sky

Note VO=virtual observatory != VO=virtual organisation
Five years ago, combining multi-wavelength observations was worth a paper in itself; the VO's goal is to make this trivial
Technical slog

- generally available from large professional archives, both project-specific and general (eg `cds.u-strasbg.fr` or `www.sdss.org`)
- bibliographic archives well-organised (`adswww.harvard.edu` and `arxiv.org`)

The apparatus used is represented in fig. 2.

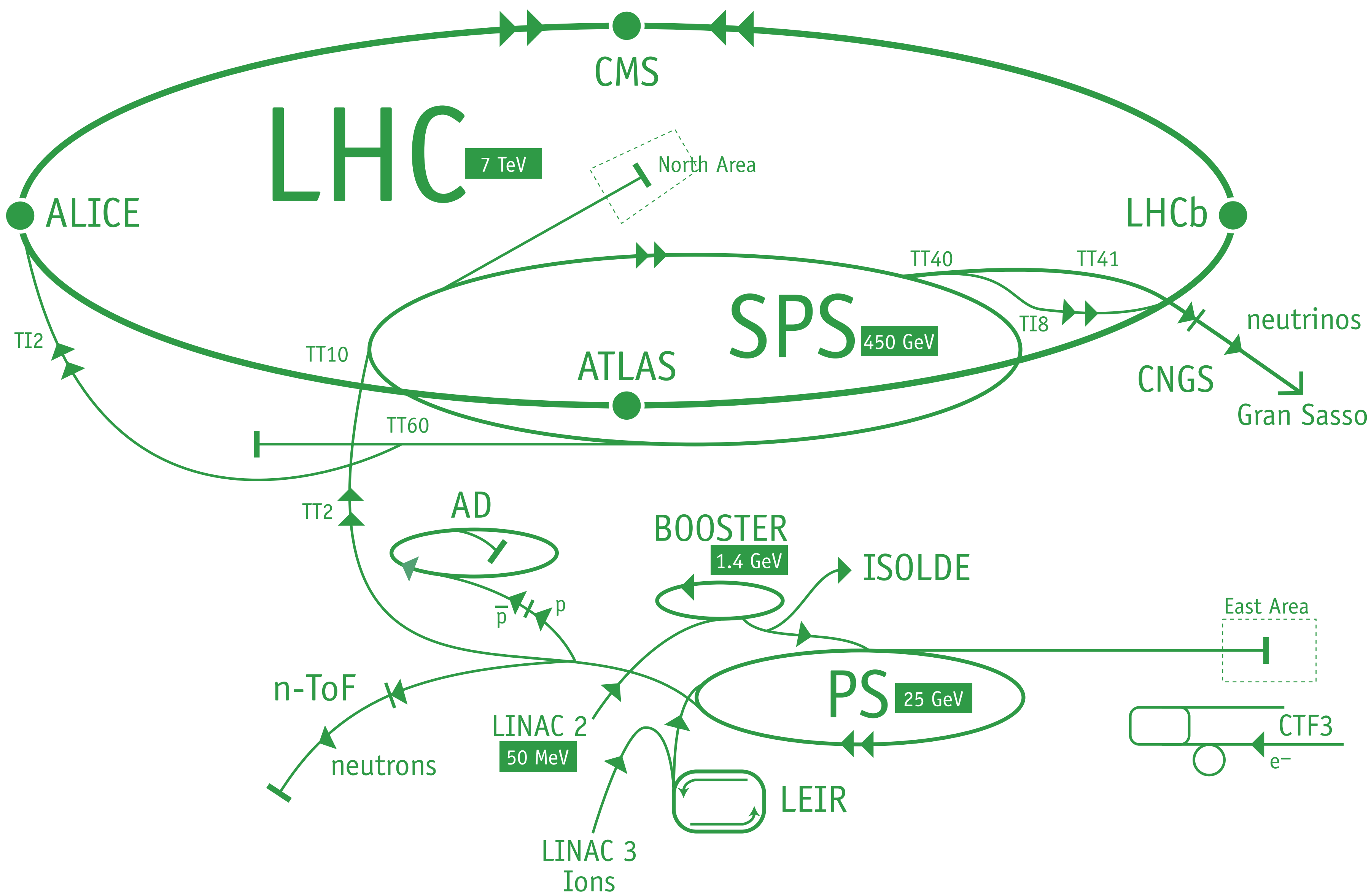
Fig. 2.

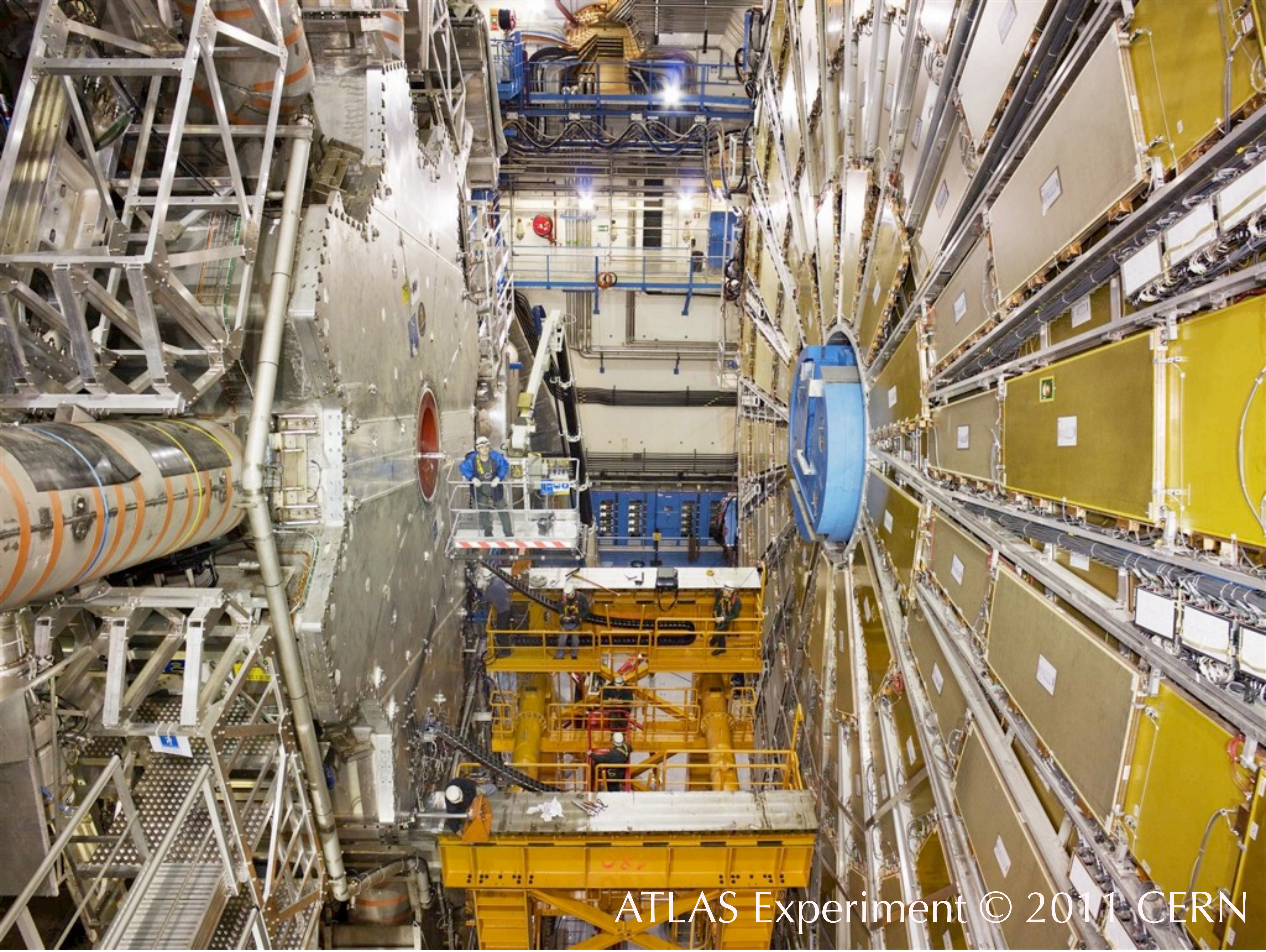


The rays from the cathode C pass through a slit in the anode A, which is a metal plug fitting tightly into the tube and connected with the earth; after passing through a second slit in another earth-connected metal plug B, they travel between two parallel aluminium plates about 5 cm. long by 2 broad and at a distance of 1·5 cm. apart; they then fall on the end of the tube and produce a narrow well-defined phosphorescent patch. A scale pasted on the outside of the tube serves to measure the deflexion of this patch.

J J Thomson, Phil. Mag (Series 5), 44(269), 293 (1897)
doi:10.1080/14786449708621070

particle physics = accelerator + detectors





End cap being moved into place (2007)

particle data at the lhc

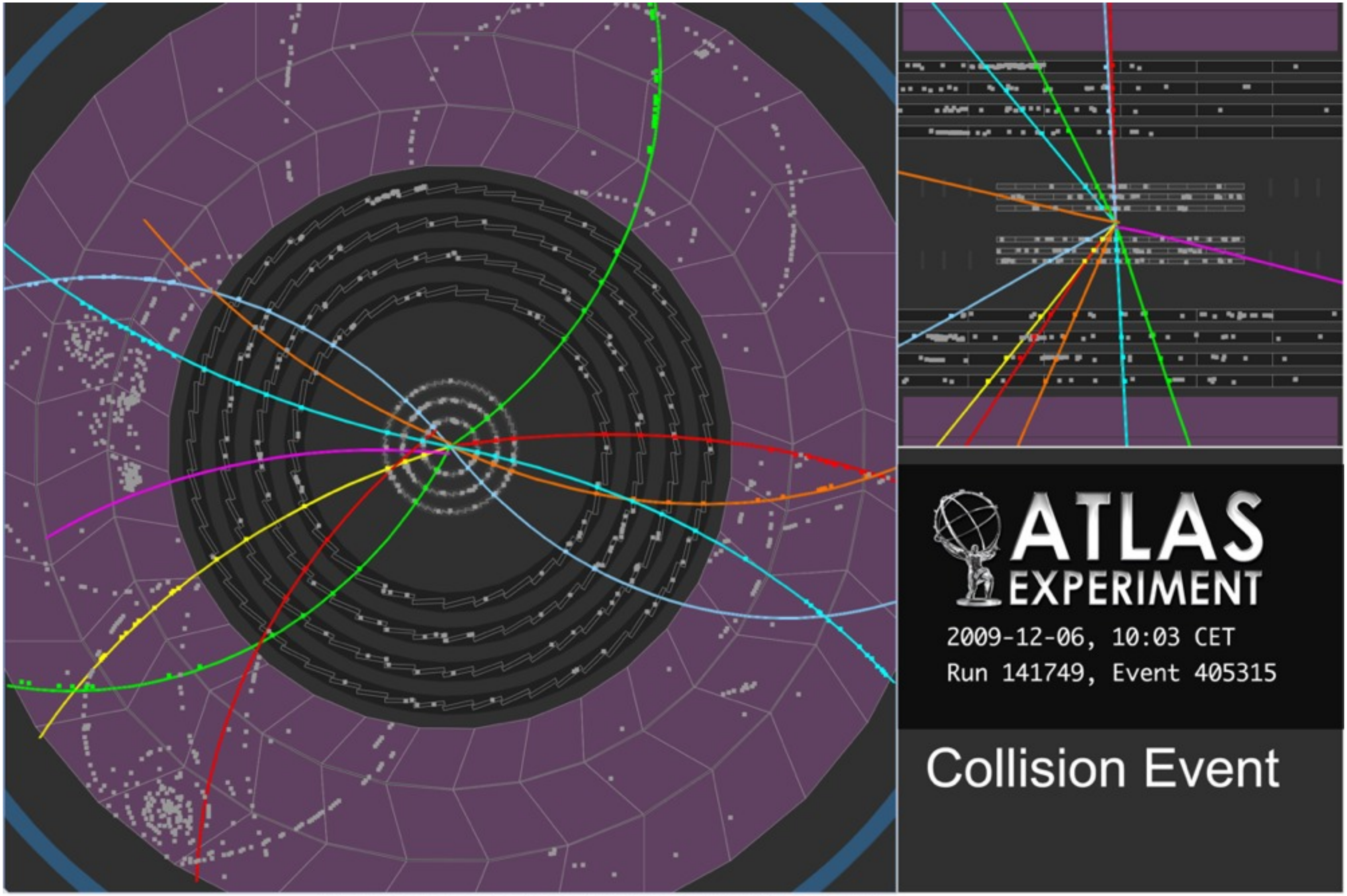
- Thousands of beam-crossings per second
- Potentially multiple PB/sec of data
- ...but most of it is thrown away (by the detector)
- ...leaving only ~10 PB/yr from ATLAS and CMS

Data decimation is done by triggers built in to the detector electronics, making snap semi-heuristic decisions in between beam-crossings.
All the data that leaves the detector is stored.

- Tier 0: CERN – spinning disks and lots of tape robots; safe copy of all of the data, and low-level data products
- Tier 1: 11 national data centres – raw data plus generation and storage of data products
- Tier 2: ~140 sub-national centres – fractions of the data plus generation of further products

http://lcg.web.cern.ch

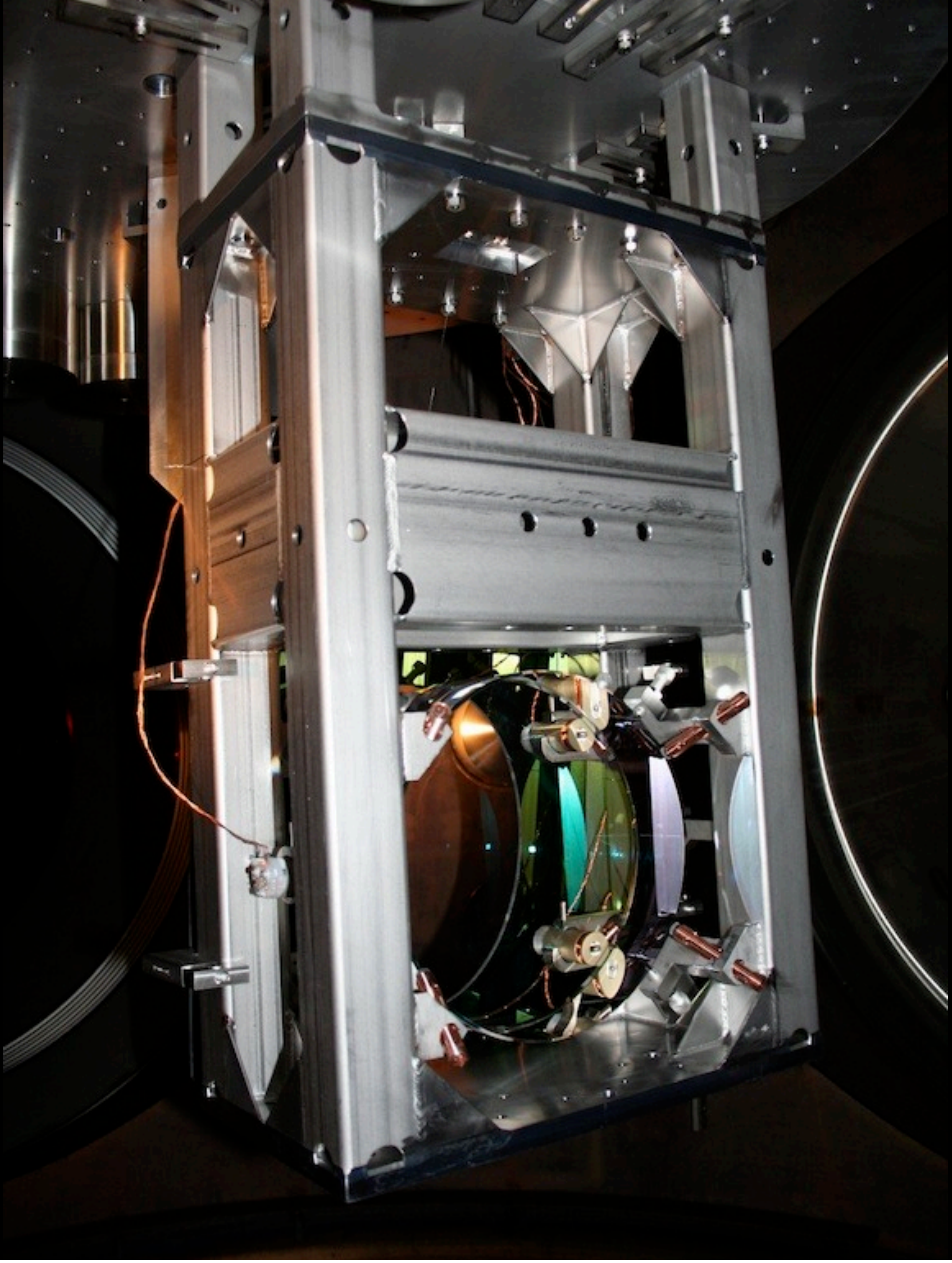
Also a ‘tier 3’ of departmental and personal stores
Tapes are good, because they use zero power when they’re offline
Mflop/W is the key measure
(note these tiers aren’t levels of product)



http://atlas.web.cern.ch/Atlas/public/EVTDISPLAY/events.html

First 900 GeV Collision Events in Stable-Beam Conditions with Inner Detector Fully Powered, December 6, 2009
Display of a collision event showing tracks in the Inner Detector system.

- raw data is *very* instrument-specific (and so essentially meaningless by itself)
- data reduction software needs huge expertise & knowledge to run and interpret (so can't easily be preserved)
- published data is *very* reduced
- raw data is *not* shared (or indeed very shareable)



Test mass assembly at one end of a LIGO interferometer arm

gravitational wave data

- astronomy, but a lot of features of HEP culture
- 800 authors in an open collaboration
- data interpretation heavily dependent on software
- an institution can join, and get access to the data, in return for personnel and resources, and accepting publication policies
- data swaps with other projects
- elaborate data release planning

stepping back: how and why should we preserve data?

"Scientists should preserve and immediately share their raw data so other scientists, and the public, can reanalyse and reuse it"

Well, no, not really...

Ultimately, yes, something like that should happen, because science is about criticism: ‘nullis in verba’, and all that.

But...

This may be simpler in biology or medicine or chemistry
But it’s not simple if your raw data consists of a petabyte of noise
Take it for granted that something like this is an aspiration: this is all about the “but...”

why data shouldn't be shared

- Data isn't free – personal and professional costs
- Raw data is generally useless
- Making data products may be very expensive
- Countering (accidental or mischievous) misinterpretations of raw data may be expensive

these arguments against are practical and cost reasons

why data shouldn't be preserved

- It may be very expensive to do so
- There might be no interest after a few years
- The Designated Community may be null in the long term

Preserving HEP data long term is nightmarish, astronomy data less so.
Preserving HEP data in the shortish term (a couple of decades) probably isn't too bad.
Old HEP data (for some value of 'old') isn't very interesting

why raw data is useless

- Large experiments are unintelligible to outsiders
- ...which means they couldn't create analysis software
- ...nor even re-run the experiment's own software
- Even simple experiments need to process raw data (in undocumented ways) before it's usable
- Also no-one, in practice, tends to ask for it

I'm not saying that data shouldn't be shared or preserved, or that raw data should be thrown away, but countering the apparent assumption (in some quarters) that it's obvious that raw data should always be preserved and shared, because there's a scientific demand for it.

Enough data!
Look at nature.

http://moodle2.gla.ac.uk/course/view.php?id=4069

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

Norman Gray – http://nxg.me.uk