cesagen

Understanding the ‘intensive’ in ‘data intensive research’:

Data flows in next generation sequencing and environmental networked sensors

Ruth McNally
Senior Research Fellow
ESRC Cesagen
Lancaster University

ESRC genomics network

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

LANCASTER UNIVERSITY

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Co-authors:

Adrian Mackenzie, Cesagen, Lancaster
Jennifer Tomomitsu, (formerly Cesagen, Lancaster)

Allison Hui, David C. Lam Institute for East-West Studies, Hong Kong Baptist University (also formerly Cesagen, Lancaster).

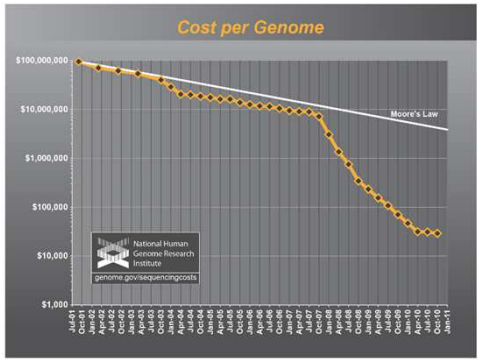
This research was undertaken with support from the e-Science Institute, Edinburgh

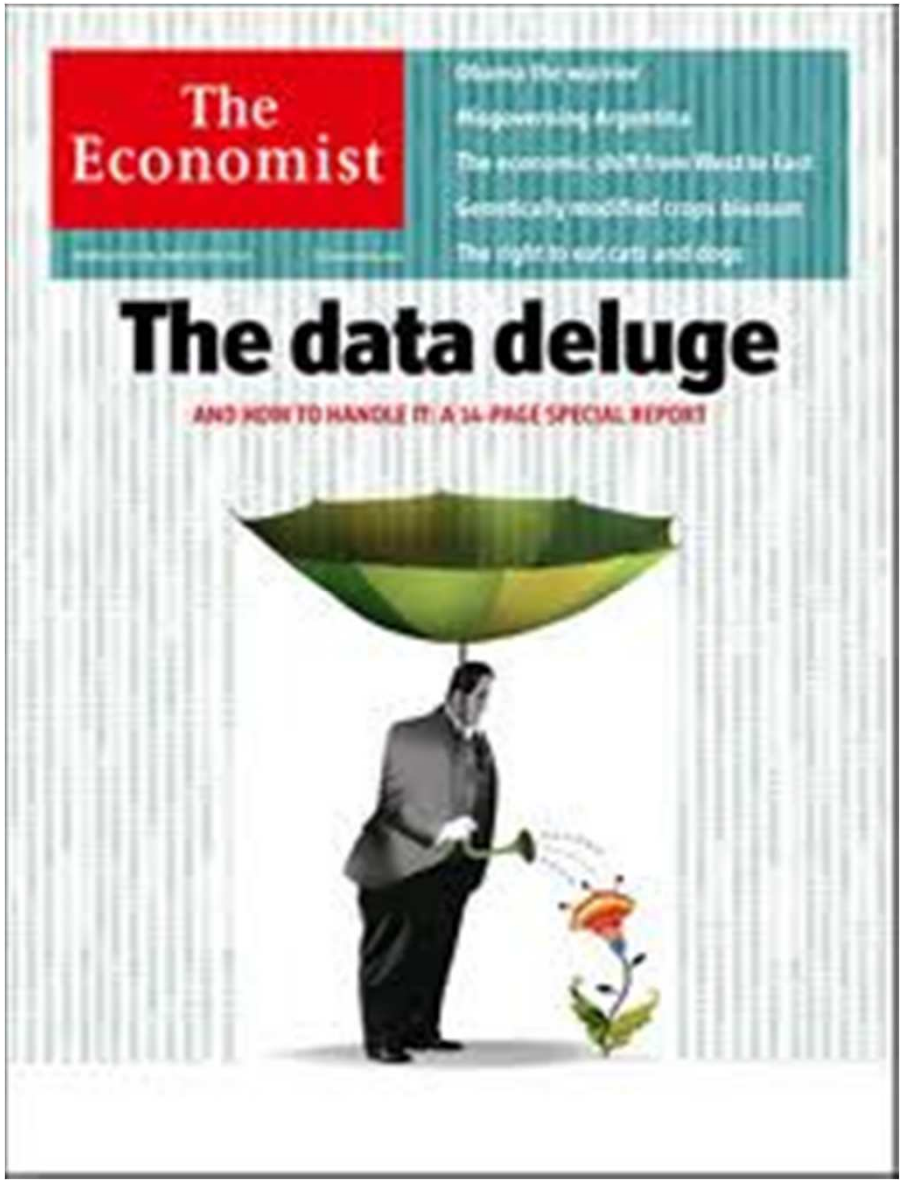
Our motivation

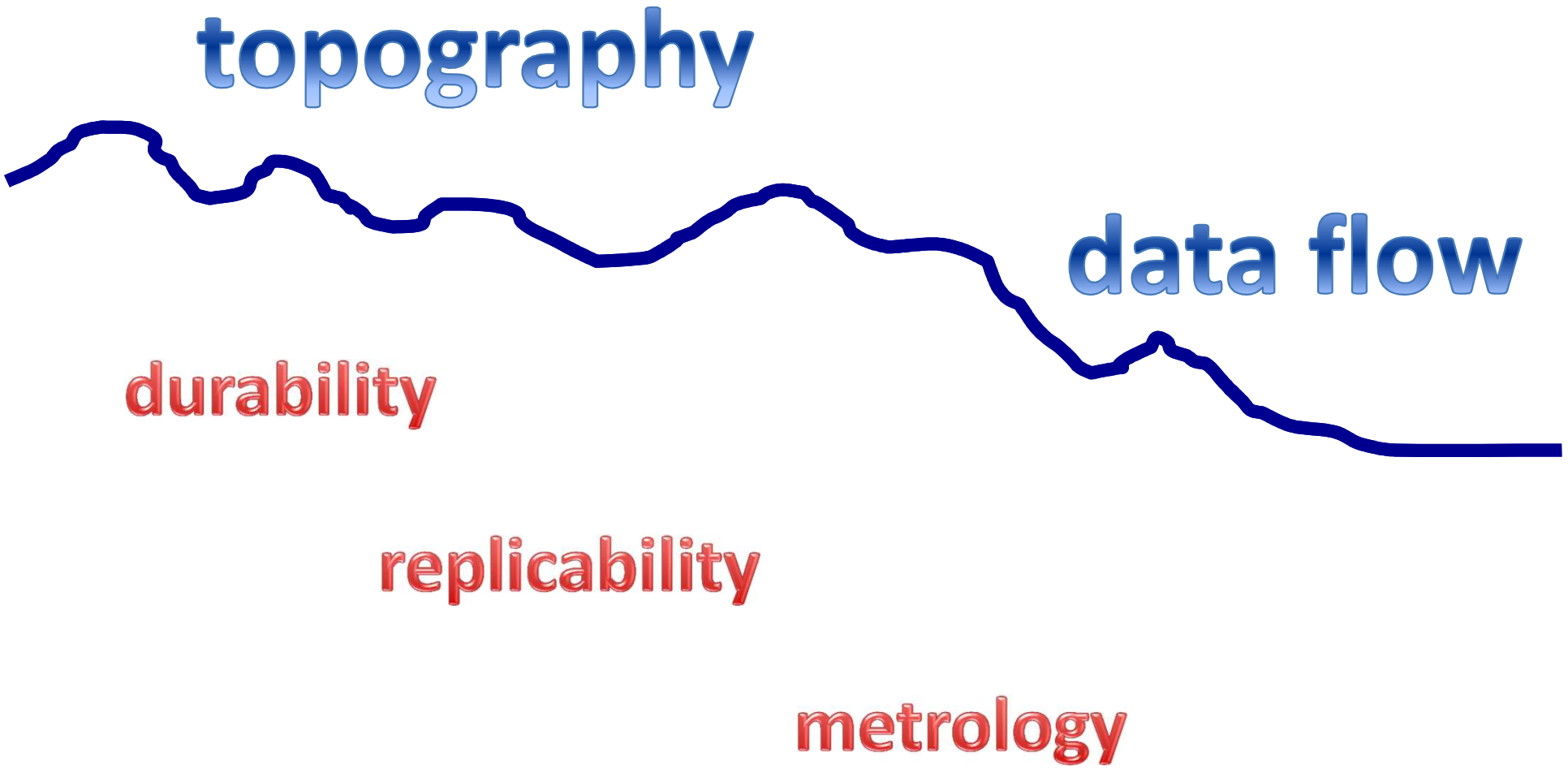
DNA Sequencing Caught in Deluge of Data

New York Times 30 Nov 2011

‘The lower cost, along with increasing speed, has led to a huge increase in how much sequencing data is being produced. World capacity is now 13 quadrillion DNA bases a year, an amount that would fill a stack of DVDs two miles high, …..’





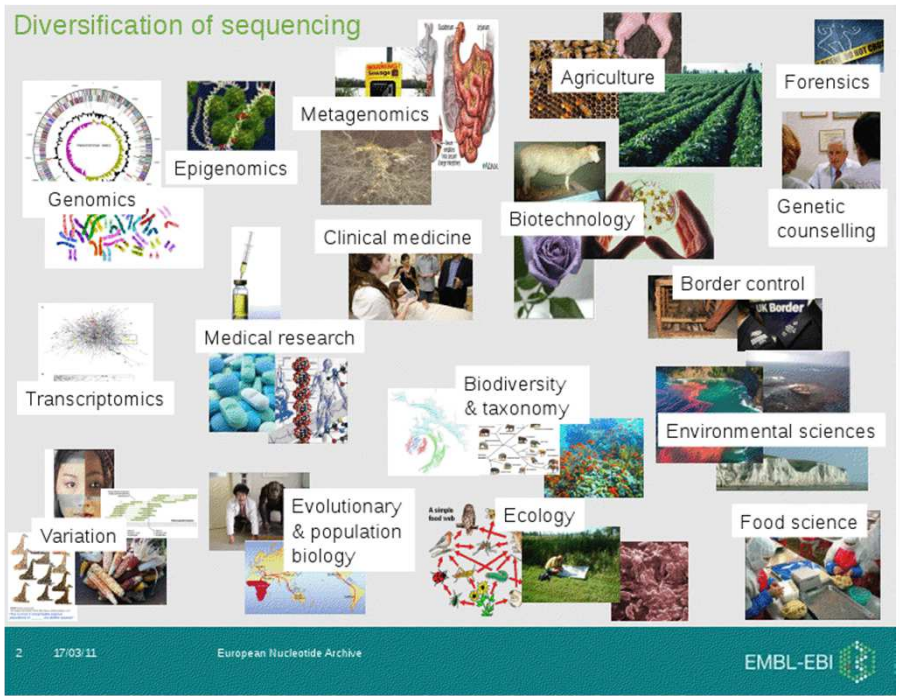

topography
data flow
durability
replicability
metrology

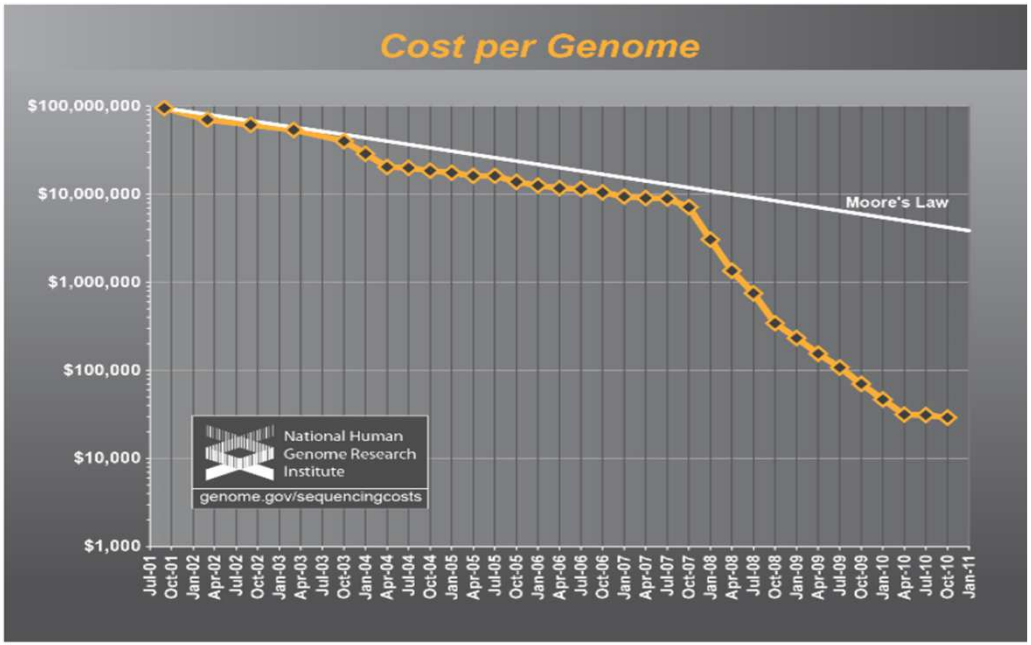
Commoditisation of sequence - Personal sequencers

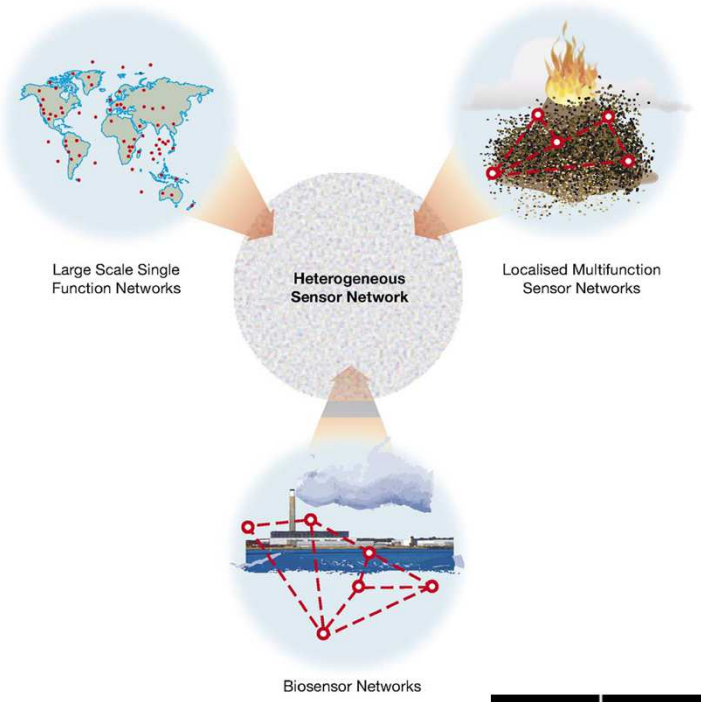
Next Generation Sequencing

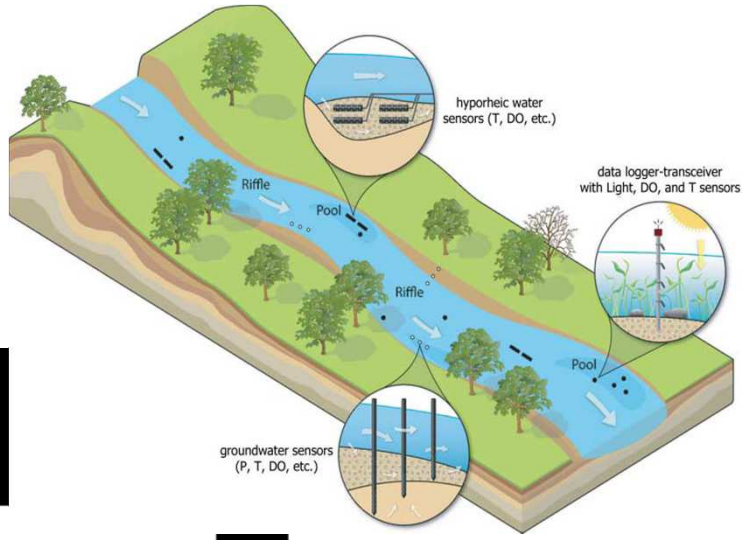
1622 total machines; inc,.712 in USA, 199 in China and 132 in UK.

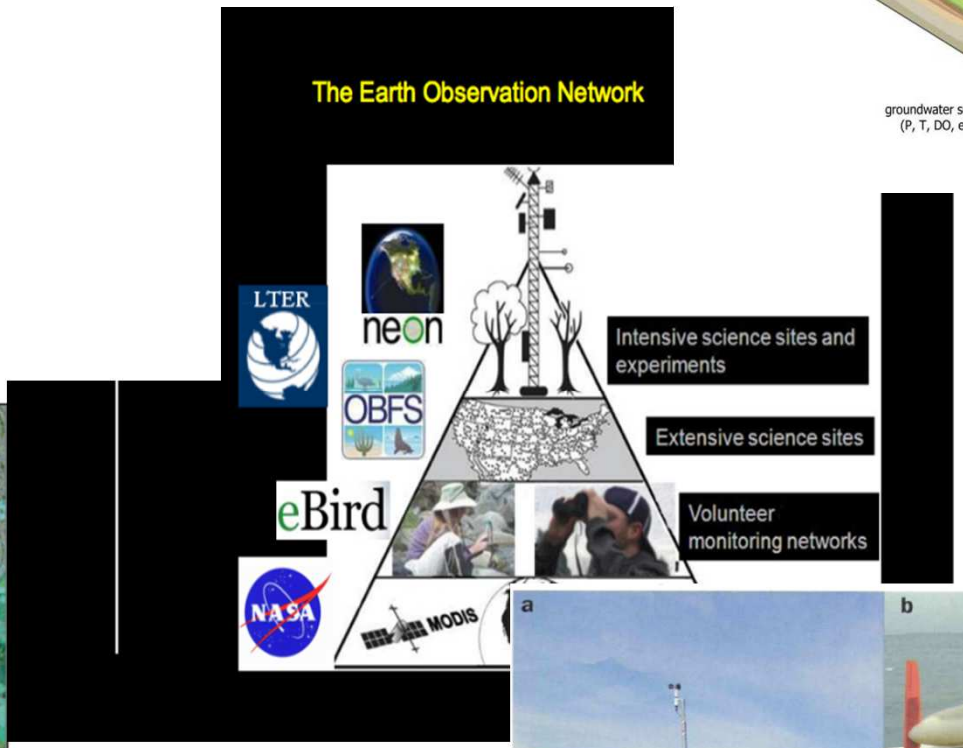
Faster benchtop machines –> results in a day
More expensive
Diagnostics 10Mb-1Gb



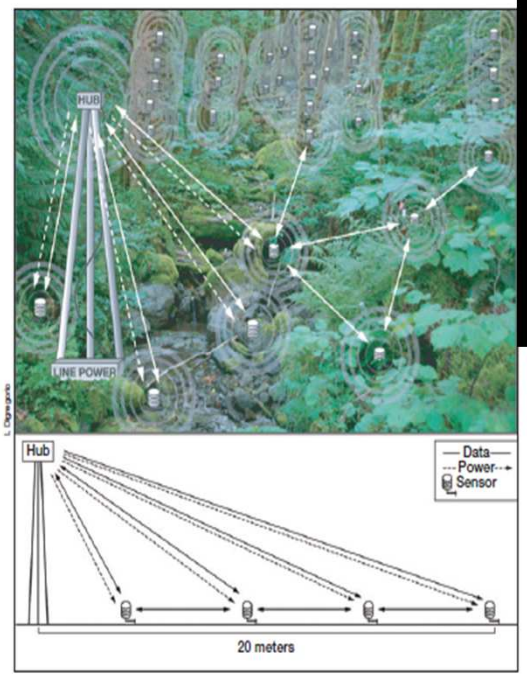


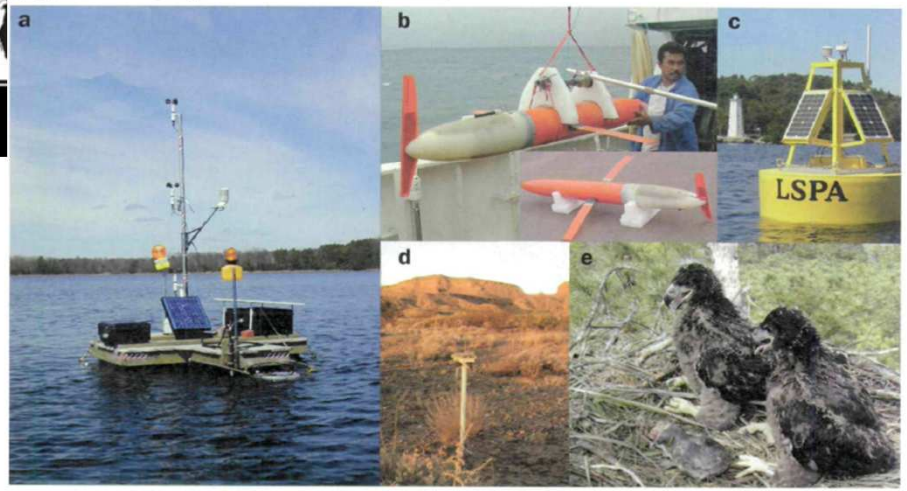




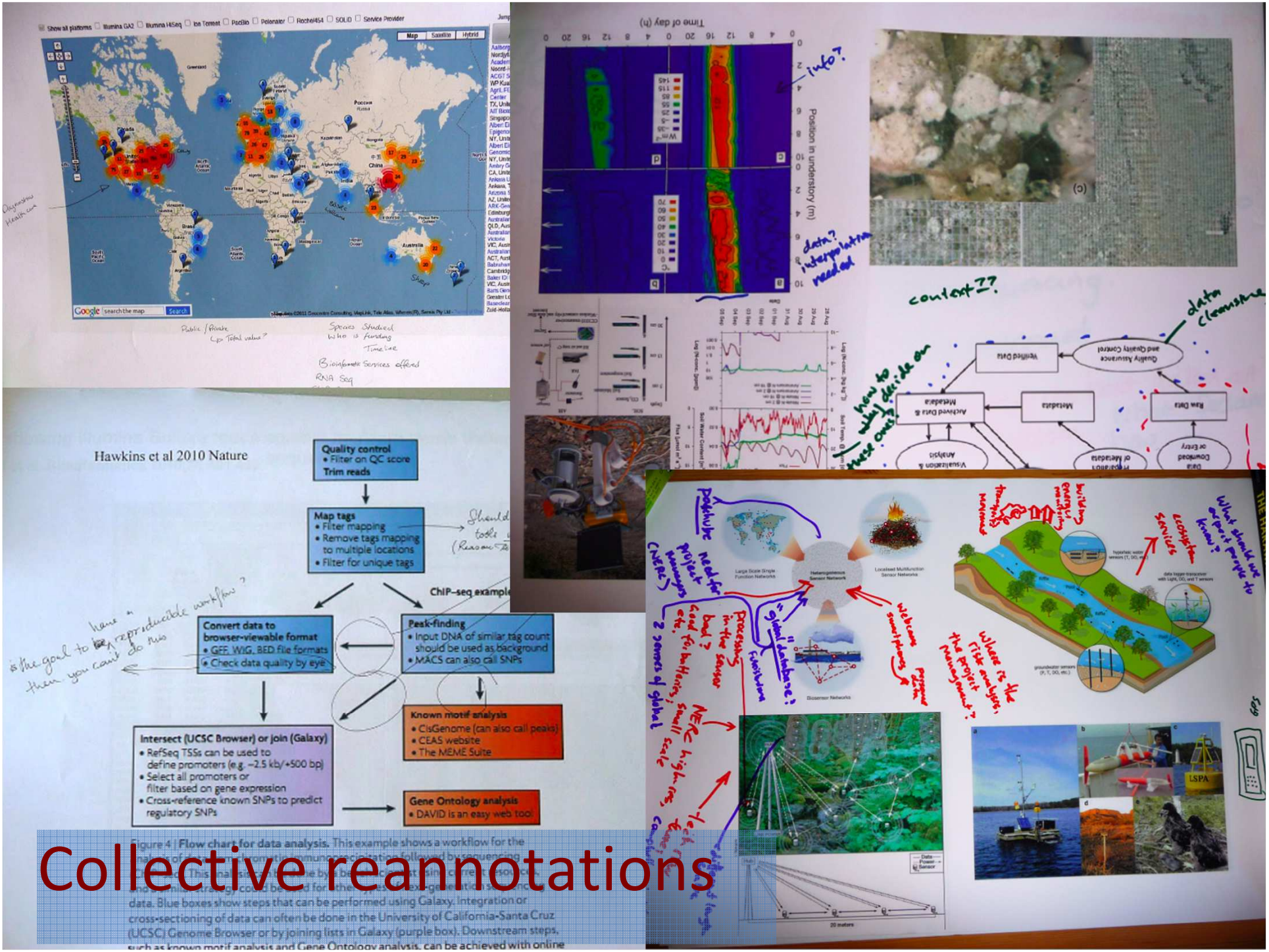


Embedded Networked Sensors





Instrumented workshops


Hawkins et al 2010 Nature
Quality control
• Filter on QC score
Trim reads
Map tags
• Filter mapping
• Remove tags mapping to multiple locations
• Filter for unique tags
ChIP-seq example
Convert data to browser-viewable format
• GFF, WIG, BED file formats
• Check data quality by eye
Peak-finding
• Input DNA of similar tag count should be used as background
• MACS can also call SNPs
Known motif analysis
• CisGenome (can also call peaks)
• CEAS website
• The MEME Suite
Intersect (UCSC Browser) or join (Galaxy)
• RefSeq TSSs can be used to define promoters (e.g. –2.5 kb/+500 bp)
• Select all promoters or filter based on gene expression
• Cross-reference known SNPs to predict regulatory SNPs
Gene Ontology analysis
• DAVID is an easy web tool
Collective reannotations

Coded transcript from our own archive of digital data

282	Some of the delays in the 1000 genomes was getting irrevocible consent. Many delays in 1000 GP is to do with sample collection – the blood. You can't take 1000 GP consent back. Participants cannot change their minds.	Durability of consent - must be irrevocible
283	A big disconnect between what we see of value and what they think can get back. Disconnect between value as seen in west and as seen in those communities.	different distributions of value
284	Some communities did say yes because there were local diseases that they could get something back for. Mark: some did see local inherited disease and so said yes. Guarantees for help.	
285	In Africa, some of the samples come from the malaria genome Malariaplx are as	
286	What about all the other human genomes that have been sequenced? What about the other genomes outside of 1000gp?	
287	If they see BAM, no problem. The most sequenced individual is a child whose parents who have been done.	
288	Craig Venter – not as good his, because his is diploid. We are trying to get variants, phasing. And getting parents with children.	
289		
290	**SHANE – 1000 genomes again. - IKG data processing pipeline**	
291	[Shane has only been in post a short while]	
292	MCCARTHY 2: The Sanger perspective. Data processing and data analysis. Data production: sequencing; QC Data processing: mapping, improvement, merging Data analysis: variant calling - SNPs, indels, other structural variants BAM production for what is released to DCC then to rest of group and then the world at large	FLOWCHART: data flows. How does this map onto Laura's Slide 12? What is DCC? - Data collection centre
293	MCCARTHY 3: METADATA TRACKING Metadata tracking The DCC releases a monthly sequence index file. Update an internal metadata tracking database. Kept in sync with internal storage hierarchy. Lanes -> Library -> Platform -> Sample -> Population -> Project API to access the metadata database DCC releasing to use a sequence index file – we update our own metadata tracking database, to sync with sequence index file. Go through update bam files for mistakes. Internal pipelines have access to the tracking database Shows an old database scheme diagram.	
294	MCCARTHY 4: LANE LEVEL OPERATIONS Having got fastq files, we put them into our mapping pipeline that is parallel. We have multiple methods for calling different aligners. We have other mappers ready to go. We merge to get bam file and record stats on the way. Then we go into BAMimprovement – give it a list of known snp sites, and it realigns, sort and recalibrate. All of this is done on each lane. Monthly releases keep it going. After the mapping they do a merge and get a bam file out.. fastq in ..... MAPPING .... bam out bam in ..... BAMIMPROVEMENT .... Stats out Freeze and then start building a release.	THIS MAPS ONTO LAURA'S SLIDE 14: QUALITY CONTROL
295	MCCARTHY 5 - BUILDING A RELEASE At a certain point, there will be a freeze – 23 November. So we freeze and start building a release. Multiple lanes and maybe even multiple libraries for an individual. Merge all lanes into a common library. MCCARTHY 6, MC7, MC8 : Building a release, we start with each lane, we then merge all lanes in a common library. A common platform BAM.	
296	MC6, MC10: So for each individual, we are going to prepare data for release to the DCC. MC11: MC12, MC13: MC14: On a per chromosome basis, we split bam for all matched reads, and a bam for all unmapped bams, and initially Chromosome 20 bam because it is so small, and gives them something to do. A BAM is an individual on a common platform. If a different platform for the same individual is used, then a different BAM. SplitBam into chromosomes and a bam for all the mapped reads and a bam for the bits that didn't map to the ref genome. Then a Bam release to the DCC.	
297	MC15: So Phase 1 release, 1094 individuals (all 2x; 1003 >3x). Of the 25Tb in 2500 FASTQ FILES, about 18Tb was generated at Sanger. BAM release data. Pilot and phase 1 finished. Phase 1 is 1094 individuals. Sanger and TGEM in Texas. BAMS also coming from TGEM and from illumina and solid VCF format too	metrics: data sic number of individuals, coverage, number of lanes, number o this slide contains an image like Jenny's slide 7 - metrics as data flow in megab
298	MC16: With all the bams in, we are going to start looking for variants VARIANT CALLING - VCF FORMAT-VARIATION CALL FORMAT Fully adopted by 1000G group as interchange format for variant calls SNPs, indels, and recently SVs Genotyping calls for all samples Annotation of variants via user-defined tags VCF APIs and tools via http://vcftools.sourceforge.net Scaling issues with VCF – BCF format in development	IMAGE OF VCF FORMAT - HAS BECOME A STANDARD IN IKG GROUP data diversity - VCF standard SEE LAURA'S SLIDE 13
	MC17-MC20: VARIANT CALLING; CHALLENGES Variant calling: challenges Traditionally BAM files have been produced per sample with all of the lanes/libraries merged Lanes -> Library -> Platform -> Sample (1 per individual) Problem: population based SNP calling needs to be aware of the reads	

Durability 1: The timing of data flow has to be synchronised with domain specific temporal dynamics

‘time can be wasted’ taking advice on experimental redesign that delays the start of data flow. (NGS)

‘In CENS [a project], big issue was time stamps – notoriously bad. Sad stories about non-synched data Data sets not synched properly’ (ENS)

Durability 2: Technical and domain scientists inhabit different ‘time zones’

‘There was only about 2 years when data was collected that was of use to the application scientists. The initial period all about battery life, sensors, networks. They realized in the middle that it was important to keep the human in the loop –that coincided with about 2 years of useful data for application scientists. At the end of that, the technology was mature enough the application scientists could take with them and use it. The technology people got bored at this point and moved on to doing mobile applications – they kicked environmental scientists out of the loop’ (ENS)

Durability 3: Projects change

‘Projects can change from being one type of project into another … People who got grants to do exome capture are now going to complete genomics to get analysis’ (NGS)

‘It is the Achilles heel of every semantic integration technology that it is not robust with changes. They use the most robust one (in practice). At the moment, in terms of reliable technology, it is not that scalable. The problem is mainly that modifications cause you to have a propagation effect on the mappings’ (ENS)

Replicability 1: Too much - too little?

Experimental replicability

'Short read sequencing is so cheap, it's a disposable item. It's cheaper make and analyse your own data than to download someone else's' (NGS)

.

Replicability 2: Too much - too little?

Code and practice

‘Bioinformaticians are doing the same things over and over again. Everyone has to continue reinventing the wheel. Rinse and repeat all over the world’ (NGS)

‘Most of these things [workflows] are moving targets – in our experience for mapping and assembly, how often do we change a version of it? Hourly seems to be the response’ (NGS)

’I don't think we will ever get to fixed workflows. You will never get around to having to write new code for projects. The driver of that is the science. Science has to be novel and therefore cannot reuse whole system. That novelty is what makes you have to write new bits of code’ (NGS)

Replicability 3: Scaling up - is repetition enough?

Collaboration and enrolment

Can't do this on your own – have to have a massive team – computer scientists, engineers, domain scientists, people to keep spirits up." (ENS)

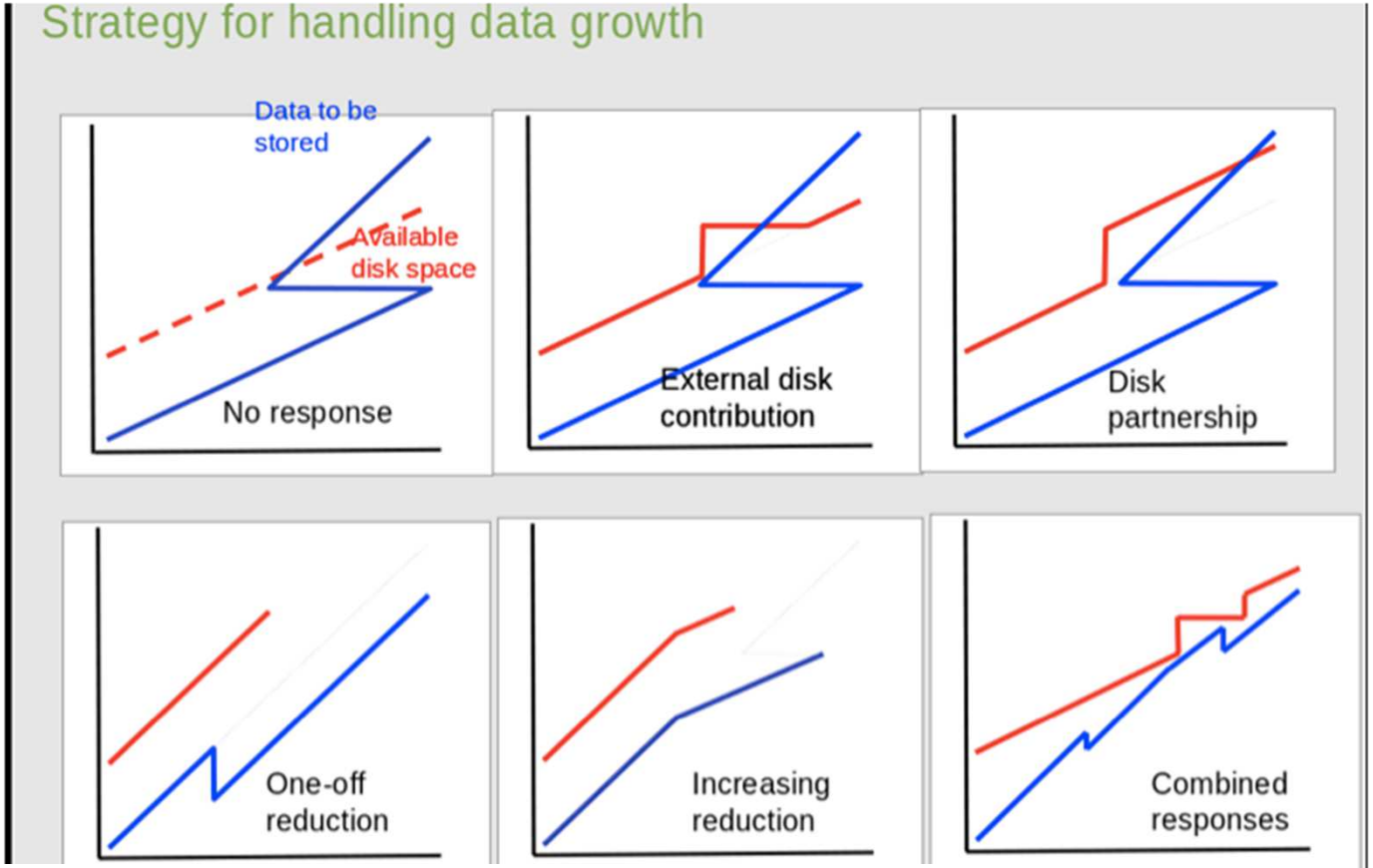
'You have to demonstrate it works as well as previous methods or better, and then wait for acceptance from the discipline before you go too far' (ENS).

System change

'One of projects – eBird – global project – concept is to get volunteers to go out and using fairly standard protocols, but standard, collect their observations of birds. … When first started, couldn't get anybody to do that. So we changed how we thought about citizen science data. Changed in 2005. Launch of eBird 2.0. Last Tuesday they collected more data than they did in 2004' (ENS)

Metrology 1

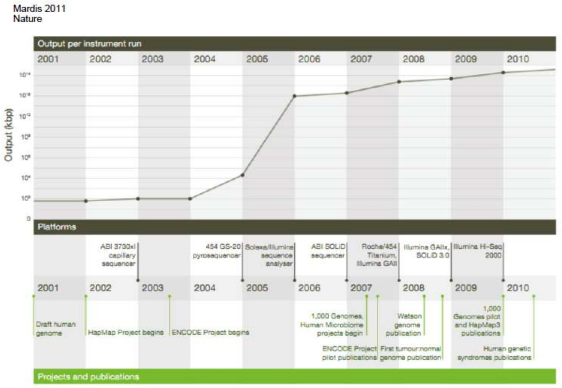
Keeping within the curve

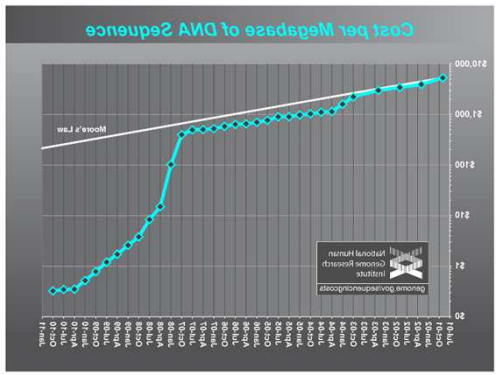


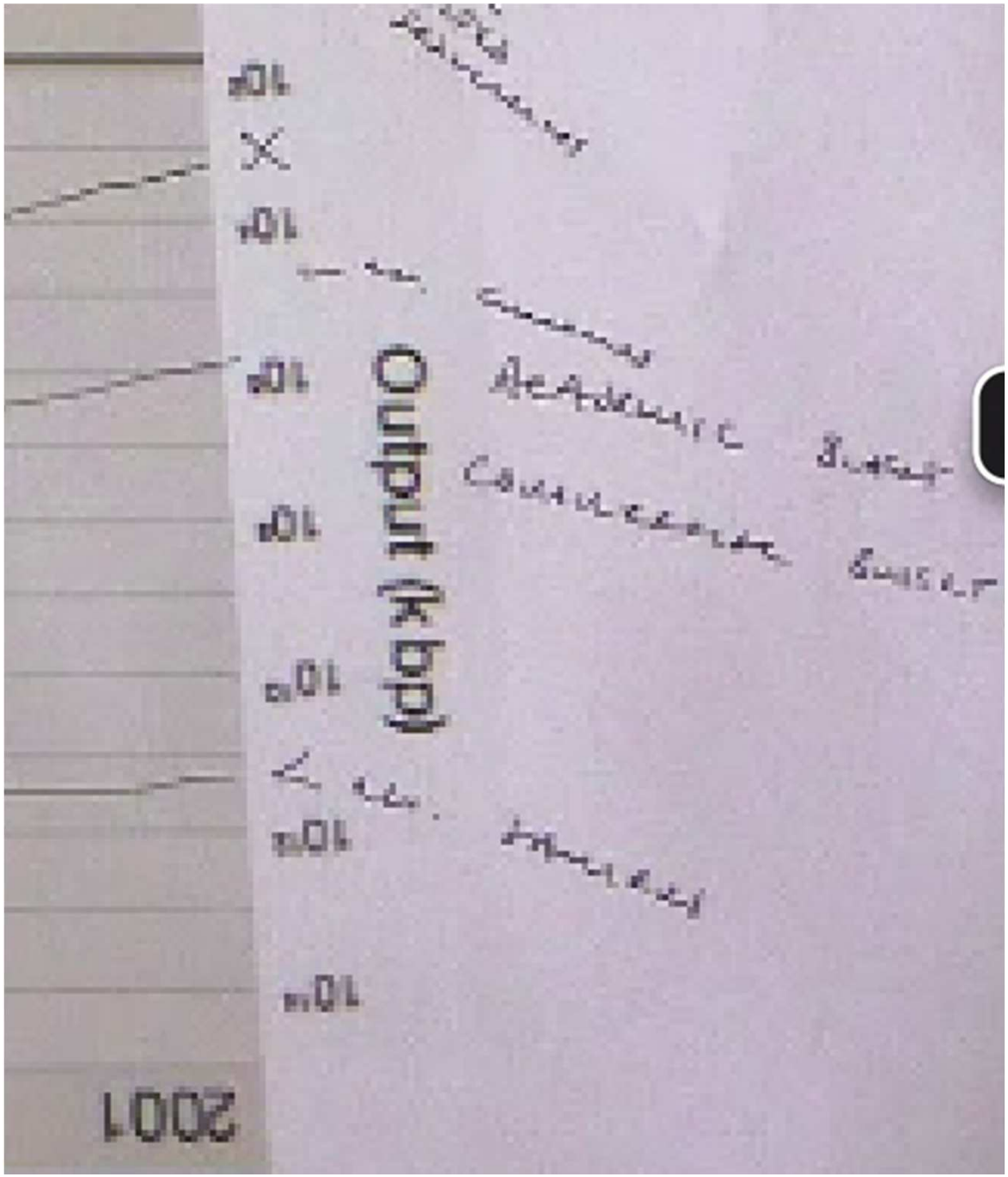
'If we have a fixed annual spend, we will be able to grow our disk space exponentially. The data is also exponential. The risk is that if we do nothing, two lines cross. I'm going to show an exercise in keeping the blue line below the red line. Either partner on diskspace. The kind of compression that will deflect this curve. There are features that will allow that to happen. The strategy is to have a set of responses to hand that we can deploy when we need to. Apply strategies of data reduction, judicious and community informed' (NGS)

Metrology 2

missing metrics



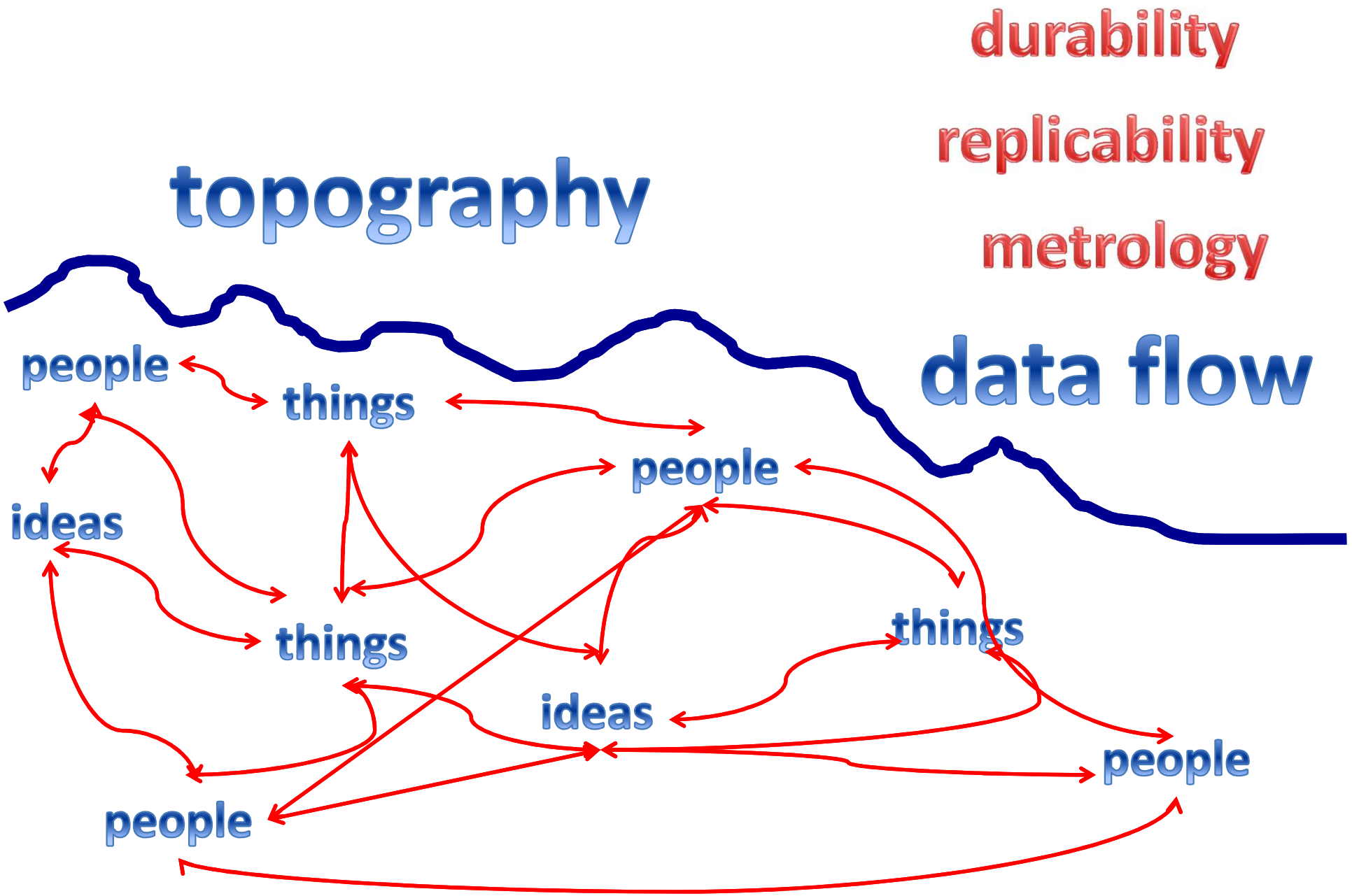




Metrology 3: Novel metrics and data bibliometrics

‘Recently an ecologist determined you could more accurately determine the onset of spring through public webcams using green divided by blue than by using remote sensing data’ (ENS)

‘Is there any benefit to having standards? You get cited more if you cite ArrayExpress. Look at ProteoRED MIAPE satisfaction survey. 95% of people like MIAPE. Papers with data in ArrayExpress get cited more than equivalent papers that don’t have data in ArrayExpress’ (NGS)


durability
replicability
metrology
topography
data flow
people
things
people
ideas
things
things
ideas
people
people

THANK YOU FOR YOUR ATTENTION

This research was undertaken with support from the e-Science Institute, Edinburgh (see http://www.esi.ac.uk/research-themes/20). The support of the Economic and Social Research Council (ESRC) is gratefully acknowledged. The work presented forms part of the programme of the ESRC Genomics Network at Cesagen.

ESRC genomics network

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

LANCASTER UNIVERSITY

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL