Lecture 2 - Domain Drivers

Professor Richard O. Sinnott
Director, eResearch
University of Melbourne
9th March 2017
rsinnott@unimelb.edu.au

Objectives

- To give the "big picture" of why we need Cluster and Cloud Computing
  - This lecture is not focused on technologies, but on giving examples of how challenges are shaping the technological landscape
    - …and how on-going/completed projects have met/are meeting those challenges
  - Many perspectives
    - Big Data – and the hype!
    - Big Compute
    - Big Distribution
    - Big Collaboration
    - Big Security
    - …

Noting…

- Often similar challenges facing many research domains
- Tools, technologies and methodologies have been/can/are evolving to tackle these challenges
  - That there is a huge amount of work still to be done
    - Don’t believe the hype!!!
    - The pace of research evolution FAR outweighs the pace of IT know-how to deal with the challenges
  - Domain knowledge!!!

焦点

Focal Point

- Examples from different research domains
  - {Computational/Data/Distributed/Collaboration/ Security} bound…
    - Bioinformatics 生物信息学
    - Electronics
    - Astrophysics 天体物理学
    - High Energy Physics
    - Arts and Humanities
    - Biomedical and bioinformatics domain

- BREAK
    - Social Sciences
    - Urban research domain
    - Clinical Domain

Project Portfolio

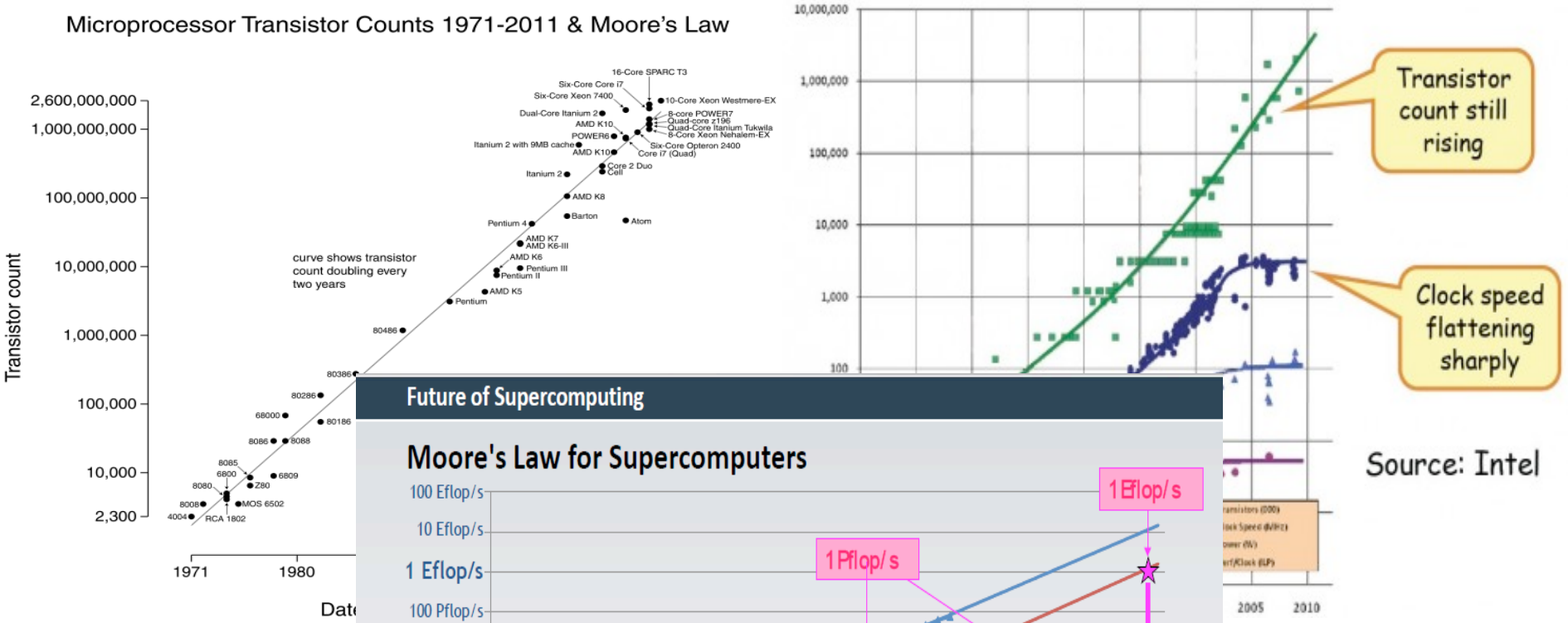
Completed

- National e-Science Centre (I, II, III)
- Dynamic Virtual Organisations for e-Science Education
- Biomedical Research Informatics Delivered by Grid Enabled Services
- GridNet, GridNet-2
- Grid Enabled Microarray Expression Profile Search
- Glasgow early adoption of Shibboleth
- Joint Data Standards Survey
- ESP-Grid
- HPC Compute cluster award // Sun industrial sponsorship
- OGC Collision
- OMII-Security Portlets // OMII-RAVE
- Integrating VOMS and PERMIS for Superior Grid Authorization
- NCeSS
- CESSDA PPP
- Pharming of Therapeutic RNA
- Grid Enabled Occupational Data Environment
- Towards an e-Infrastructure for e-Science Digital Repositories
- Grid enabled Biochemical Pathway Simulator
- Virtual Organisations for Trials and Epidemiological Studies
- A European e-Infrastructure for e-Science Repositories
- Modelling, Inference and Analysis for Biological Systems up to the Cellular Level
- Drug Discovery Portal
- Parliamentary Discourse
- Scots Words and Placenames
- Qvolution stress management survey system
- Advanced Grid Authorisation through Semantic Technologies ShinTau
- AlstromUK VRE
- Grid-enabled Virtual Safe Settings
- Clinical Streaming Transcription Software
- Enhancing Repositories for Language and Literature Researchers (ENROLLER)
- Proxy Credential Auditing Infrastructure for the NGS
- Scottish Bioinformatics Research Network (SBRN)
- Generation Scotland Scottish Family Health Study
- Breast Cancer Tissue Biobank
- Data Management through e-Social Science (DAMES)
- Meeting the Design Challenges of nanoCMOS Electronics (nanoCMOS)
- EU FW7 AvertIT
- EU FW7 EuroDSD
- NeSC Research Platform (NRP)
- NeSC Information Network (NIN)
- ESF Network for Study of Adrenal Tumors
- Scottish Health Informatics Platform for Research (SHIP)
- National E-Infrastructure for Social Simulation (NeISS)
- EU R4SME Diagnosis of Parkinsons Disease (DiPAR)
- Automating River Pollution Detection (CAPIM)
- Endocrine genomics Virtual Laboratory (endoVL)
- DSDNetwork Australasia

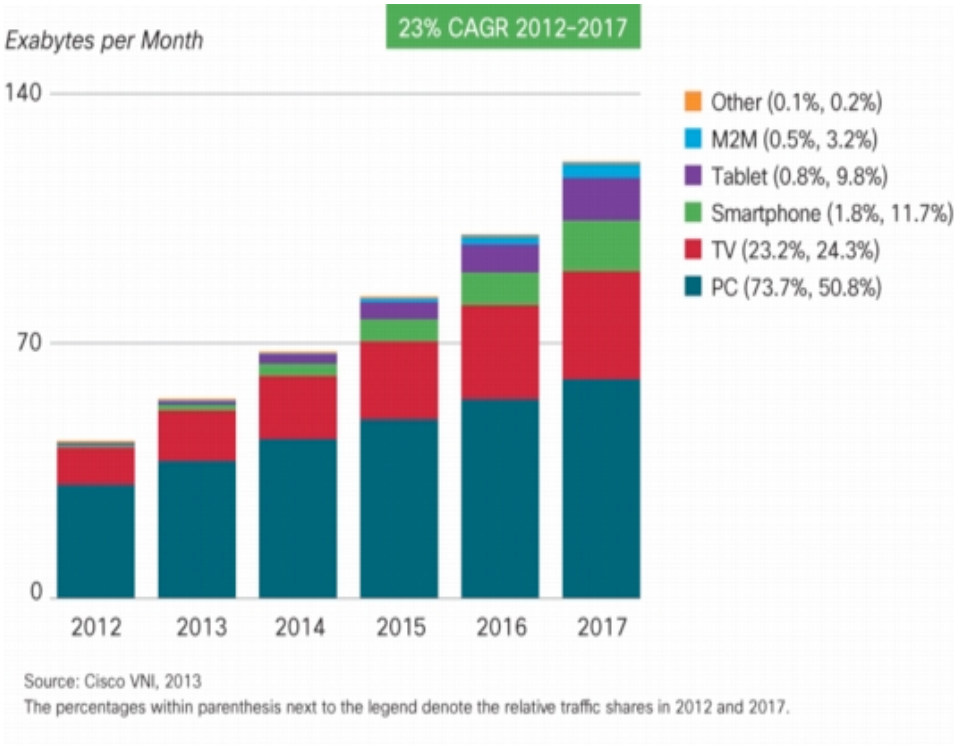
On-Going

- EU European Platform for Study of Wolfram, Alstrom, Bardet Biedl (EuroWABB)
- Multicenter prospective study of biochemical profiles of monoamine-producing tumors (PMT Study)
- European Society of Hypertension Study on Pheo/PGL
- International DSD
- EU FW7 European Network for Study of Adrenal Tumors Cancer Research Platform (ENSAT-CANCER)
- VicHealth Health Indicators and Spatial Objective Data
- National Spinal Injury Research Platform
- Australian Urban Research Infrastructure Network (AURIN)
- Epilepsy e-Learning portal
- Type-1 Diabetes study of environmental factors on onset of T1D
- Australian Diabetes Data Network (ADDN)
- International Niemann-Pick A, B and C Registry
- Carlton Connect Data Journalism in the Big Data Era
- FAMIAN - Combined 18F-fluorodeoxyglucose positron emission tomography and 123I-Iodometomidate Imaging for Adrenal Neoplasia
- Melbourne Genomics Health Alliance (variant DB)
- NeCTAR Cloud Encryption/Decryption and Secure Deletion
- CRE for Protection of Pancreatic Beta Cells
- Airbox (Atmospheric Physics and Climate Research)
- NESP Clean Air and Urban Environments
- Application of omics-based strategies for improved diagnosis and treatment of endocrine hypertension
- Youth alcohol consumption database and mobile app
- LIDAR Data Analytics Research Environment
- Type-1 Diabetes Clinical Research Network
- American Asian Australian Adrenal Alliance
- International League Against Epilepsy
- Platform for Research Software Solutions (PRESS)
- Mobile applications for the Environmental Determinants of Islet Autoimmunity
- Secure Data Solutions for the Biomedical Communities of the Cloud
- Metabolomics Sample Management and Processing Platform
- Linked Data PolicyHub Stage II: Urban & Regional Planning & Communications
- Australian Genomics Health Alliance
- Melbourne Genomics Health Alliance
- Australian Diabetes Data Network – Phase II (ADDN2)
- Helicopter advanced training system, Australian Department of Defence
- Public Records Office Victoria Data Management Solutions
- Complex System Modelling Platform and GPU utilisation
- Public Records Office Victoria Data Management Solutions Follow-Up Grant
- VicHealth 2016 Indicators API
- Helicopter advanced training system Phase II, Australian Department of Defence
- Twitter data analytics for business
- Mobile Applications for Patients with Neuroendocrine Tumours
- Systems Genomics Support Platform
- SWARM: Smartly-aggregated Wiki-style ARgument Marshalling (SWARM)
- ORCA Cognitive Assessment Platform
- VicSpin Victoria-wide Flu Surveillance System
- ElectraNetLIDAR // VectorNZ Lidar

Compute Scaling



Network Scaling



Source: Cisco VNI, 2013

The percentages within parenthesis next to the legend denote the relative traffic shares in 2012 and 2017.

Table 1. The VNI Forecast Within Historical Context

Year	Global Internet Traffic
1992	100 Gigabytes per Day
1997	100 Gigabytes per Hour
2002	100 Gigabytes per Second
2007	2,000 Gigabytes per Second
2012	12,000 Gigabytes per Second
2017	35,000 Gigabytes per Second

Source: Cisco VNI, 2013

Table 6. Table A-1 Global IP Traffic, 2012-2017

IP Traffic, 2011-2016							
	2012	2013	2014	2015	2016	2017	CAGR 2012-2017
By Type (PB per Month)							
Fixed Internet	31,339	39,295	47,987	57,609	68,878	81,818	21%
Managed IP	11,346	14,679	18,107	21,523	24,740	27,668	20%
Mobile data	885	1,578	2,798	4,704	7,437	11,157	66%
By Segment (PB per Month)							
Consumer	35,047	45,023	56,070	68,418	82,683	98,919	23%
Business	8,522	10,530	12,822	15,417	18,372	21,724	21%
By Geography (PB per Month)							
Asia Pacific	13,906	18,121	22,953	28,667	35,417	43,445	26%
North America	14,439	18,788	23,520	28,667	34,457	40,672	23%
Western Europe	7,722	9,072	10,568	12,241	14,323	16,802	17%
Central and Eastern Europe	3,405	4,202	5,167	6,274	7,517	8,844	21%
Latin America	3,397	4,321	5,201	5,975	6,682	7,415	17%
Middle East and Africa	701	1,049	1,483	2,013	2,659	3,465	38%
Total (PB per Month)							
Total IP traffic	43,570	55,553	68,892	83,835	101,055	120,643	23%

Source: Cisco VNI, 2013

Data Past

- From tablets, to papyrus (纸莎草), to books
  - (quite adequate for several thousand years)
- Enter silicon transistors (硅晶体管) circa (大约) 1960
  - punch cards,
  - punched streamer tape,
  - magnetic tape,
  - floppies,
  - …
    - ~RIP!

Data Present

- Data Storage today
  - CDs,
  - DVDs,
  - local (computer) hard disks,
  - shared storage,
  - tape storage.
  - mobile storage,
  - The Internet!
    - Dropbox
    - Google
    - Clouds
    - …









泛濫

Data Deluge

- The combi … at approxima …

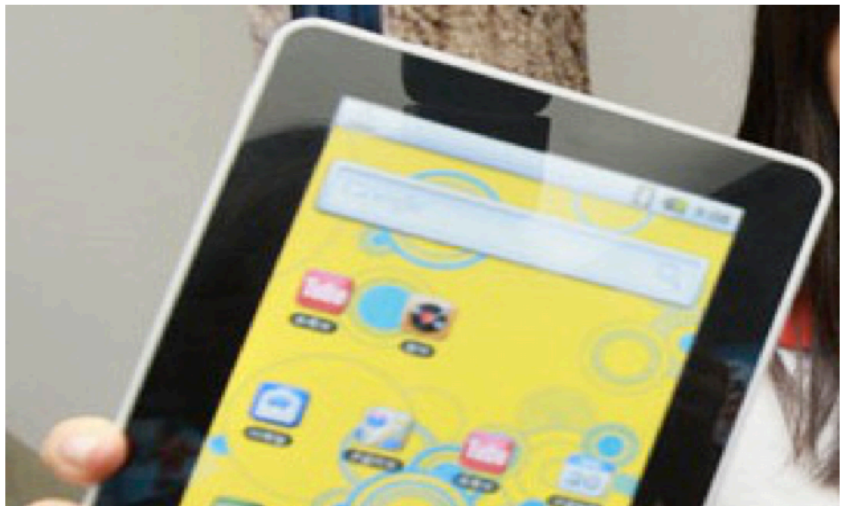
… (March 2007) … igital Universe

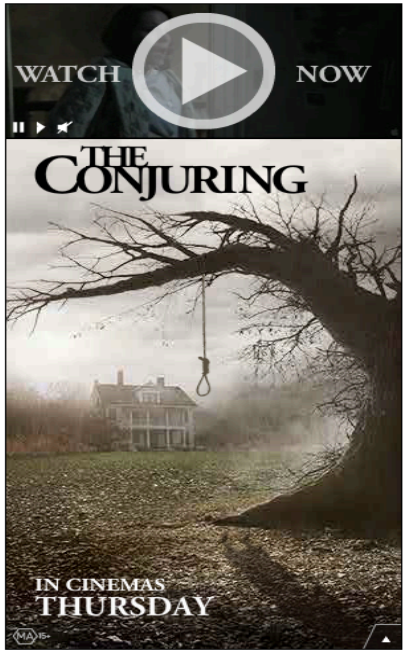
- The total a… This is 48… 2012.

Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003

MG SIEGLER

Wednesday, August 4th, 2010 — 2 Comments





… ta Corporation

- In 2008, A… …anuary 2010). …n Consumers”

- By 2015, … …uTube, IPTV, and … …s larger than it was …

… h_zettabyte/

Naturally, all of this information helps Google. But he cautioned that just because companies like his can do all sorts of things with this information, the more pressing question now is if they *should*. Schmidt noted that while technology is neutral, he doesn't believe people are ready for what's coming.

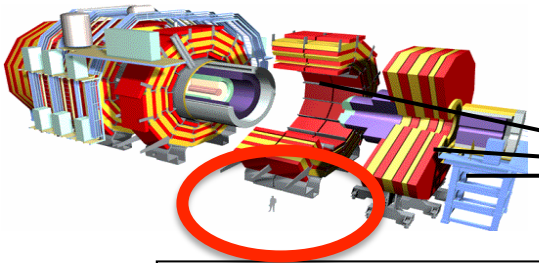
"*I spend most of my time assuming the world is not ready for the technology revolution that will be happening to them soon*," Schmidt said.

Data Intensive / Data driven Research

- Researchers need tools, methodologies
  - To search for/discover data
  - To use/analyse data
  - To share data
  - To store data
  - To track data
  - To destroy data
  - To move data around
  - To check authenticity of data
  - To visualise data
  - To overcome issues of data heterogeneity
  - …

    … and this should be tailored to the researchers needs!!!

Compute Infrastructure for High Energy Physics



~25 PB per year

There is a "bunch crossing" every 25 nsecs.
There are 100 "triggers" per second
Each triggered event is ~1 MByte in size

Online System

1 TIPS is approximately 25,000 SpecInt95 equivalents

Offline Processor Farm
~20 TIPS

Tier 0

CERN Computer Centre

HPSS

Tier 1

France Regional Centre

Germany Regional Centre

Italy Regional Centre

FermiLab ~4 TIPS

ScotGrid

~622 Mbits/sec

Tier 2

Caltech
~1 TIPS

Tier2 Centre
~1 TIPS

Physicists work on analysis "channels".
Each institute will have ~10 physicists working on one or more channels; data for these channels should be cached by the institute server

Physics data cache

Institute
~0.25TIPS

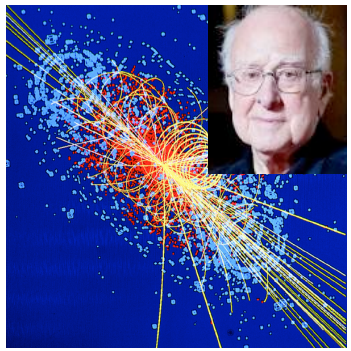
~1 MBytes/sec

Tier 4

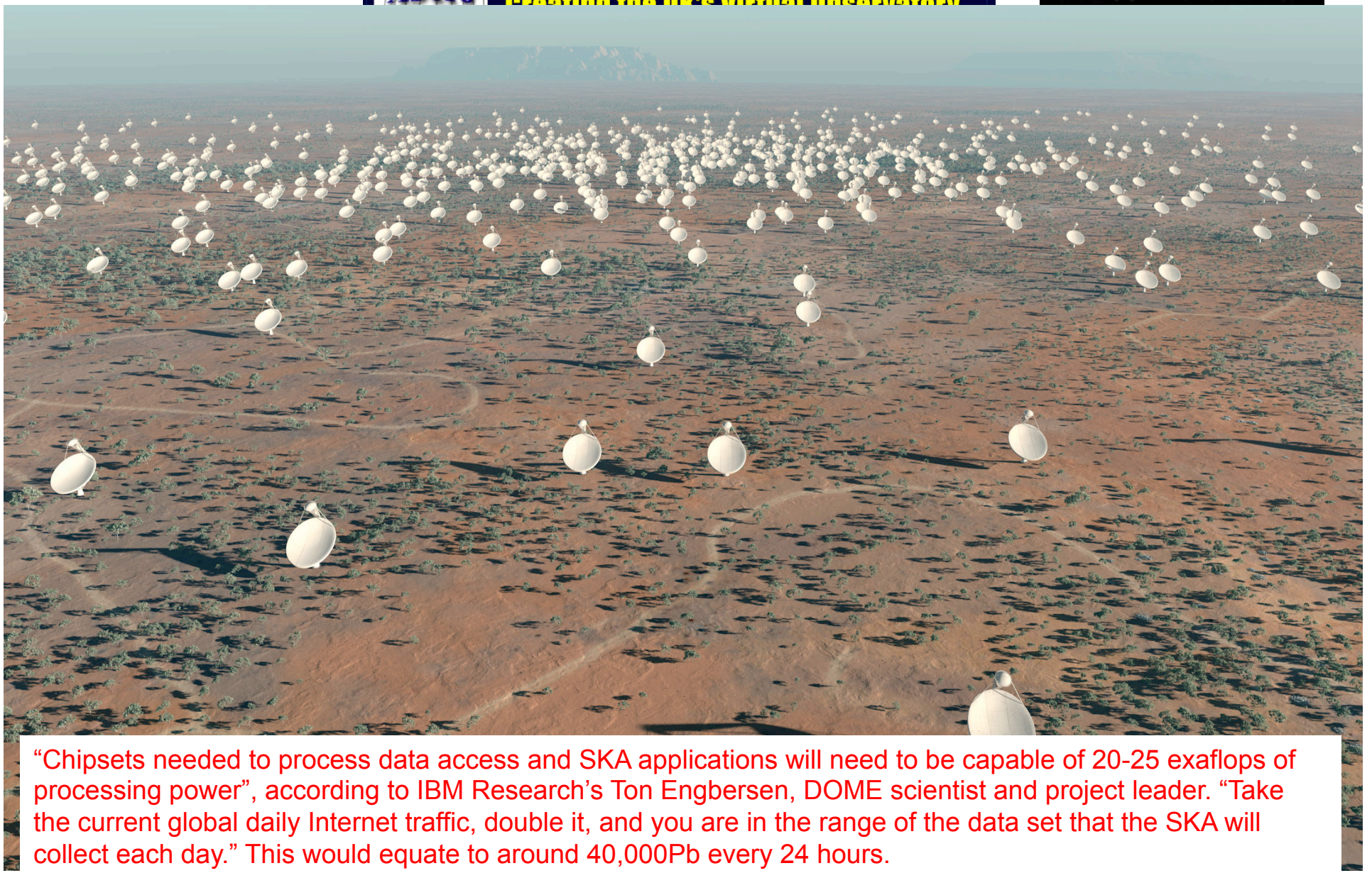
Raw data per event ~1 Mb, produced at a rate of about 40 million events per second.

- Basic filtering reduces to 100,000 events per second
- Advanced filtering to 100 or 200 events per second

This two-stage data re-processing is performed several times a year on all data acquired since the LHC start-up.

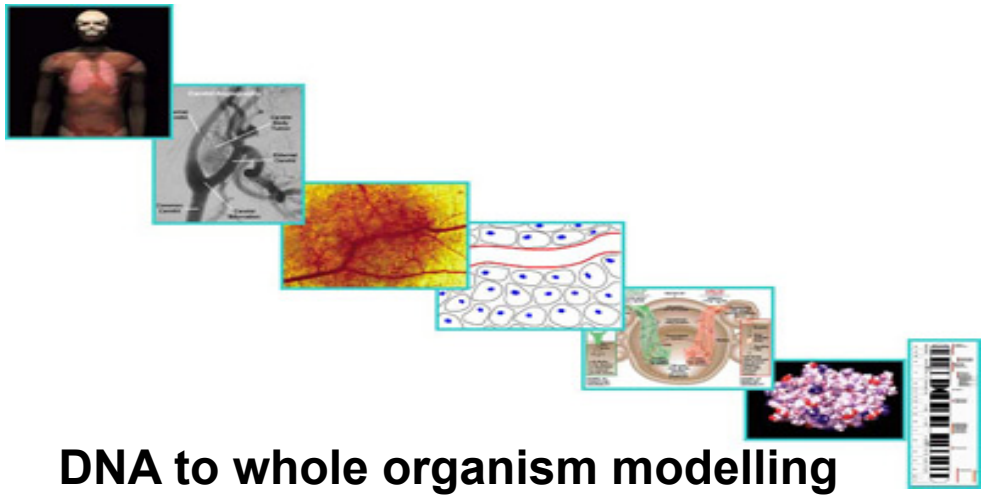


Mapping the Skies

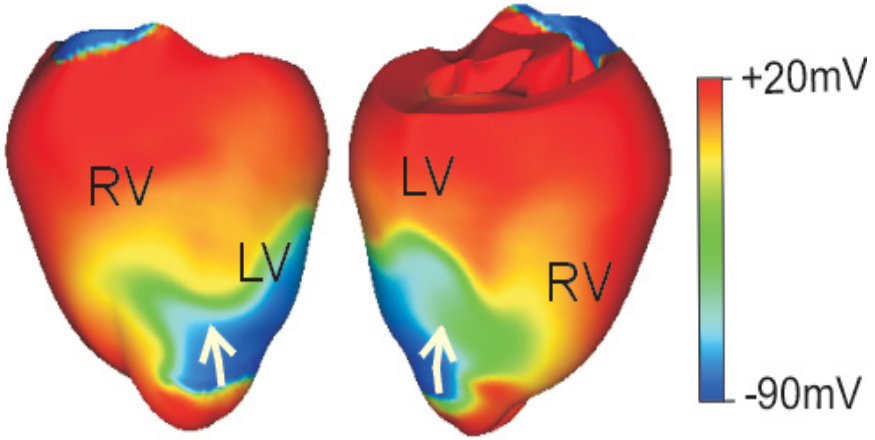


"Chipsets needed to process data access and SKA applications will need to be capable of 20-25 exaflops of processing power", according to IBM Research's Ton Engbersen, DOME scientist and project leader. "Take the current global daily Internet traffic, double it, and you are in the range of the data set that the SKA will collect each day." This would equate to around 40,000Pb every 24 hours.

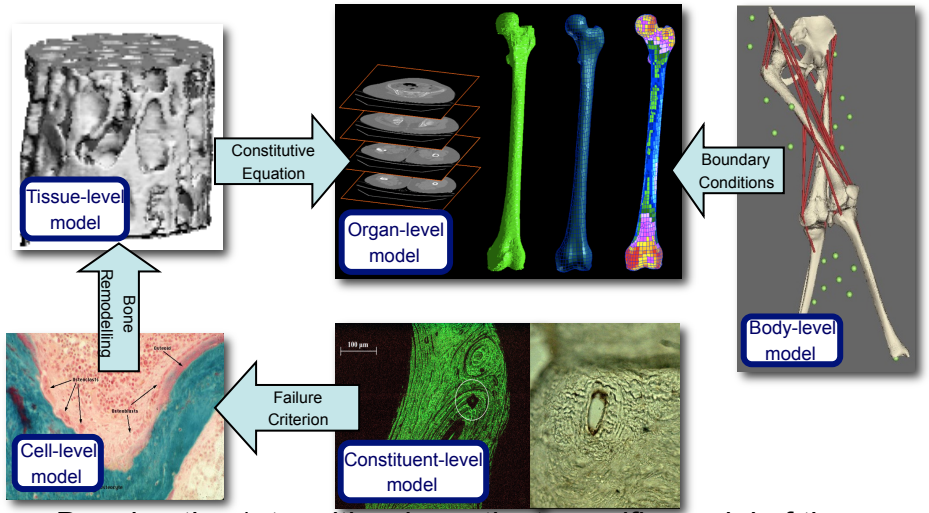
Macro-micro Simulations



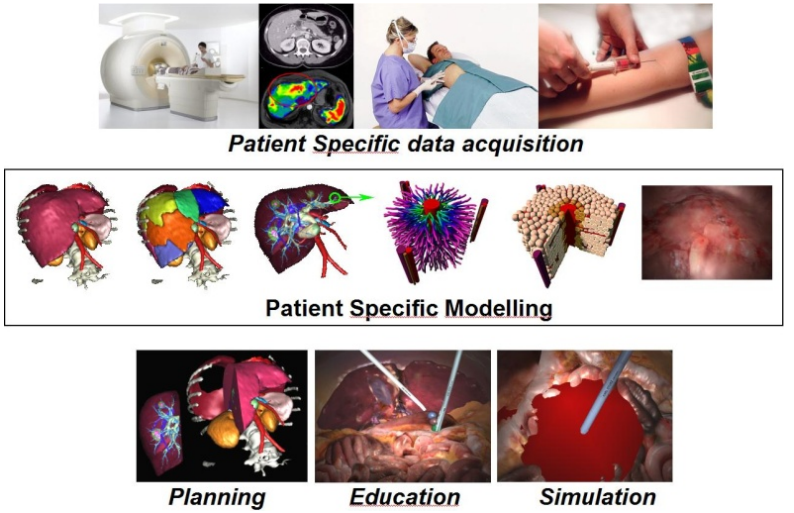
DNA to whole organism modelling



Direction of wave propagation

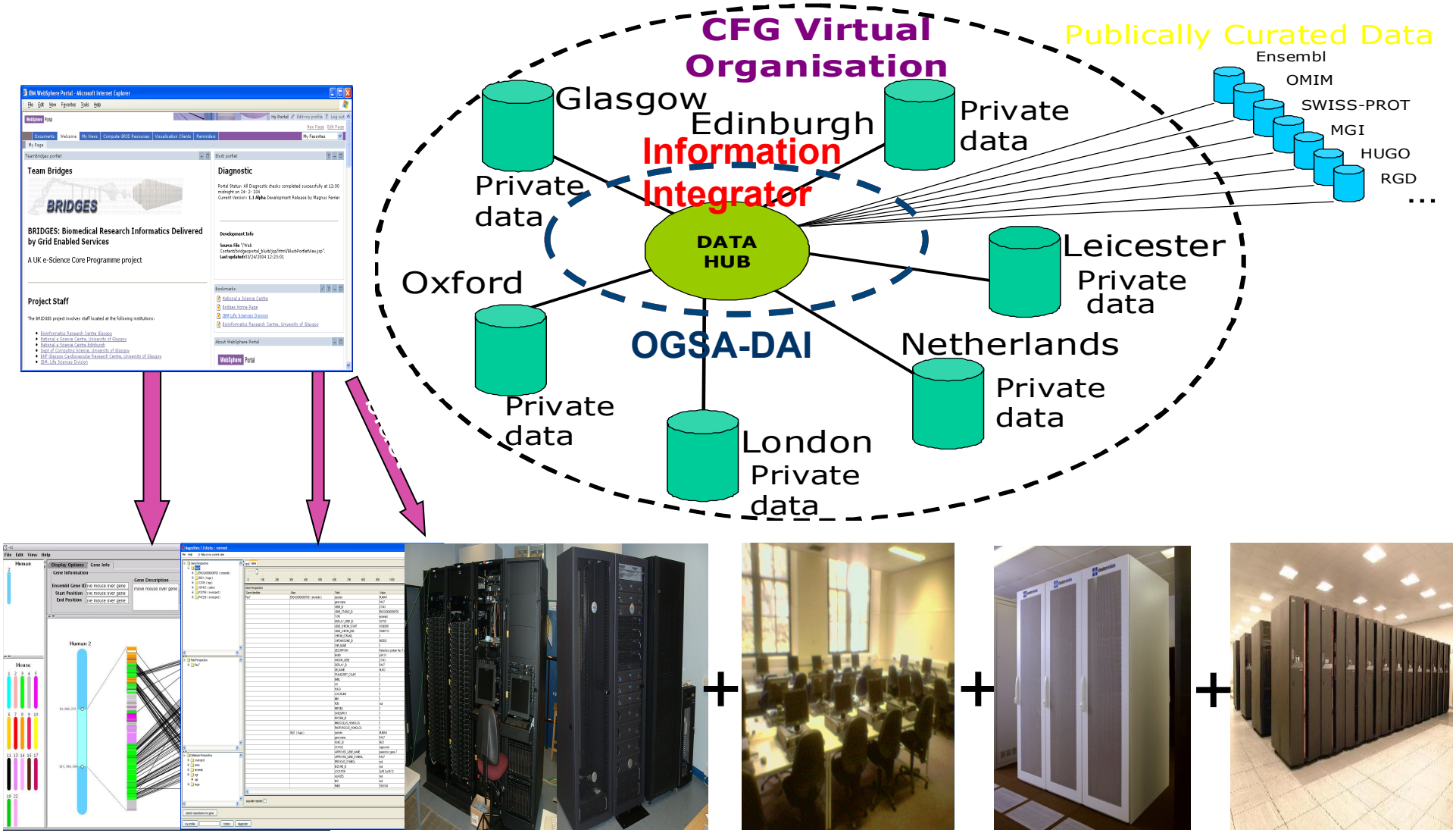


Develop the 1st multiscale, patient-specific model of the musculoskeletal system and use model for diagnosis, prognosis and treatment of osteoporotic fractures

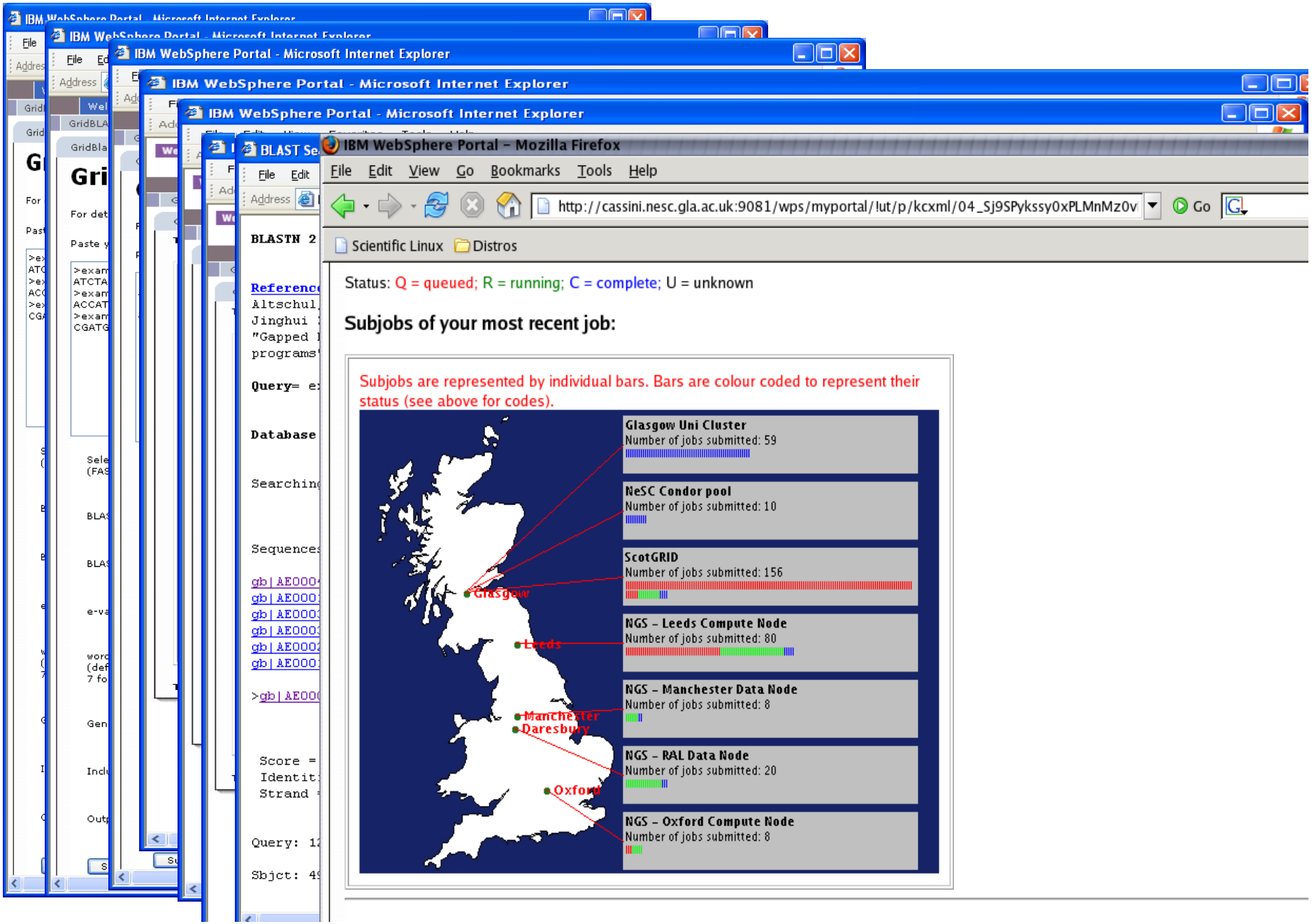


PAtient Specific Simulation and PreOperative Realistic Training for liver surgery

Example of BRIDGES Project



Grid Blast Interface



Meeting the Design Challenges of Nano-CMOS Electronics

e-Science Pilot Project (EPSRC)

R. Sinnott (e-Science Director)

Resources

£3.7M EPSRC; £4.4M FEC
£5.8M incl. industrial contributions

11 PDRAs (7 + 4)
7 PhD

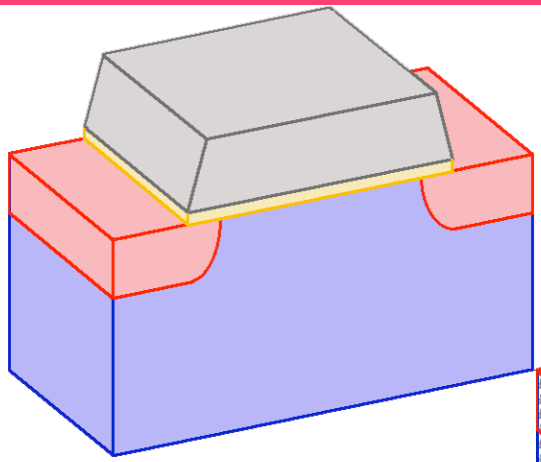
Industrial partners

ARM, wolfson microelectronics, SYNOPSYS, freescale semiconductor, National Semiconductor, FUJITSU

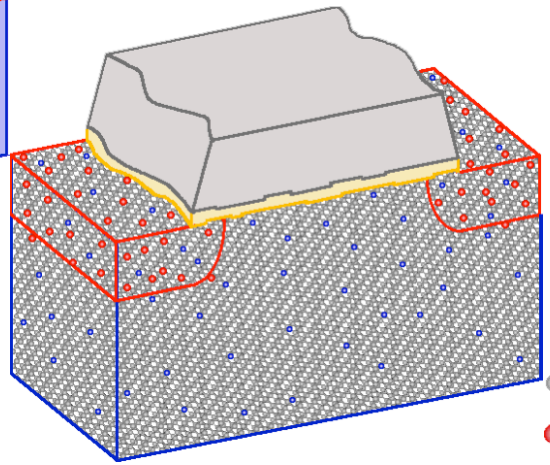
University partners

半导体

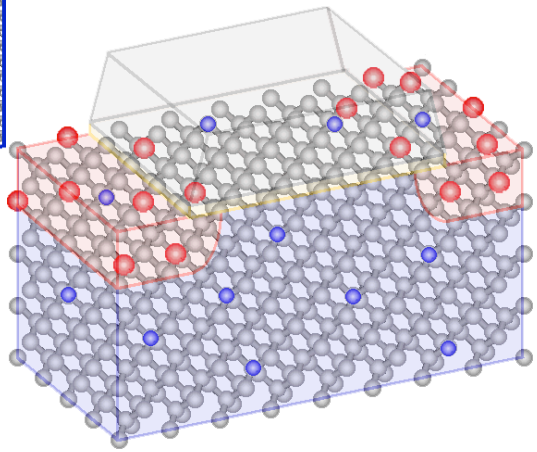
Semiconductor device variability



Historic simulation paradigm

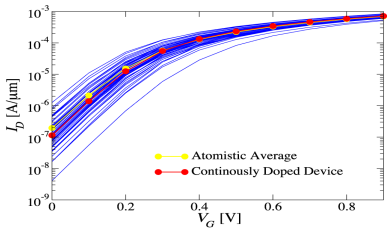


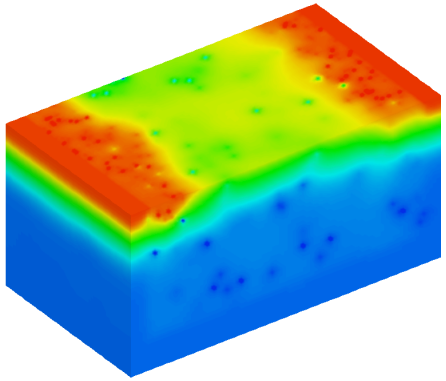
A 22 nm MOSFET
In production 2011

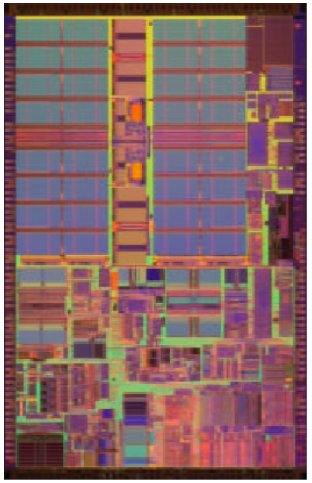


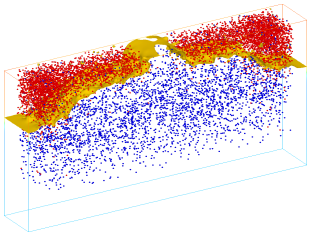
A 4.2 nm MOSFET
In production 2023

Challenges of NanoCMOS Design





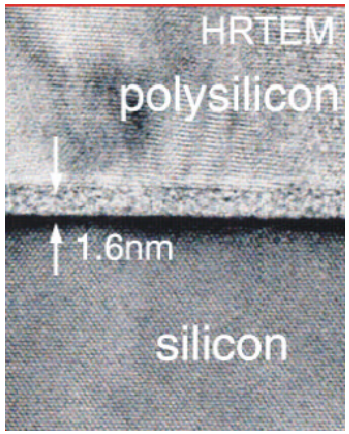


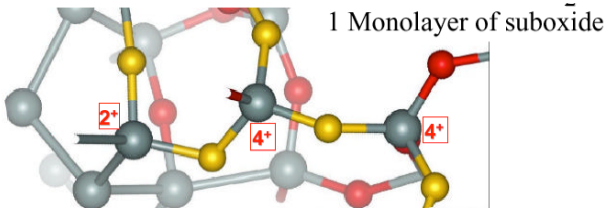


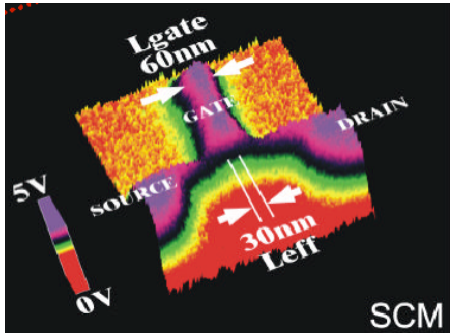
3D

+

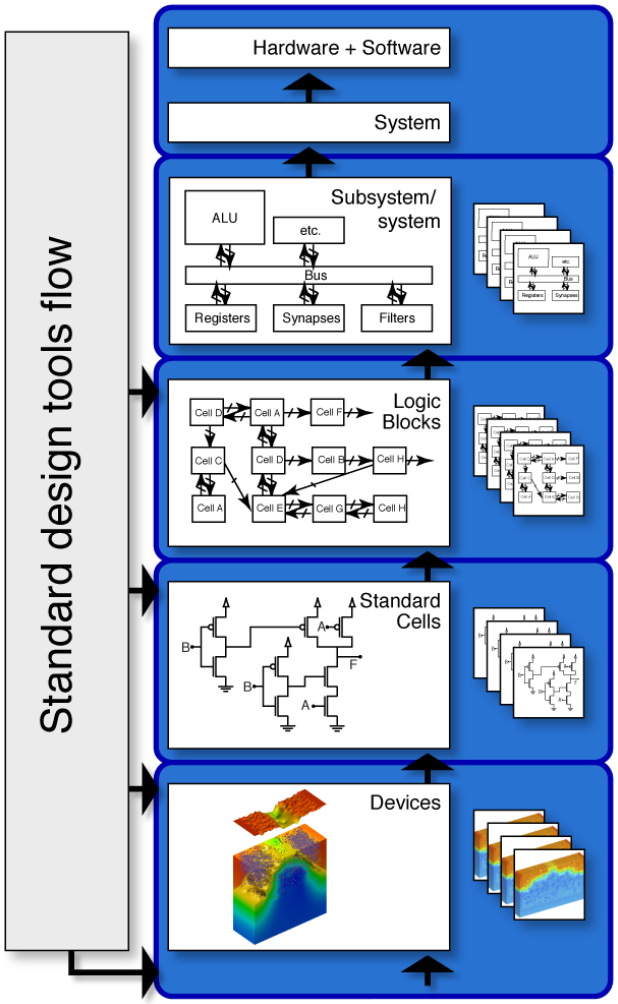
Statistical







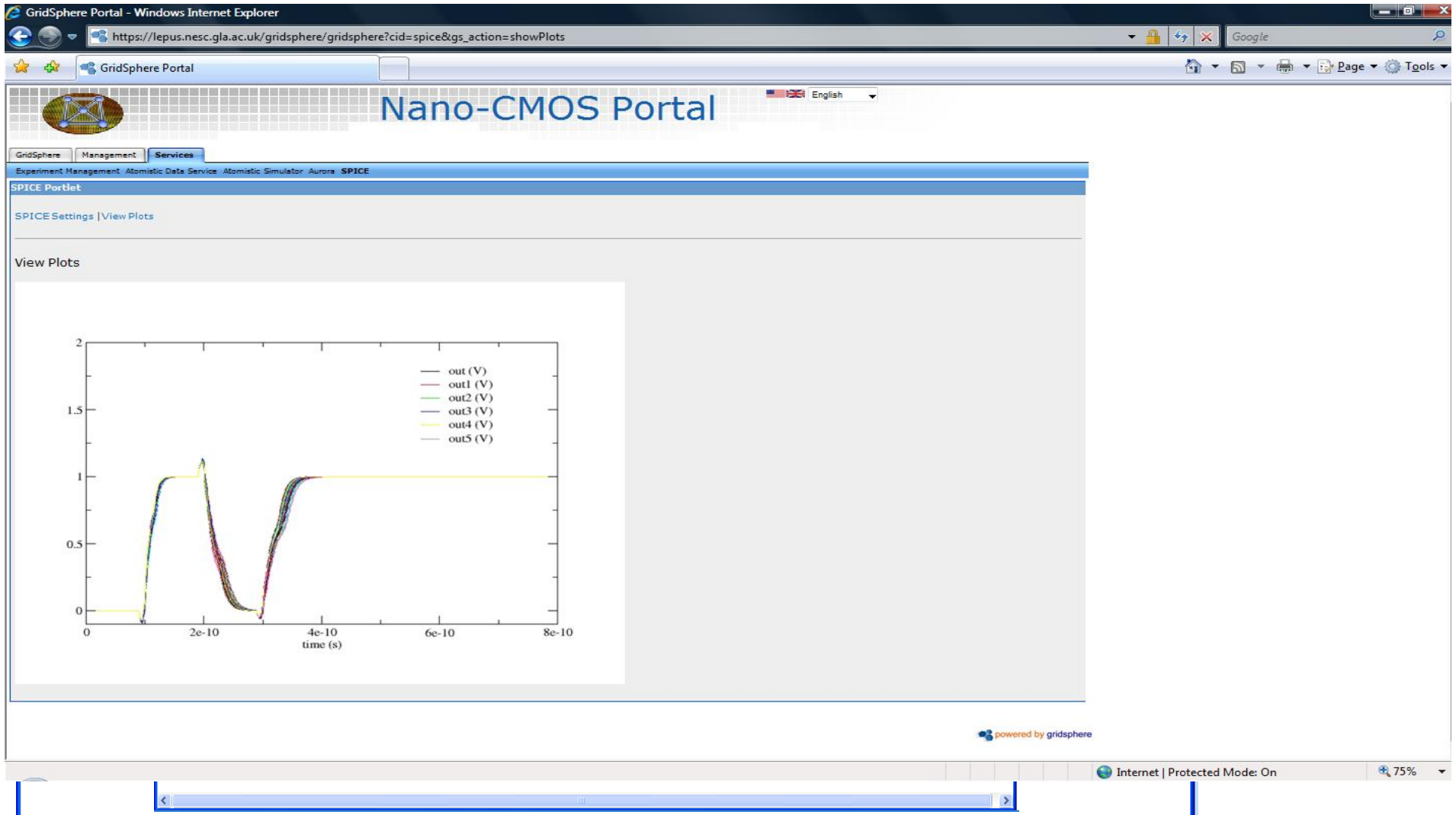
Challenges Hierarchical statistical system simulations



- Very large device and circuit simulations
  - *3D devices*
  - *10^5 circuit components*
- Large statistical samples
  - *1000 - 100000 3D simulations - 4D*
  - *1000 - 100000 circuit simulations*
- Complex flow and storage of data
  - *Many files per simulation*
  - *Metadata capture and data provenance*
- Collaboration between 5 partners
  - *Multidisciplinary background*
  - *Complex data exchange*
- Stringent security requirements
  - *Commercial IP*
  - *Expensive software licenses*

E-Experiences

- We started with a secure portal and a wiki!!!”



But ended up with…

- … a command-line based solution

This community

Security solution

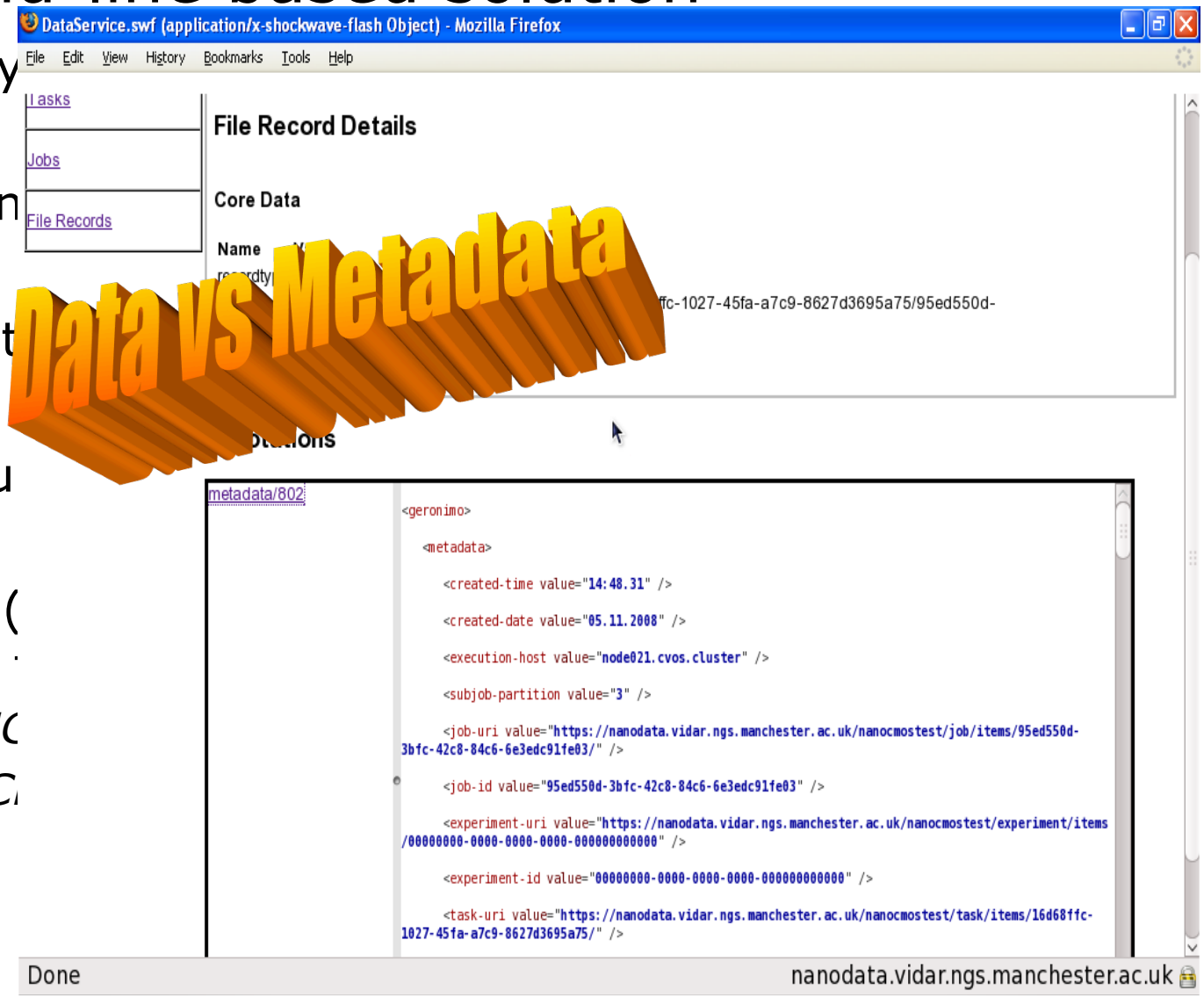
Secure, distribut

Meta-data captu

Job submission (
massive (at the

- *ScotGrid, NC*
- *Millions of C*

The –g flag!!!

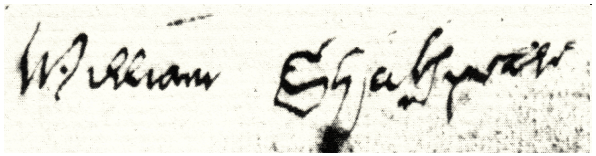


Arts and Humanities

- Data resources galore 大量
- Language and literature research
  - Linguists
  - Philologists
  - Historians
  - Societal observers
  - Culturists
  - …
- Written resources
- Spoken resources
- Video resources

Good old Bill

- To be or not to be…
  - Shakespeare, Shakespere, Shakespear, Shakspeare, Shackspeare, Shakspere, Shackespeare, Shackspere, Shackespere, Shaxspere, Shexpere, Shakspe~, Shaxpere, Shagspere, Shaksper, Shaxpeare, Shaxper, Shake-speare, Shakespe, Shakp
  - Inside London, Outside London,
  - Printed, cited, referenced
    - 8 signatures – all different!



Most “Excellent” Dude

- seldlic syndorlic til wlitig special breme gradely noble thriven and thro burly singular dainty gentle proper beforepassing goodly daintiful thriven thriving priceprise vounde virtuous curious principal fine gay rare singlar egregious gallant eximious jolly jelly braw brave stamming surprising excelling phoenix royal of worth rare admirable sublimated valiant excellent twanging topgallant lovely prestantious splendid sterling spanking hogenhogan pure licking rattling tearing soaring famous yrare pure and daisy immense capital elegant trimming gallows budgeree crack dandy smicksmack

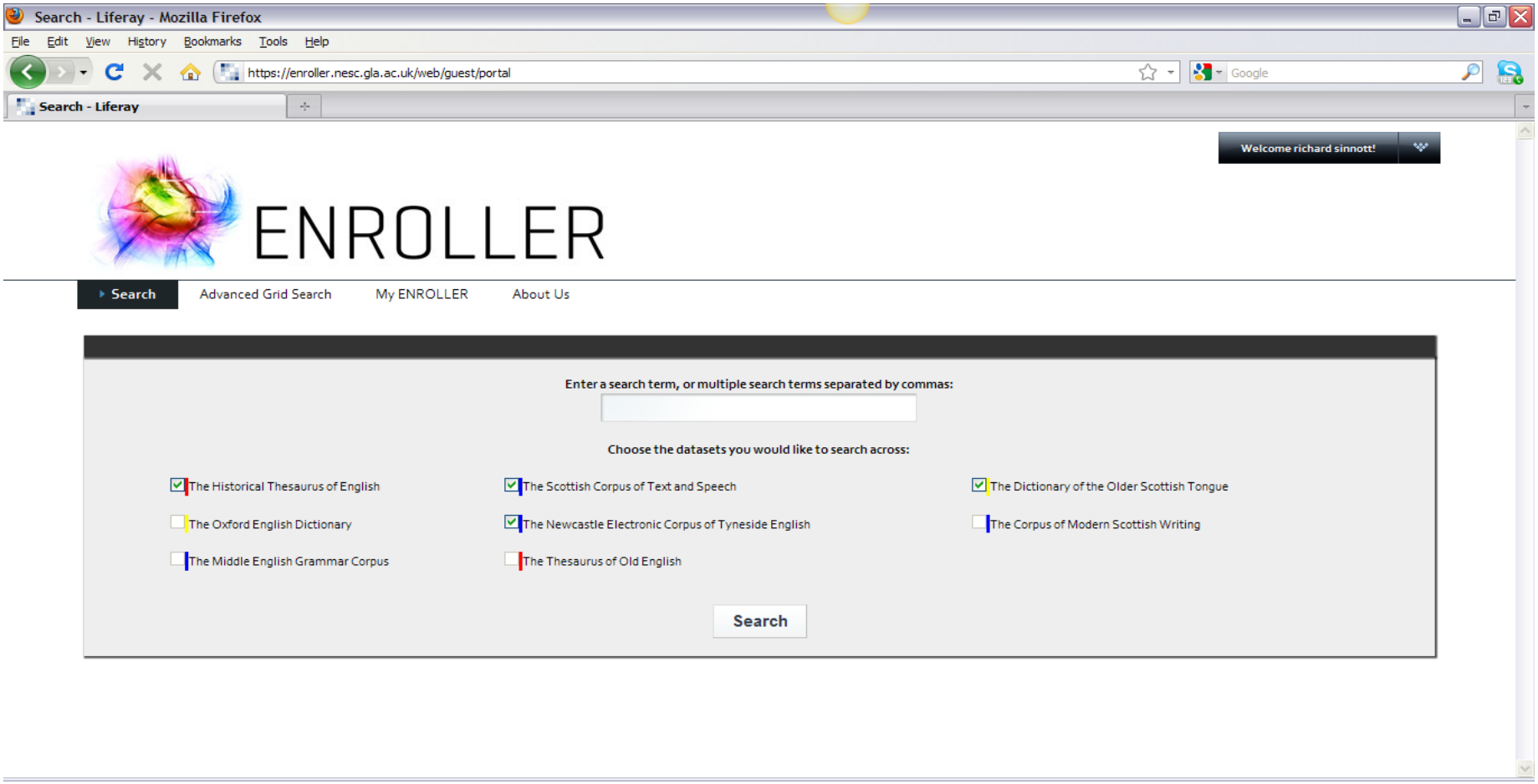
- 02.01.15.07.08.06 adj, Historical Thesaurus of English

I ain’t “Through” yet…

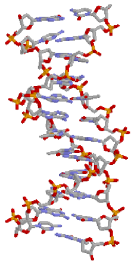
Thro(u)ch, Through(e, Throw(e, Thorow(e, prep. Also: throuche, throwch(e, throwgh(e, thro(u)cht, throwcht, throught, threwch, thrwch, throuthe, thruch(t, thrugh(t, thruth(t, throche, throgh(e, throiche, throicht, throchge, thro(u)g, throu, throue, threu, threw, thru, thrw, thro, troch(t, trouch, trew, troithe, trhow, thorowch, thorycht, thorrow, thoro, thurrow

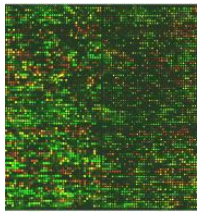
Thro(u)ch (preposition), DOST

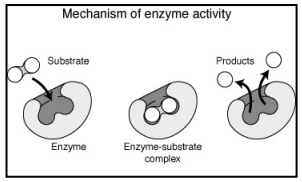
ENROLLER

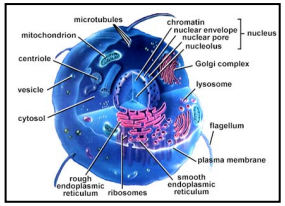

Search - Liferay - Mozilla Firefox
File Edit View History Bookmarks Tools Help
https://enroller.nesc.gla.ac.uk/web/guest/portal
Search - Liferay
Welcome richard sinnott!
ENROLLER
Search
Advanced Grid Search
My ENROLLER
About Us
Enter a search term, or multiple search terms separated by commas:
Choose the datasets you would like to search across:
The Historical Thesaurus of English
The Scottish Corpus of Text and Speech
The Dictionary of the Older Scottish Tongue
The Oxford English Dictionary
The Newcastle Electronic Corpus of Tyneside English
The Corpus of Modern Scottish Writing
The Middle English Grammar Corpus
The Thesaurus of Old English
Search

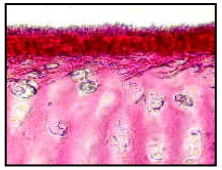
The e-Health Future…

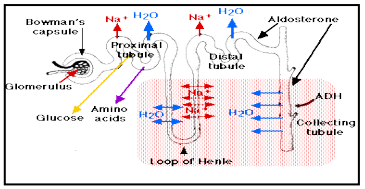






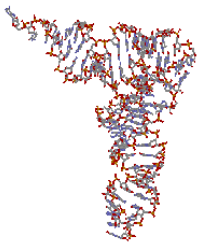


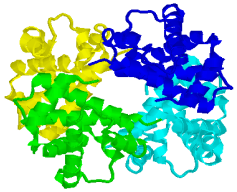


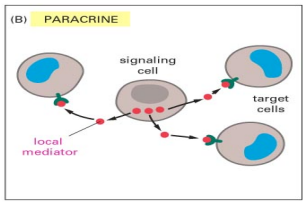


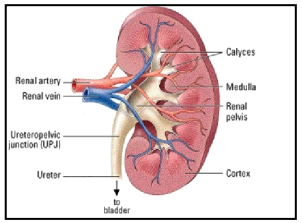
Nucleotide sequences
Nucleotide structures
Gene expressions
Protein Structures
Protein functions
Protein-protein interaction (pathways)
Cell
Cell signalling
Tissues
Organs
Physiology
Organisms
Populations

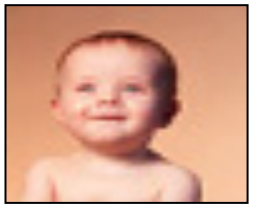
Security

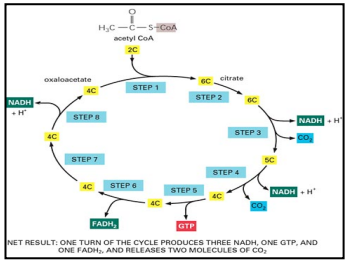












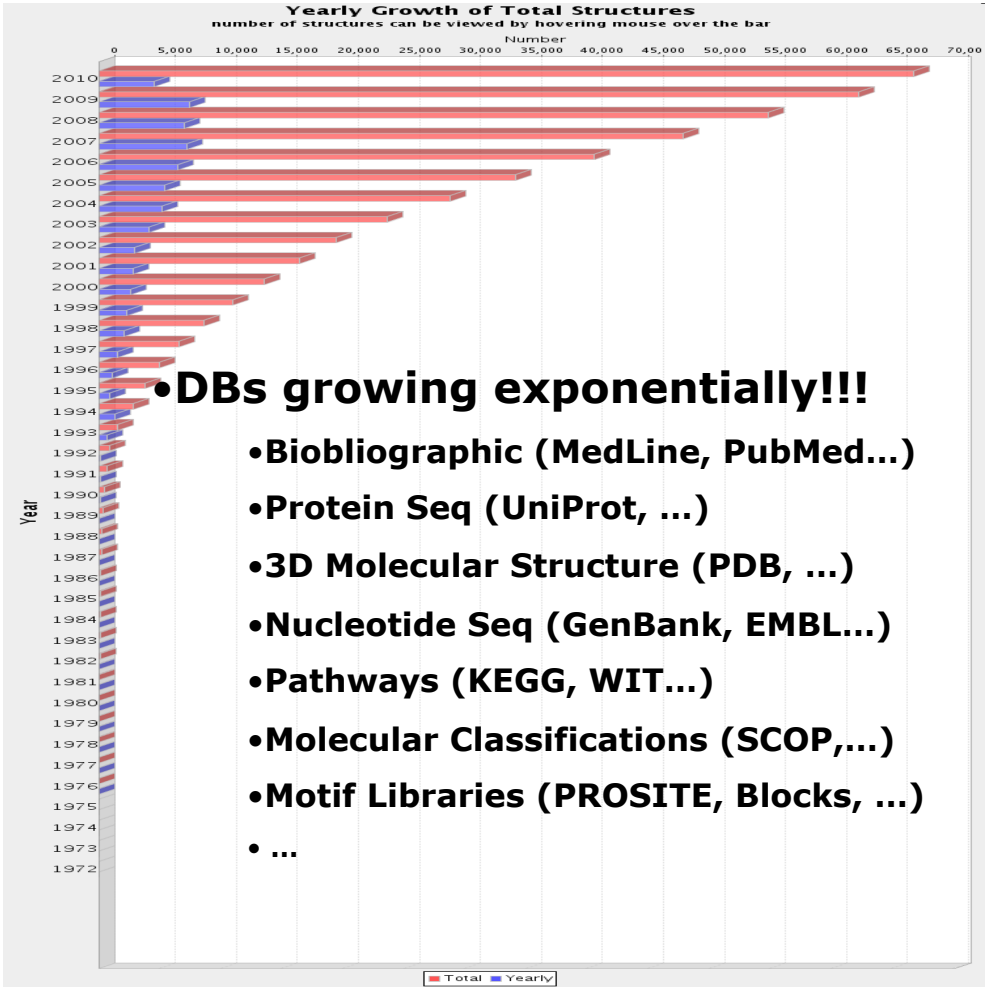
+ environmental, social, geographic …

Life Sciences

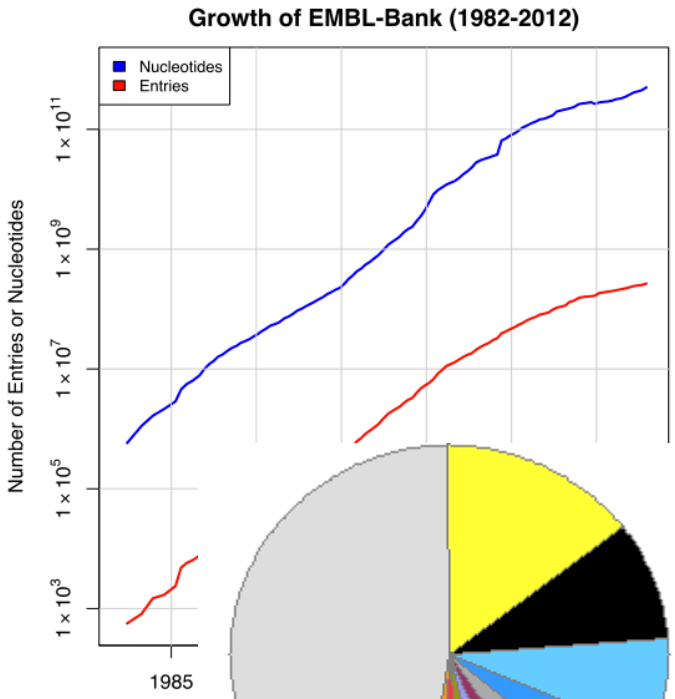
- Extensive Research Community
  - Parkville Precinct for example
- Many people care about them
  - Health, Food, Environment – truly interdisciplinary!
- Interacts with virtually every discipline
  - Physics, Chemistry, Maths/Stats, Nano-engineering, …
- Thousands of databases relevant to bioinformatics (and growing!)
  - Heterogeneity, Interdependence, Complexity, Change, …
- Some of the Big Questions/Challenges
  - How does a cell work?
  - How does a brain work?
  - How does an organism develop?
  - Why do people who eat less tend to live longer?
  - …

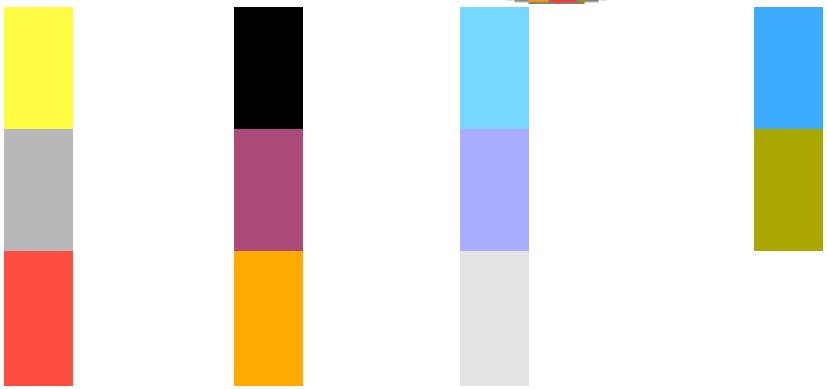
Database Growth

PDB Content Growth



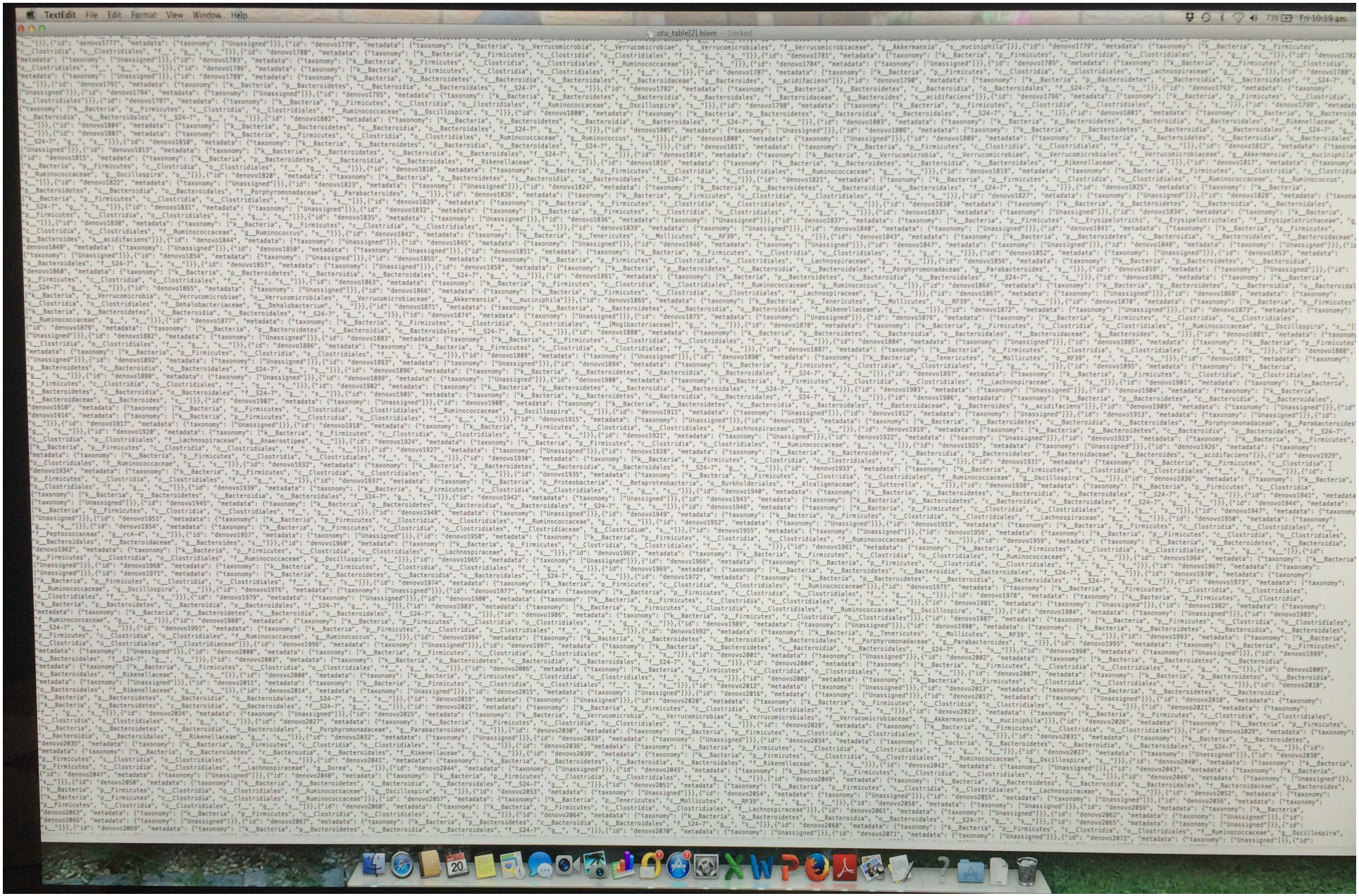
- **DBs growing exponentially!!!**
  - **Biobliographic (MedLine, PubMed…)**
  - **Protein Seq (UniProt, …)**
  - **3D Molecular Structure (PDB, …)**
  - **Nucleotide Seq (GenBank, EMBL…)**
  - **Pathways (KEGG, WIT…)**
  - **Molecular Classifications (SCOP,…)**
  - **Motif Libraries (PROSITE, Blocks, …)**
  - **…**





More (and more and more) genomes…

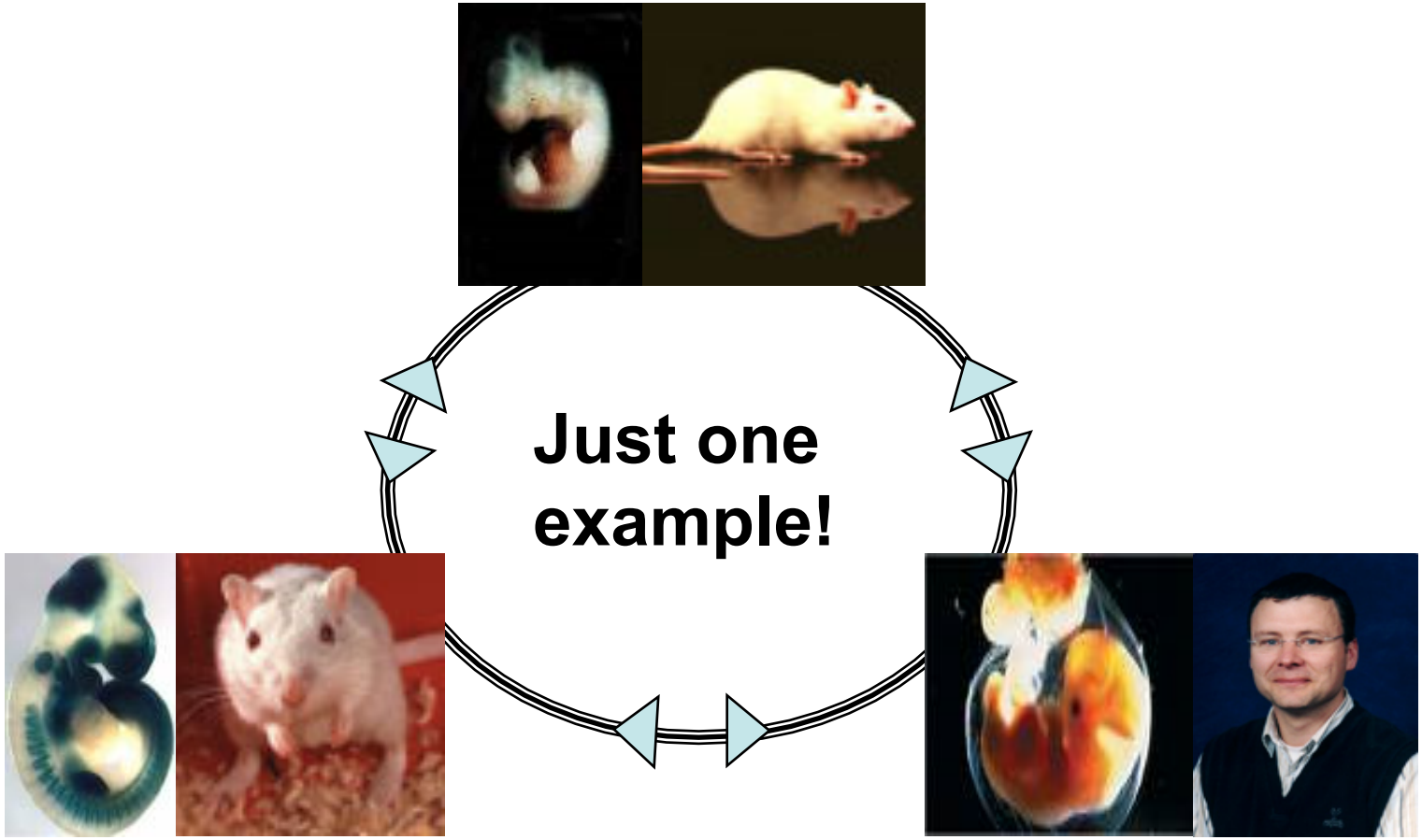
Distributed, completely heterogeneous data



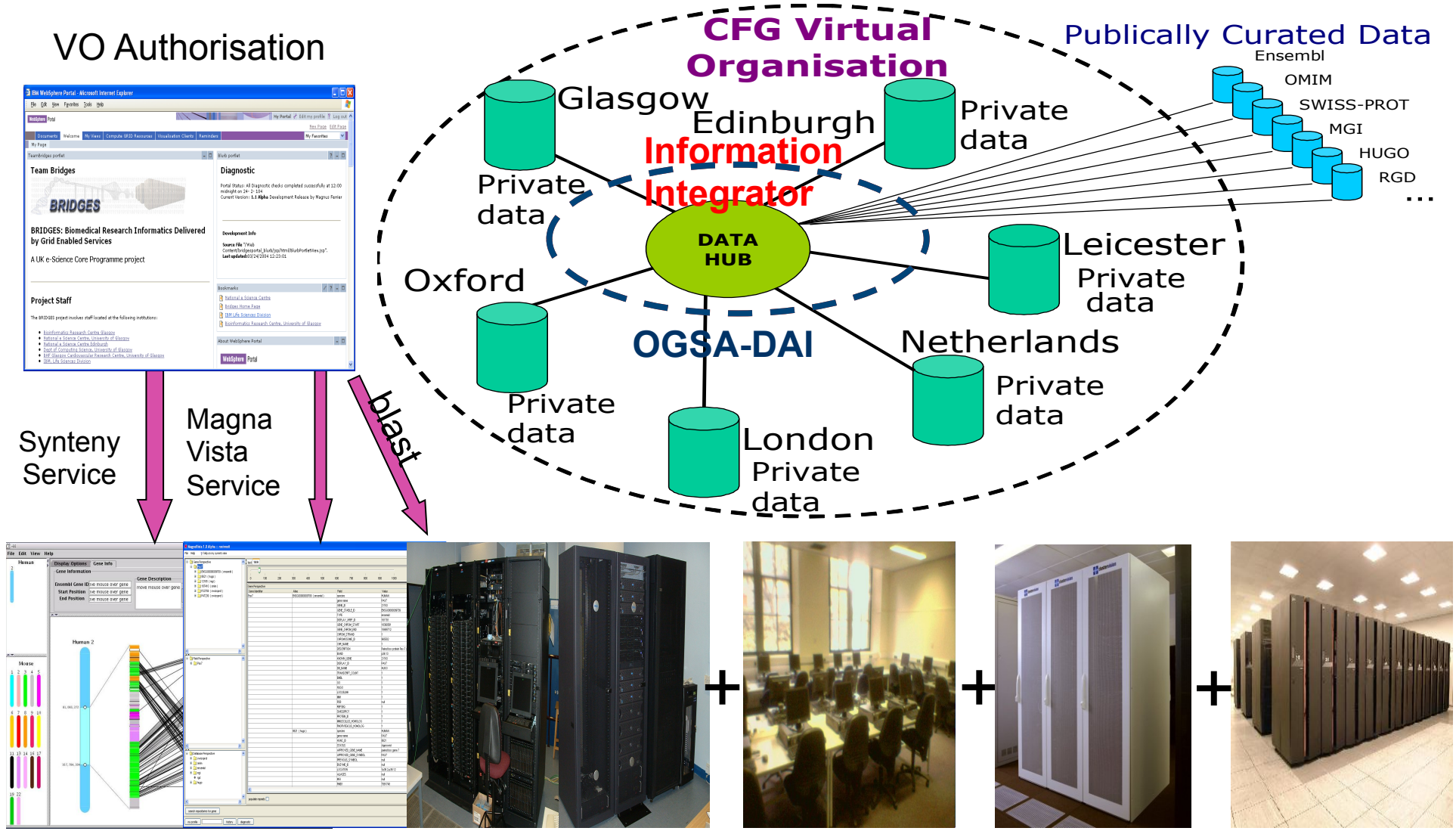
Messy


GPL96 - Notepad
File Edit Format View Help
Next Generation Sequencers
~1TB data

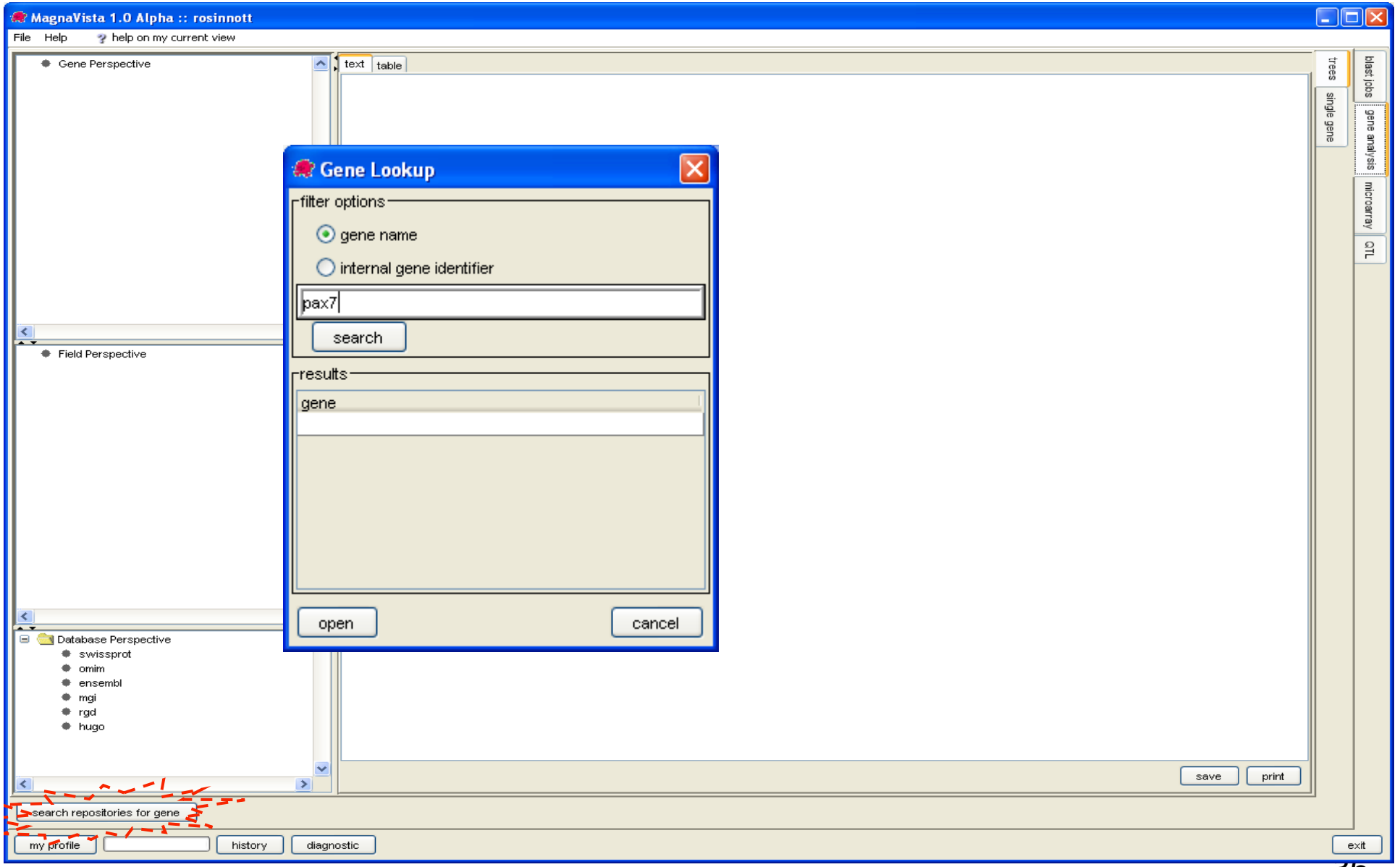
Translational Research



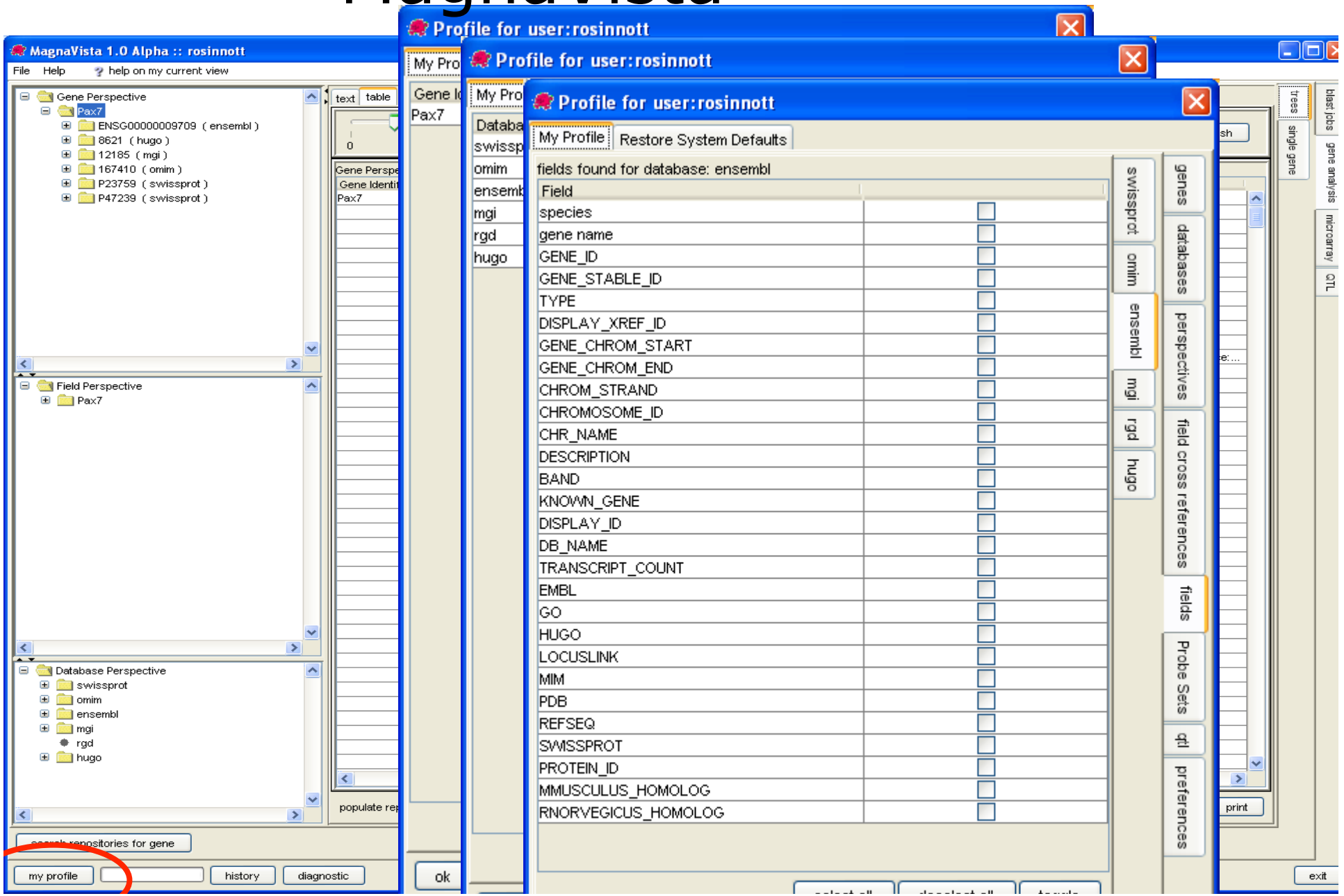
BRIDGES Project (again)



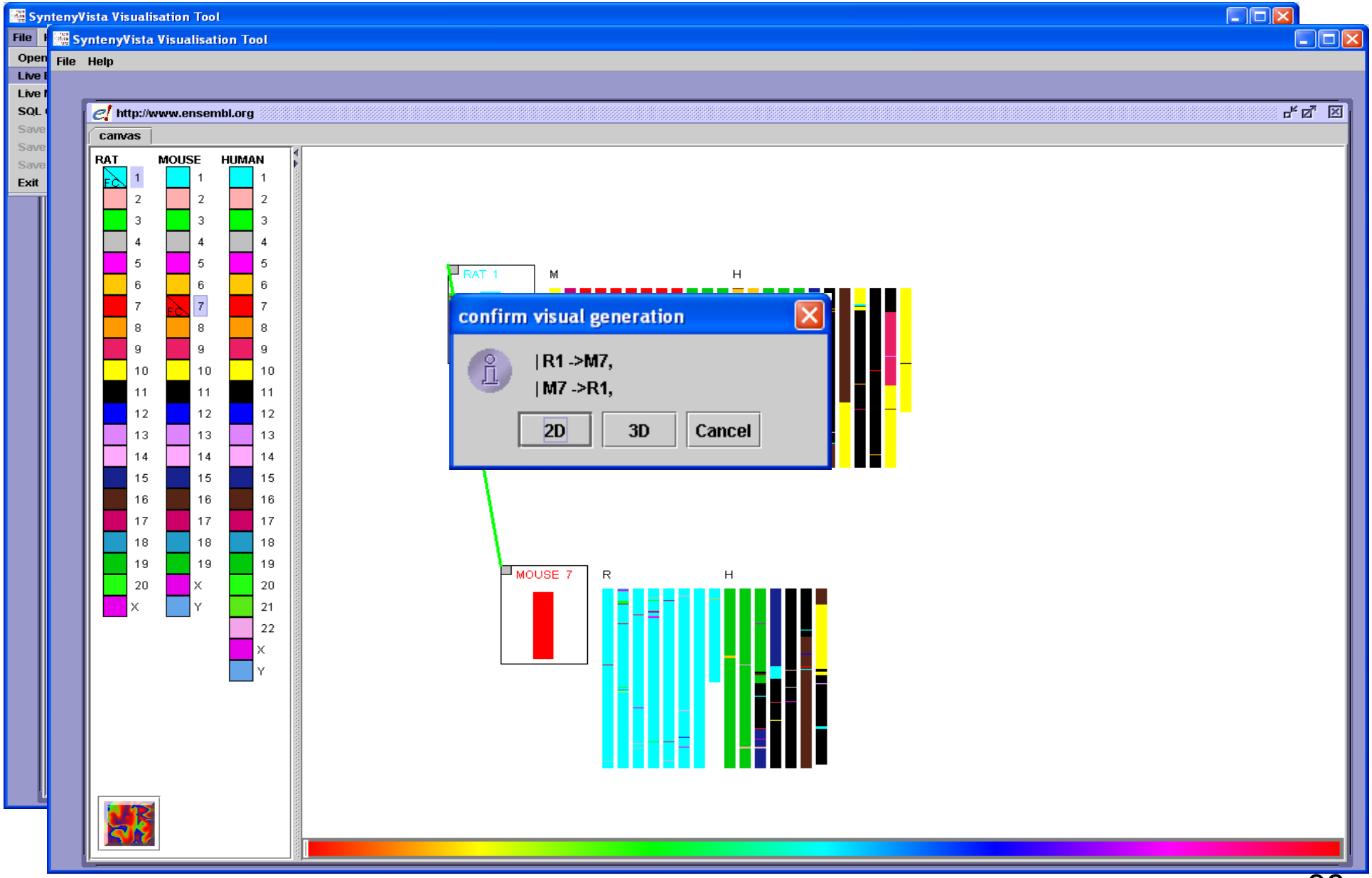
MagnaVista

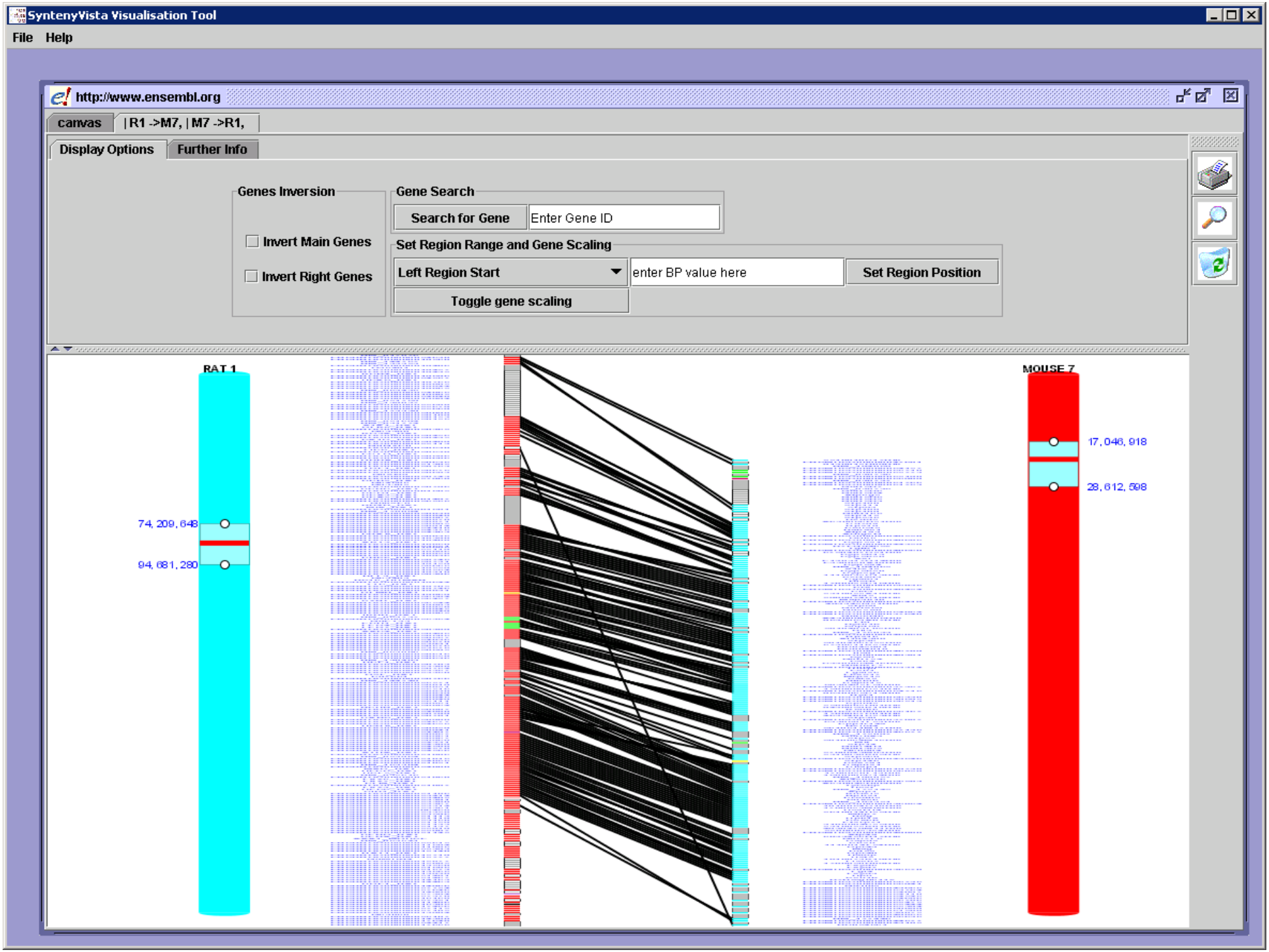

MagnaVista 1.0 Alpha :: rosinnott
File
Help
help on my current view
Gene Perspective
Field Perspective
Database Perspective
swissprot
omim
ensembl
mgi
rgd
hugo
text
table
trees
single gene
blast jobs
gene analysis
microarray
QTL
Gene Lookup
filter options
gene name
internal gene identifier
pax7
search
results
gene
open
cancel
save
print
search repositories for gene
my profile
history
diagnostic
exit

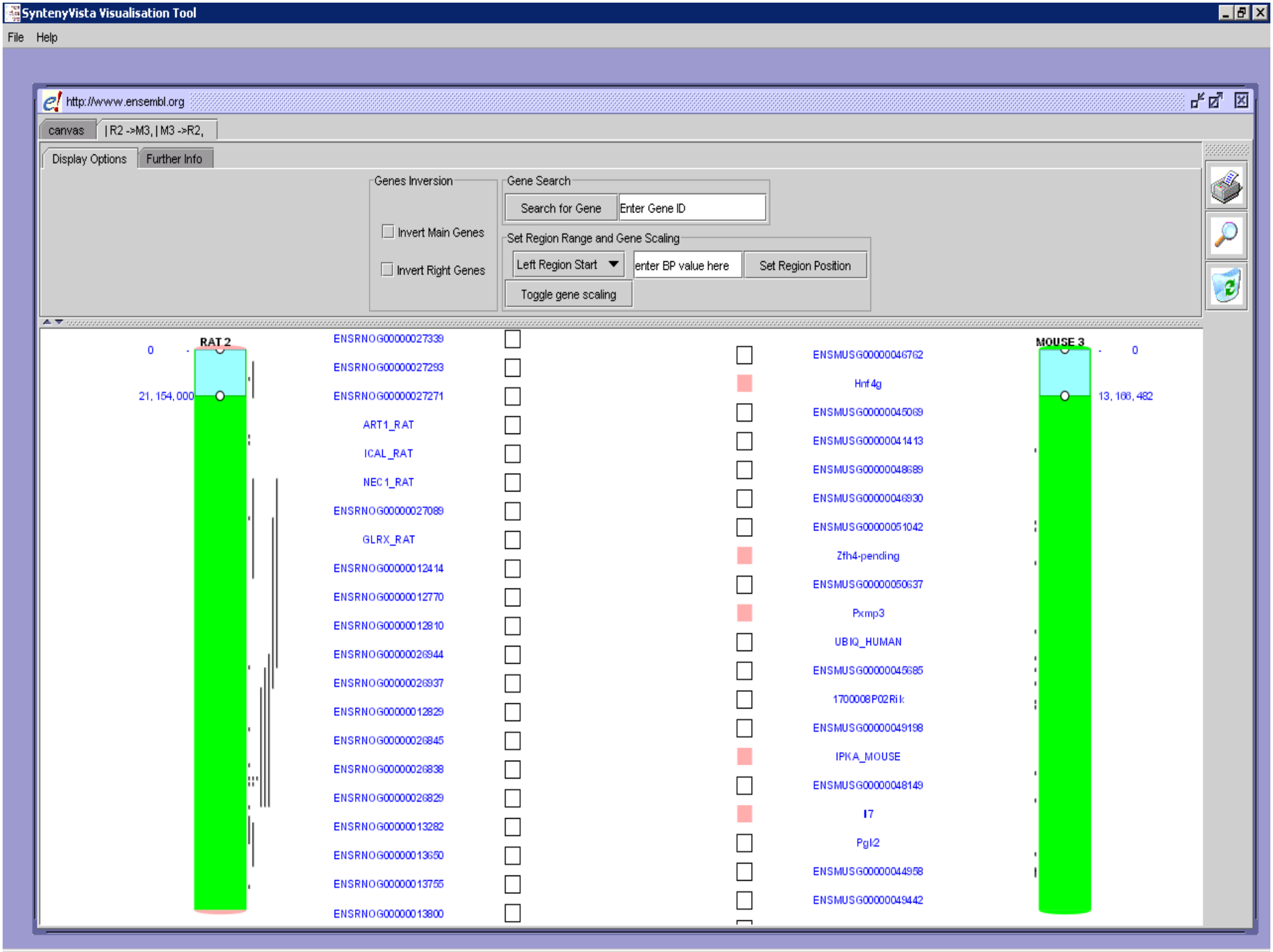
MagnaVista


MagnaVista 1.0 Alpha :: rosinnott
File
Help
help on my current view
Gene Perspective
Pax7
ENSG00000009709 (ensembl)
8621 (hugo)
12185 (mgi)
167410 (omim)
P23759 (swissprot)
P47239 (swissprot)
Field Perspective
Pax7
Database Perspective
swissprot
omim
ensembl
mgi
rgd
hugo
text
table
my profile
history
diagnostic
Profile for user:rosinnott
My Profile
Restore System Defaults
fields found for database: ensembl
Field
species
gene name
GENE_ID
GENE_STABLE_ID
TYPE
DISPLAY_XREF_ID
GENE_CHROM_START
GENE_CHROM_END
CHROM_STRAND
CHROMOSOME_ID
CHR_NAME
DESCRIPTION
BAND
KNOWN_GENE
DISPLAY_ID
DB_NAME
TRANSCRIPT_COUNT
EMBL
GO
HUGO
LOCUSLINK
MIM
PDB
REFSEQ
SWISSPROT
PROTEIN_ID
MMUSCULUS_HOMOLOG
RNORVEGICUS_HOMOLOG
swissprot
omim
ensembl
mgi
rgd
hugo
genes
databases
perspectives
field cross references
fields
Probe Sets
qtl
preferences
ok
print
exit

Importance of Data Visualisation




SyntenyVista Visualisation Tool
File Help
http://www.ensembl.org
canvas
| R1 ->M7, | M7 ->R1,
Display Options
Further Info
Genes Inversion
Invert Main Genes
Invert Right Genes
Gene Search
Search for Gene
Enter Gene ID
Set Region Range and Gene Scaling
Left Region Start
enter BP value here
Set Region Position
Toggle gene scaling
RAT 1
74, 209, 648
94, 681, 280
MOUSE 7
17, 046, 918
28, 612, 598


SyntenyVista Visualisation Tool
File
Help
http://www.ensembl.org
canvas
| R2 ->M3, | M3 ->R2,
Display Options
Further Info
Genes Inversion
Invert Main Genes
Invert Right Genes
Gene Search
Search for Gene
Enter Gene ID
Set Region Range and Gene Scaling
Left Region Start
enter BP value here
Set Region Position
Toggle gene scaling
RAT 2
0
21, 154, 000
ENSRNOG00000027339
ENSRNOG00000027293
ENSRNOG00000027271
ART1_RAT
ICAL_RAT
NEC1_RAT
ENSRNOG00000027089
GLRX_RAT
ENSRNOG00000012414
ENSRNOG00000012770
ENSRNOG00000012810
ENSRNOG00000026944
ENSRNOG00000026937
ENSRNOG00000012829
ENSRNOG00000026845
ENSRNOG00000026838
ENSRNOG00000026829
ENSRNOG00000013282
ENSRNOG00000013650
ENSRNOG00000013755
ENSRNOG00000013800
ENSMUSG00000046762
Hnf4g
ENSMUSG00000045069
ENSMUSG00000041413
ENSMUSG00000048689
ENSMUSG00000046930
ENSMUSG00000051042
Zfh4-pending
ENSMUSG00000050637
Pxmp3
UBIQ_HUMAN
ENSMUSG00000045685
1700008P02Rik
ENSMUSG00000049198
IPKA_MOUSE
ENSMUSG00000048149
I7
Pgl2
ENSMUSG00000044958
ENSMUSG00000049442
MOUSE 3
0
13, 168, 482

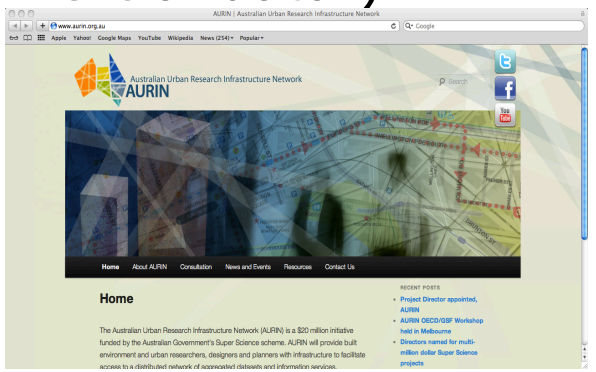
BREAK

Social Sciences

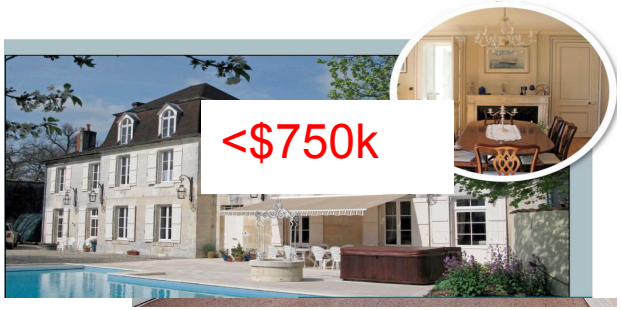
- Data, data everywhere
  - Researchers exploring all kinds of areas with societal impact
  - Many major resources distributed around the globe
    - Data Archives 档案
      - literally thousands of surveys/studies crossing society
      - Most countries have their own similar archives
    - Office of National Statistics, Australian Bureau of Statistics, … and the Census
      - ONS : 1971/81/91/01/11/…
      - ABS : …/01/06/11/16…
    - Geospatial information resources
      - Country profiles, regional profiles, city profiles, street

Australian Urban Research Infrastructure Network (AURIN)

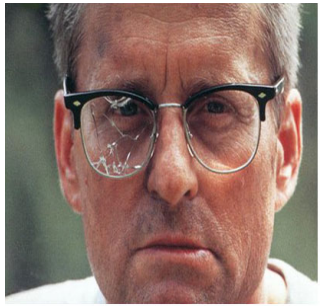
- EIF/NCRIS federally funded project
  - DIISRTE -> Innovation -> Education
  - $35m+ project
  - www.aurin.org.au
    - University of Melbourne are lead agent
- Establishing an e-Infrastructure for Urban and Built Environment Researchers
  - Distributed, (completely!) heterogeneous datasets
  - Data interrogation services
  - Security (unit level data, health data, commercial data!)
  - Online analysis tools
  - Collaboration!!!



AURIN Context



- Urban and build environment is extremely broad
  - health,
  - transport,
  - future population,
  - liveability,
  - crime,
  - housing,
  - design,
  - …

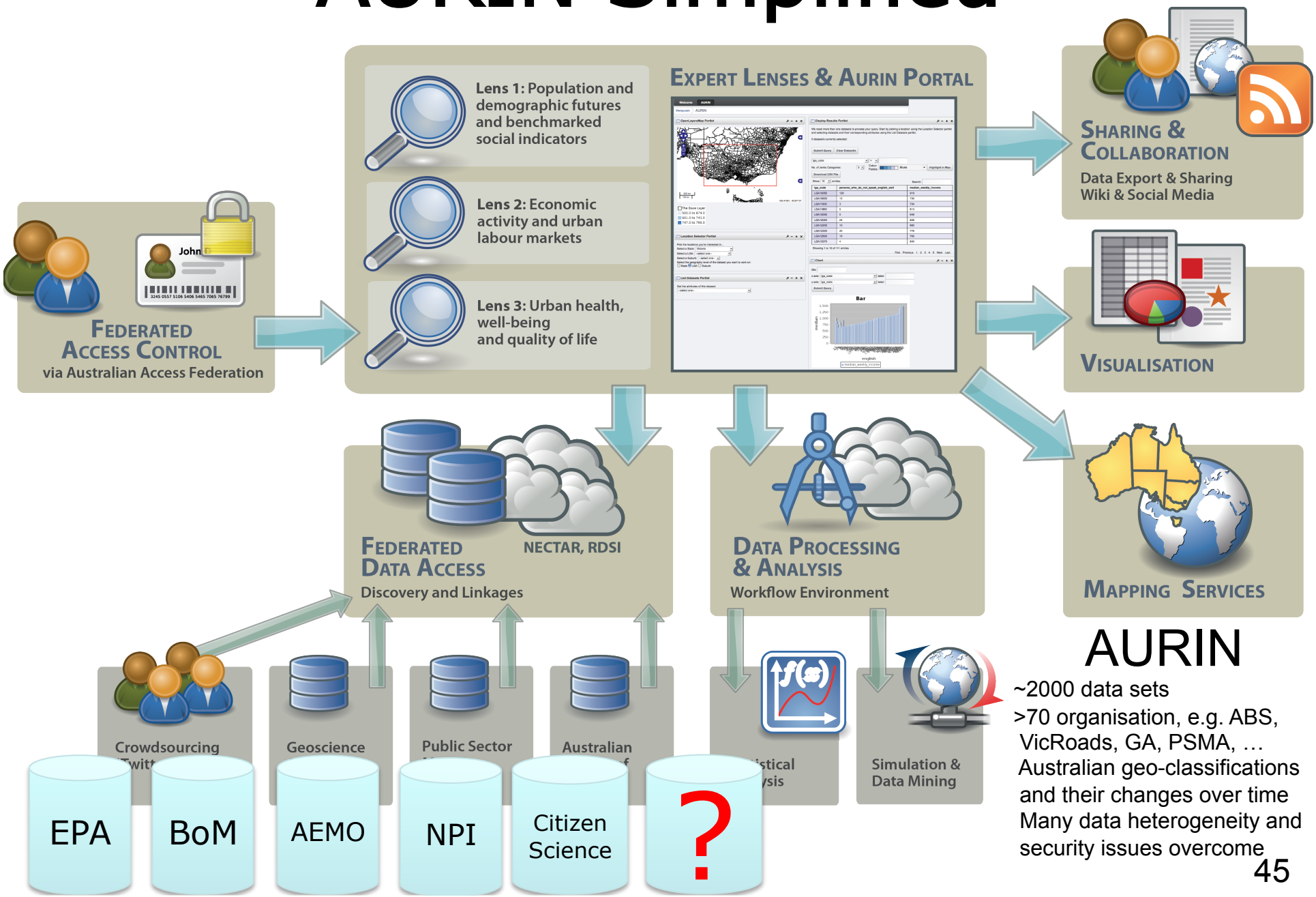


- Much research depends on access t
  - There is LOTS (and LOTS) of data of
  - Completely heterogeneous, e.g. geo
    …
    - Data is more often than not silo'd

- Requires tools to find, interrogate, analyze and visualize data and enforce good research methodologies
  - Consolidate tools and best practice/community know-how!
  - Allow researchers to share results, interact and collaborate
    - No single expert!
- Allow data providers to keep control of their data and its use
  - Authentication and authorisation (and auditing/accounting)

巩固

AURIN Simplified



AURIN

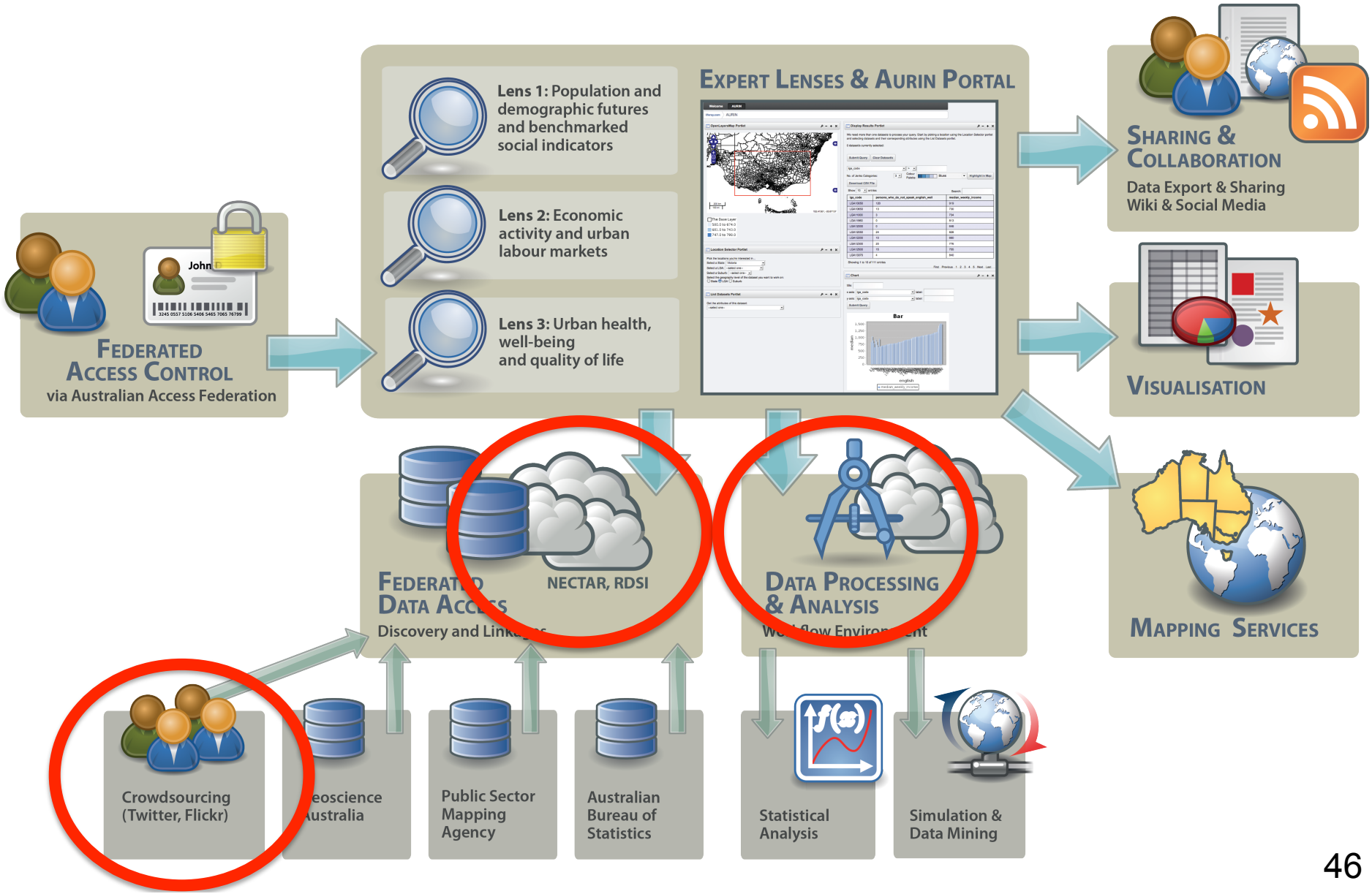
~2000 data sets

>70 organisation, e.g. ABS, VicRoads, GA, PSMA, …

Australian geo-classifications and their changes over time

Many data heterogeneity and security issues overcome

AURIN Architectural Components



Demonstration

(note – Assignment II)

Clinical Sciences

- Research into:
  - cancer
    - breast, bone, pancreatic, liver…
  - paediatric endocrinology
    - disorders of sex development, adrenal tumours, hormonal imbalances, …
  - brain
    - neurological disorders, brain trauma, …
  - …
- Pretty much every hospital, GP, clinical system I've come across is different

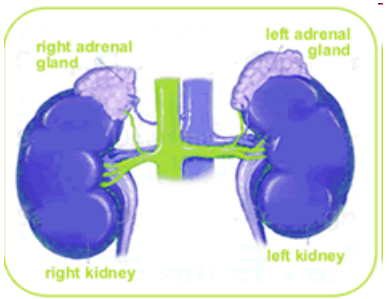
Data Sharing and Ethics

- **Clinical data sharing differences globally**
  - Ethics
    - Committees / bodies
    - Structures and processes
  - Consent
    - Opt-in
    - Opt-out
  - Data sharing
    - Local only
    - National only
    - Project collaborators only
    - All international researchers
- **Need for completely discretionary data sharing models**
  - Driven by patient/parental consent
  - Ethics, oversight, …

Designed with worst case scenarios in mind!

ENSAT-CANCER

- European Network for Study of Adrenal Tumours (ENSAT – www.ensat.org) CANCER
  - 6m Euro; 5-year project;
    - Started 2011
    - Funds one techie (@UniMelb)
  - Focus on 4 types of tumour
    - Adrenocortical carcinoma (ACC)
    - Aldosterone Producing Adenoma (APA)
    - Non-aldosterone cortical adrenal adenomas (NAPACA)
    - Pheochromocytomas and related paragangliomas (Pheo)
      - These are rare (ACC ~1-2 cases per million pop.)
  - e-'ing
    - Secure web-based tumour databases
    - Biobanking
    - Imaging
    - Clinical trials
    - Seamless interactions between all of these



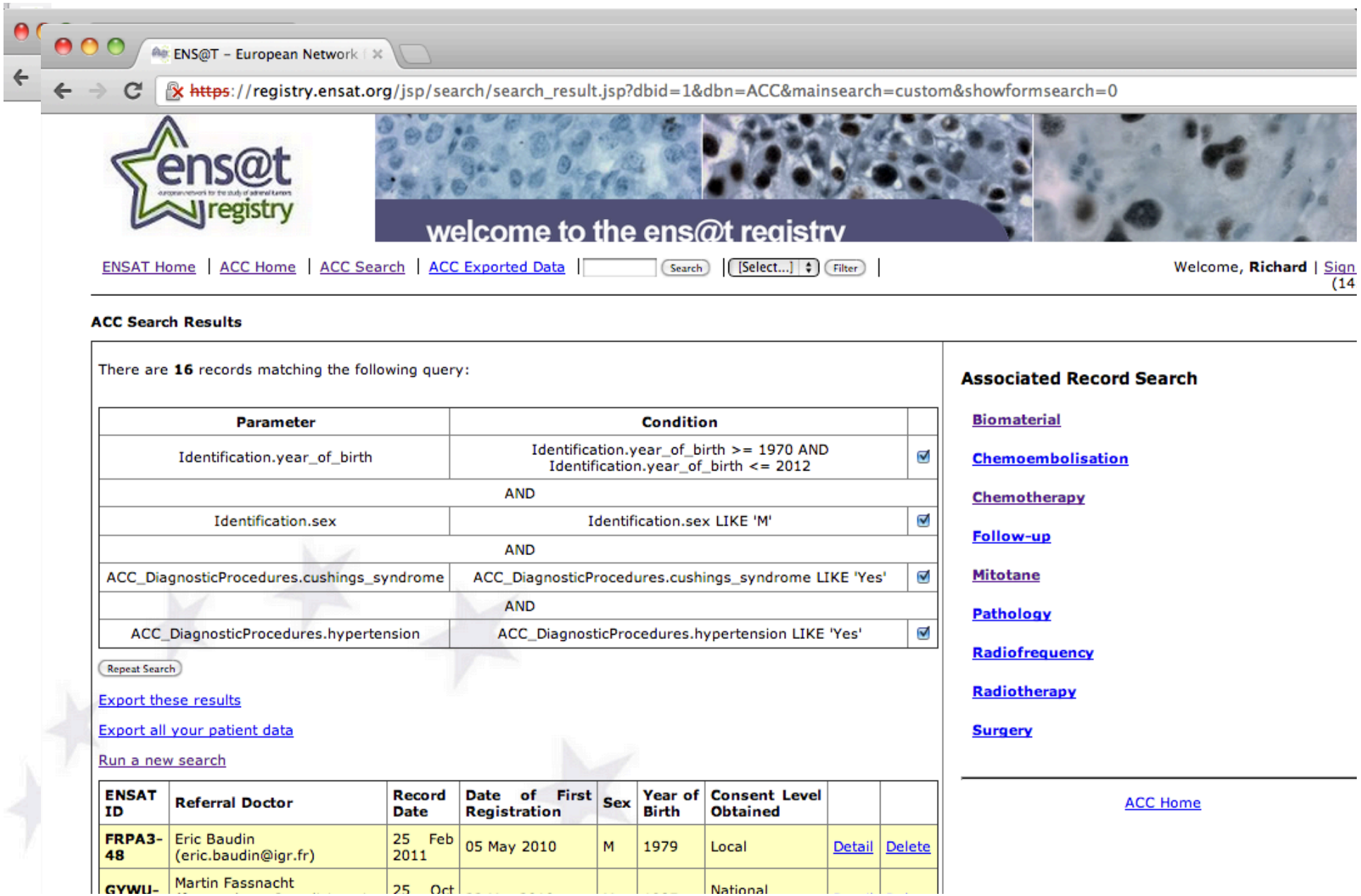
Associated Study and Registry Distribution

	ACC	Pheo	NAPACA	APA	Total	Principal Investigator	Study Protocols	Study sites/eCRFs
EURINE-ACT	515	7	1678	229	2429	Wiebke Arlt	EURINE-ACT	
Ki-67	36	0	1	0	37	Martin Fassnacht		
Stage III/IV ACC	35	0	0	0	35	Eric Baudin		
PMT	0	229	133	2	364	Graeme Eisenhofer	PMT	PMT
TMA	3	188	0	0	191	Ronald de Krijger		
Long-term PHPGL	0	242	0	0	242	Pierre-Francois Plouin		
AVIS-2	0	0	0	2	2			
PMT3	0	12	0	0	12	Graeme Eisenhofer		
ADIUVO	46	0	0	0	46	Massimo Terzolo		ADIUVO
ADIUVO Observational	61	0	0	0	61	Massimo Terzolo		
HairCo-2	9	0	1	0	10	Marcus Quinkler		
FIRST-MAPPP	0	41	0	0	41	Eric Baudin		FIRST-MAPPP
German Cushing Registry	0	1	62	0	63	Martin Reincke		
German Conn Registry	0	0	0	609	609	Martin Reincke		German Conn Registry
CHIRACIC	0	0	3	0	3			
ACC Molecular Marker	35	0	0	0	35			
UK Pheo Audit	0	669	0	0	669	Srirangalingam Umasuthan		
Lysosafe	318	0	0	0	318	Felix Beuschlein		Lysosafe
FIRMACT	5	0	0	0	5	Martin Fassnacht	FIRMACT	FIRMACT
MAPP-Prono	0	127	0	0	127	Eric Baudin	MAPP-Prono	
MIBG Impact	0	115	0	0	115	Henri Timmers		
Predict Ancillary FIRM-ACT	2	0	0	0	2	Eric Baudin		
FAMIAN	0	0	3	0	3	Stefanie Hahner		FAMIAN

Summary

	Total
Records	10726
Biosample	22059
Clinical Ar	72539
Annotatic	6.76
Biosample	2.05
Active Ce	91

ENSAT-CANCER



Prospective Monoamine-Producing Tumor (PMT) study

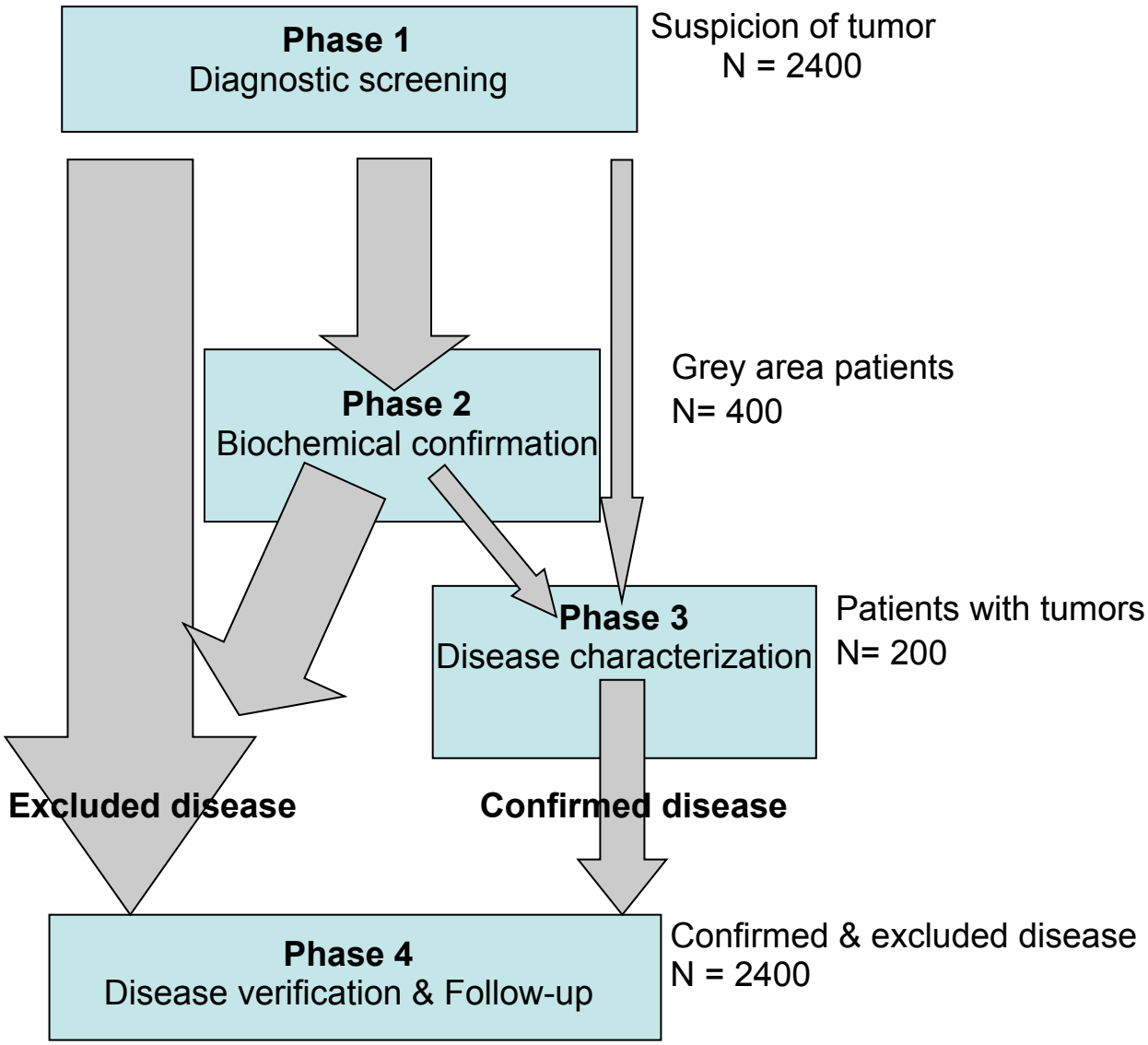
Aims:

生物标记

- Diagnostic biomarkers - miRNA & utility of plasma and urinary free metanephrines
- Follow-up testing - effective strategy for follow-up testing
- Genotype-phenotype relationships
- Characterize catecholamine meta-bolomic & secretory phenotypes in relation to cardiovascular and other manifestations of disease
- LC-MS/MS based metabolomic profiling of bioenergetic signatures
- Novel biomarkers of malignancy

Contact for patient enrolment:

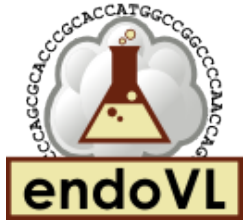
Dr. Roland Därr
Roland.Daerr@uniklinikum-dresden.de



Endocrine Genomics VL

- Recently funded NeCTAR VL project
  - $1.03m (ran for 1 year!)
  - Platform for range of endocrine related disorders with associated -omics analytics
  - Includes
    - Type-1 & Type-2 Diabetes (paediatric and adult) and rare forms of diabetes related disorders
      - Working with APEG, JDRF & ADS
    - Thyroid-related and severe obesity-related disorders
      - Working with Australian Thyroid Foundation
    - Neuroendocrine and adrenal disorders
      - Supporting clinical oncology society of Australia
    - Bone related disorders
      - Supporting Australia and New Zealand Bone and Mineral Society and Osteoporosis Australia
    - Disorders of Sex Development
      - Supporting the Australasian DSD Network

Endocrine Genomics VL



- Beyond ENSAT (registry/trials)
  - Integration of workflows for –omics data analysis
    - Whole genome/exome sequencing
    - Sharing how analytics is undertaken (repeatable data analysis pipelines)
    - Leveraging Australia-wide expertise in this space
      - Life Science Computation Centre (Victoria)
      - Australia Genetics Research Foundation
      - Bioplatforms Australia
      - National ICT Australia
  - Multiple –omics groups offering multiple omics analysis
  - Comparison of resulting analysed data in integrated system
  - Use of Cloud for data and analytics
    - …and overcoming security challenges this gives rise to!

vlsci
LIFE SCIENCES COMPUTATION

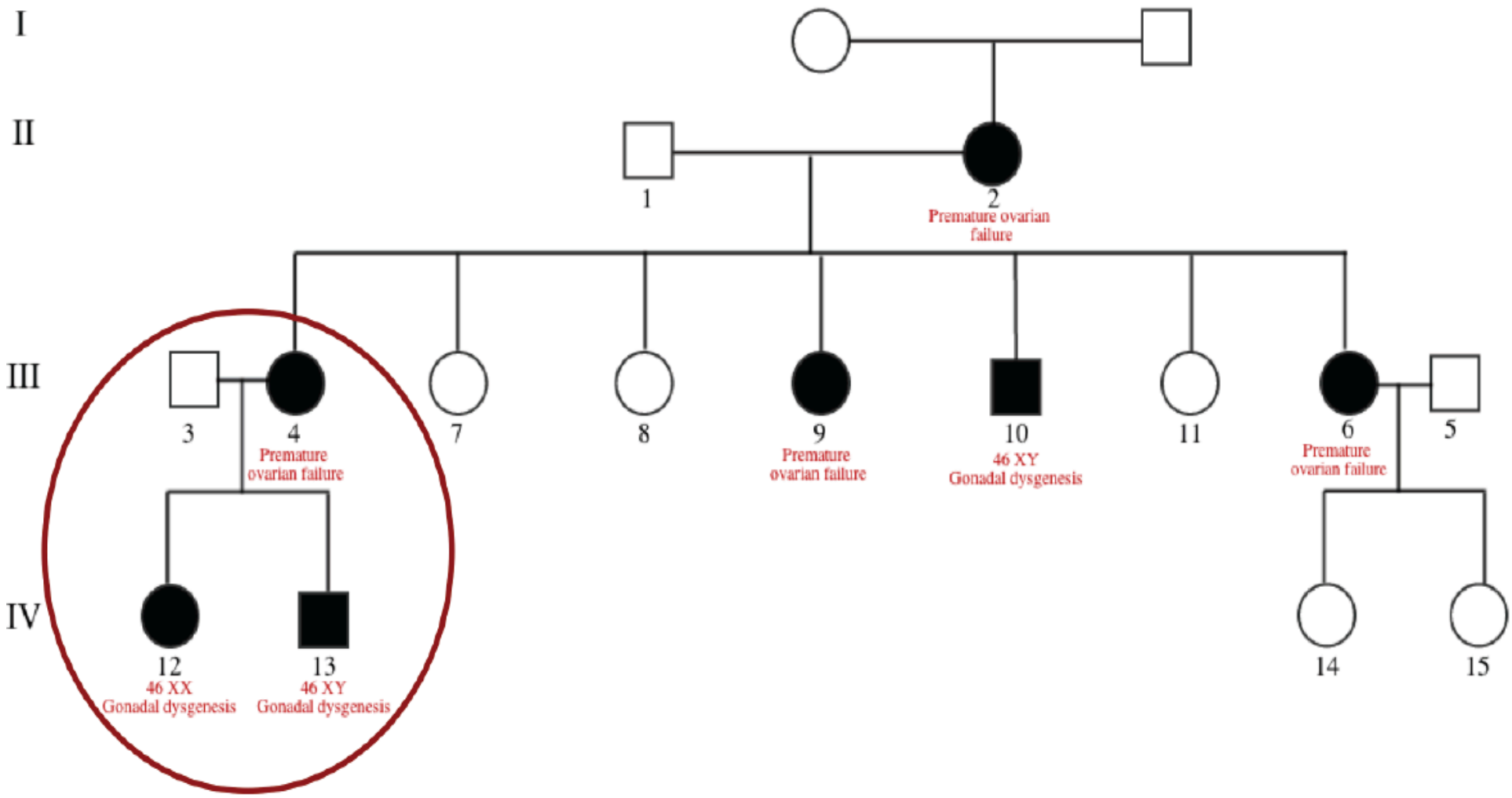
agrf
Australian Genome Research Facility Ltd

CENTRE FOR COMPARATIVE GENOMICS
Western Australia

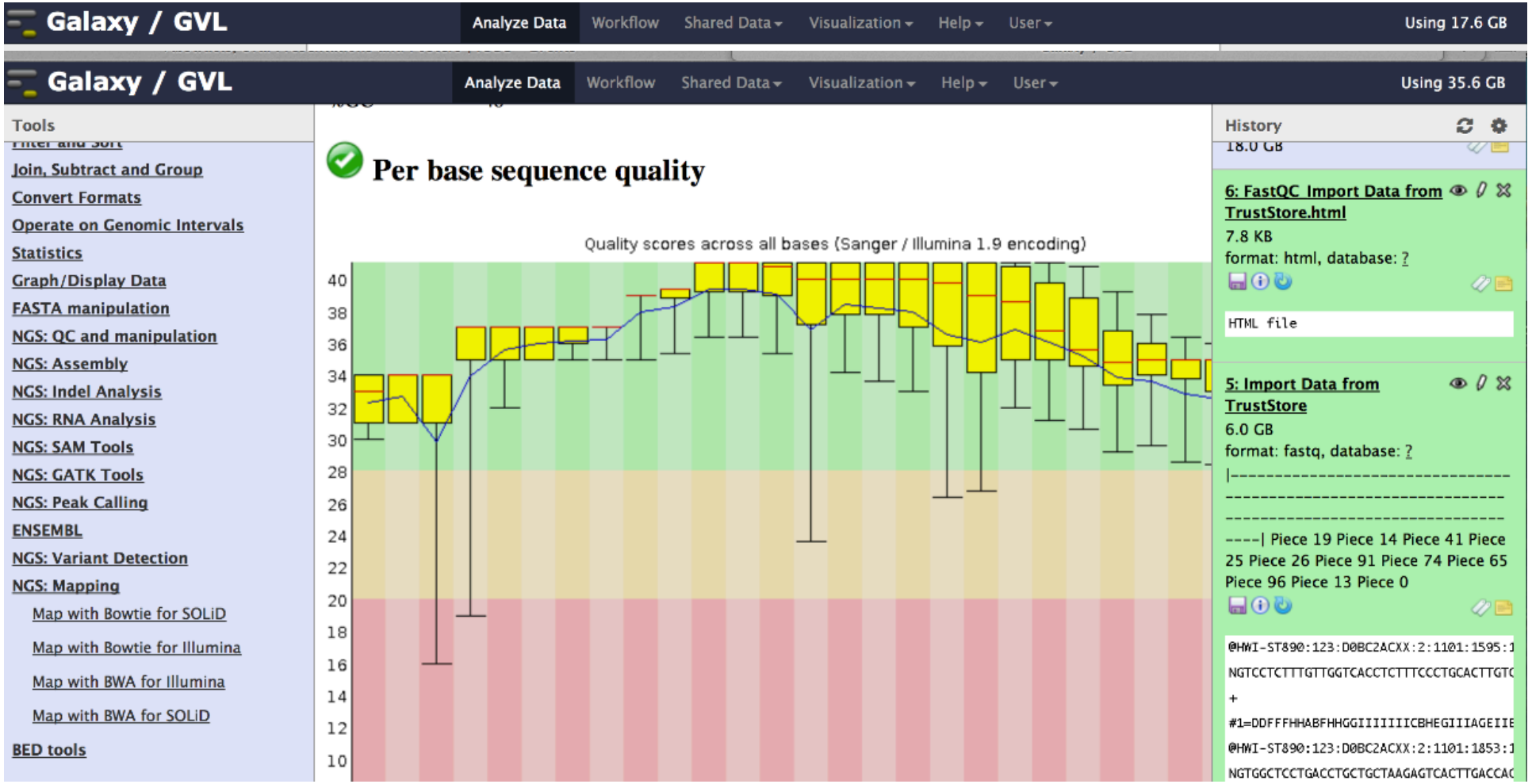
NICTA

CSIRO

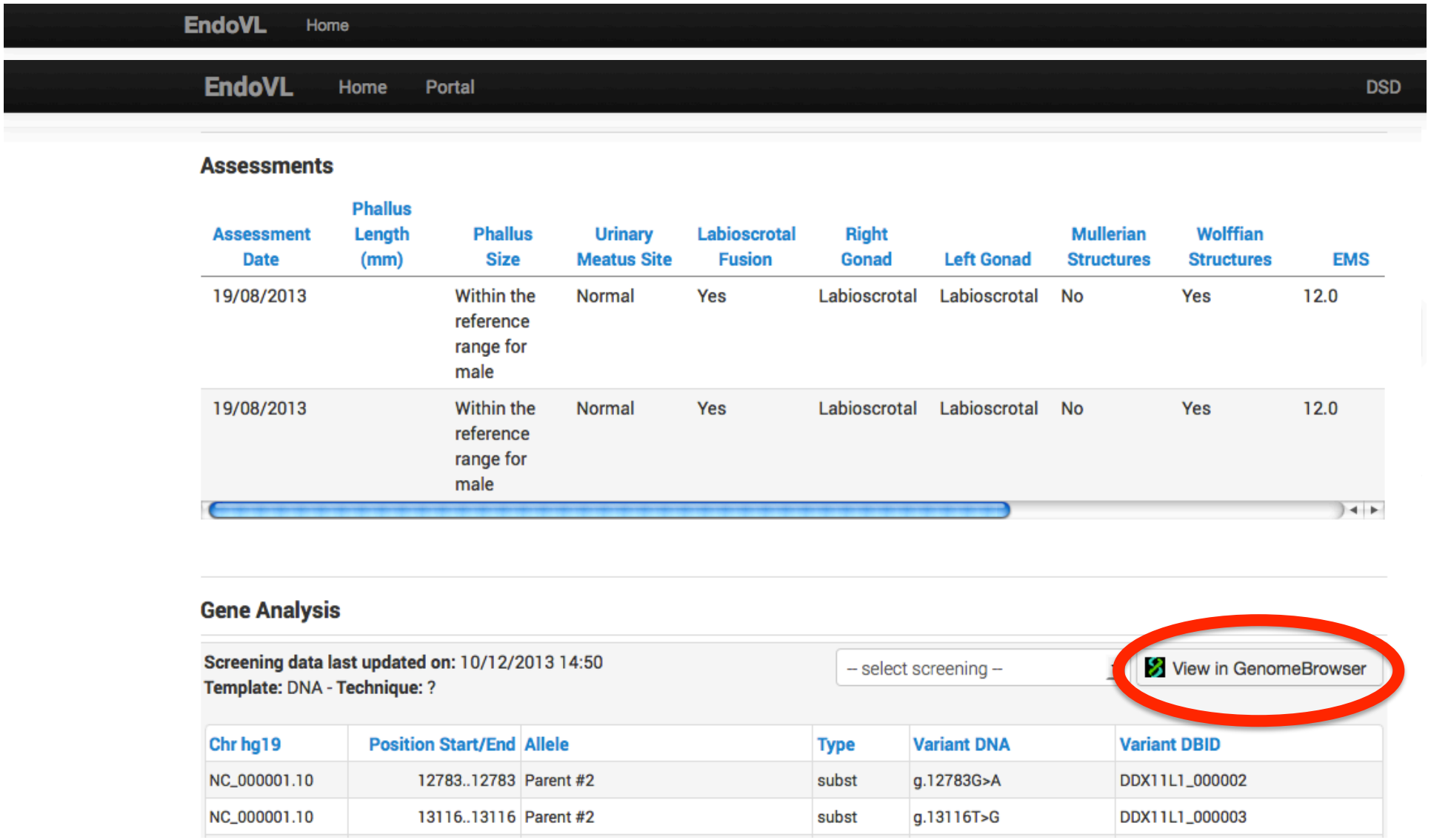
Case Study (Blind) DSD –omics analysis



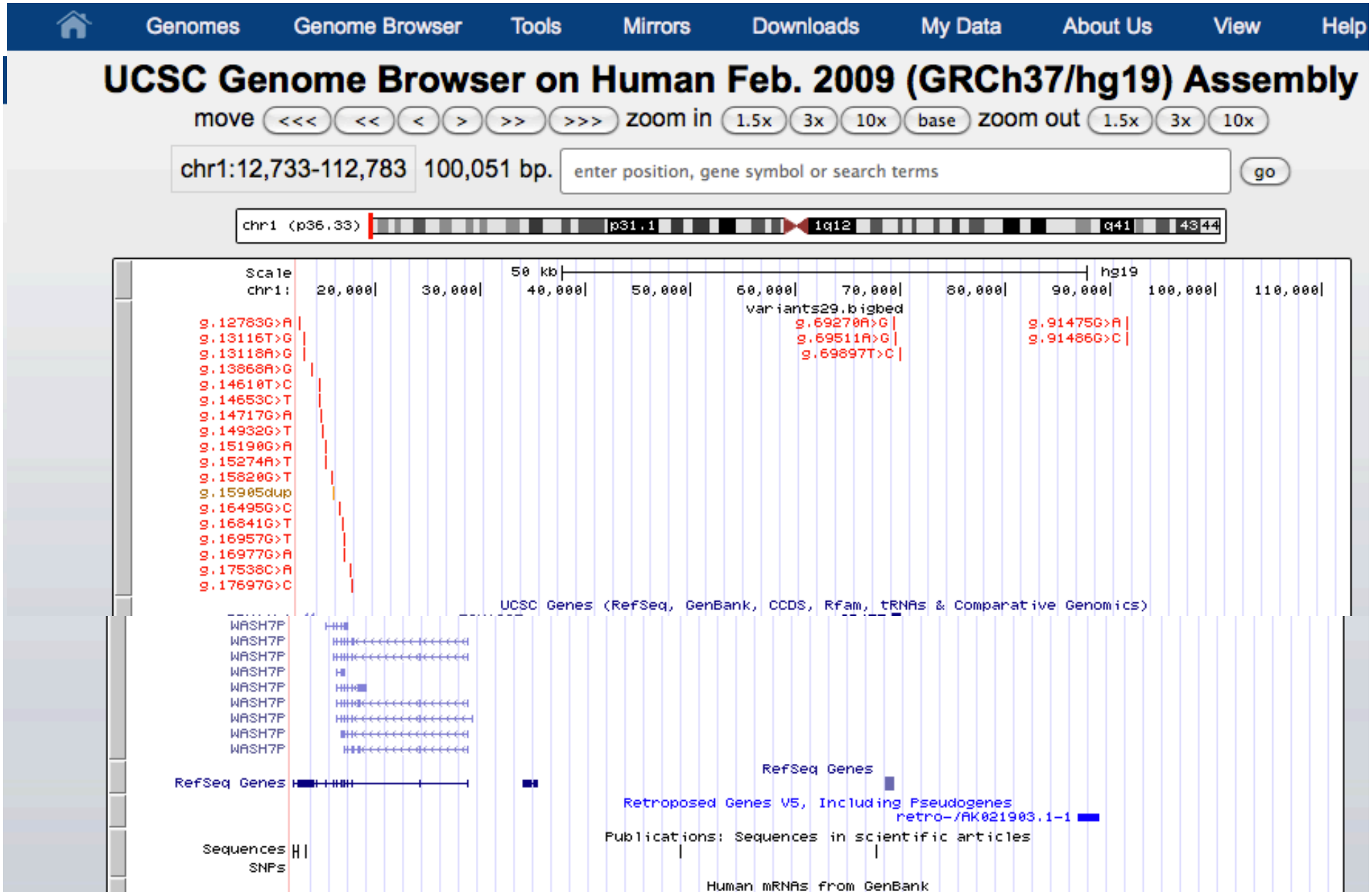
Cloud-based -omics Analytics Pipelines



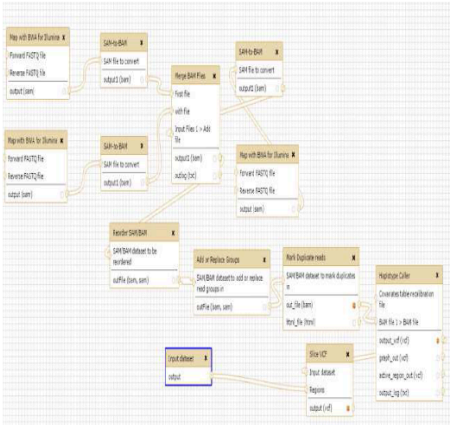
Integration of Phenotype & Genotype

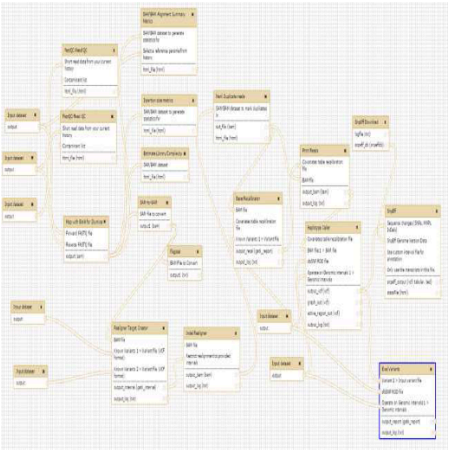


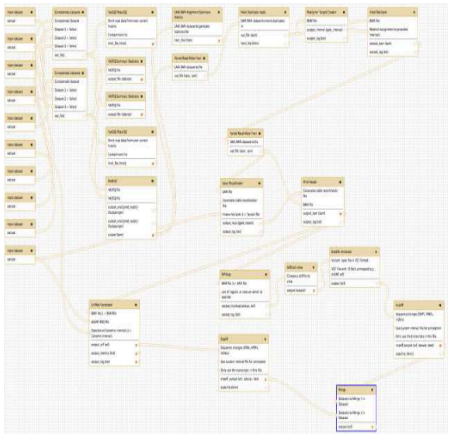
WGS Visualisation

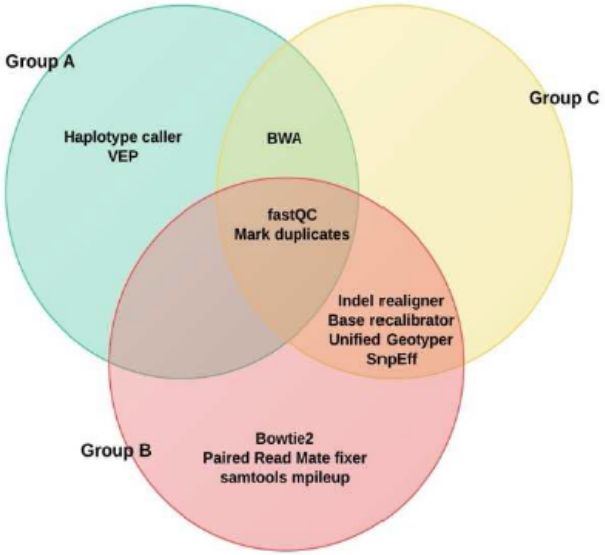


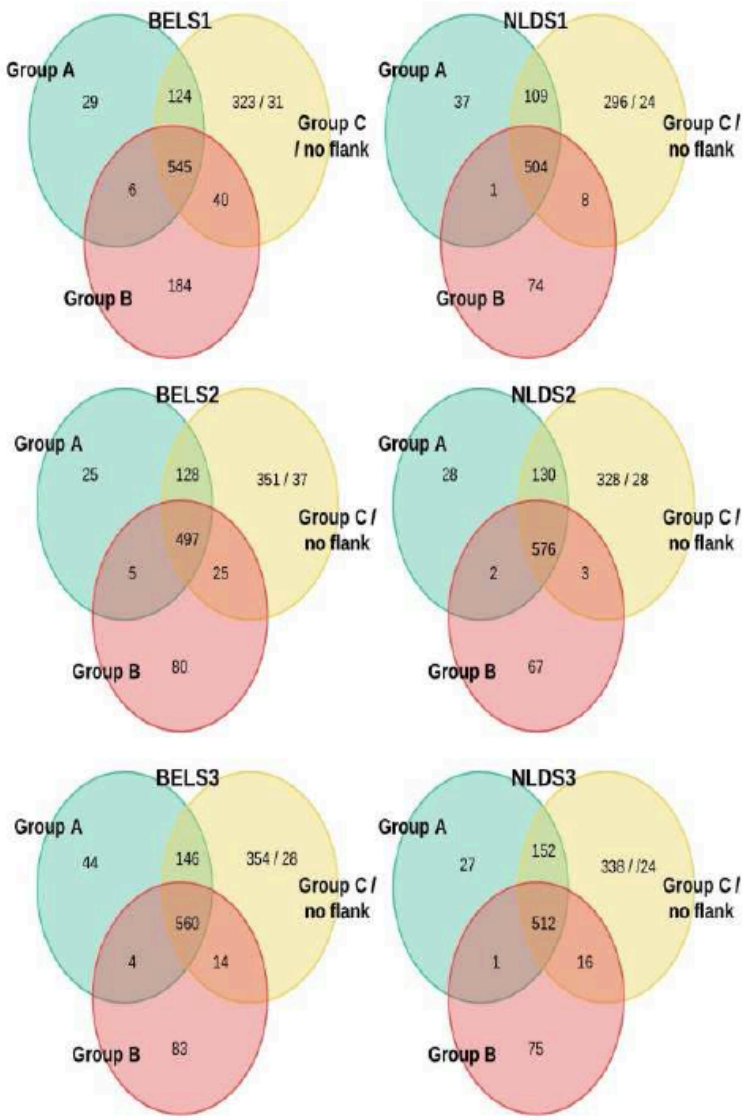
And the winner is…











Conclusions

- All application domains facing data “issues”
- How do we build e-Infrastructures that support this?
- Specialised research specific Google-farms?
  - Yes but life sciences
    - EBI,
    - SANGER,
    - NCBI,
    - …
  - The diversity
    - researchers interested in rat, mouse, human, barley, jellyfish, cancer, diabetes, Tasmanian devil, …
    - Sonic Hedgehog, ARSE, 18wheeler, pray for elves, pokemon, werewolf, BRCA1, BRCA2, …

Conclusions…ctd

- Importance of community driven standards
  - But forever a moving target
    - Ontologies
    - Communities often not aligned/mature
    - …
- The toothbrush culture!!!
  - Big Pharma/Grants…
- Security!!!
  - User-driven
  - Provider-driven
  - Virtual Organisations
    - Simple enough?
- Technology is still often specialised/hard
  - But it can be achieved …if you know how! ;o)

Questions ...?