Domain Drivers – Lecture 3-4

Professor Richard O. Sinnott
Director, Melbourne eResearch Group
University of Melbourne
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!!!

Focal Point

- Examples from different research domains
  - {Computational/Data/Distributed/Collaboration/Security} bound…
    - High Energy Physics
    - Astrophysics
    - Electronics
    - Bioinformatics
    - Biomedical and bioinformatics domain

- BREAK
    - US Intelligence
    - Social Sciences/Urban research domain
      - (prelude to workshop and assignment 2)

Project Portfolio

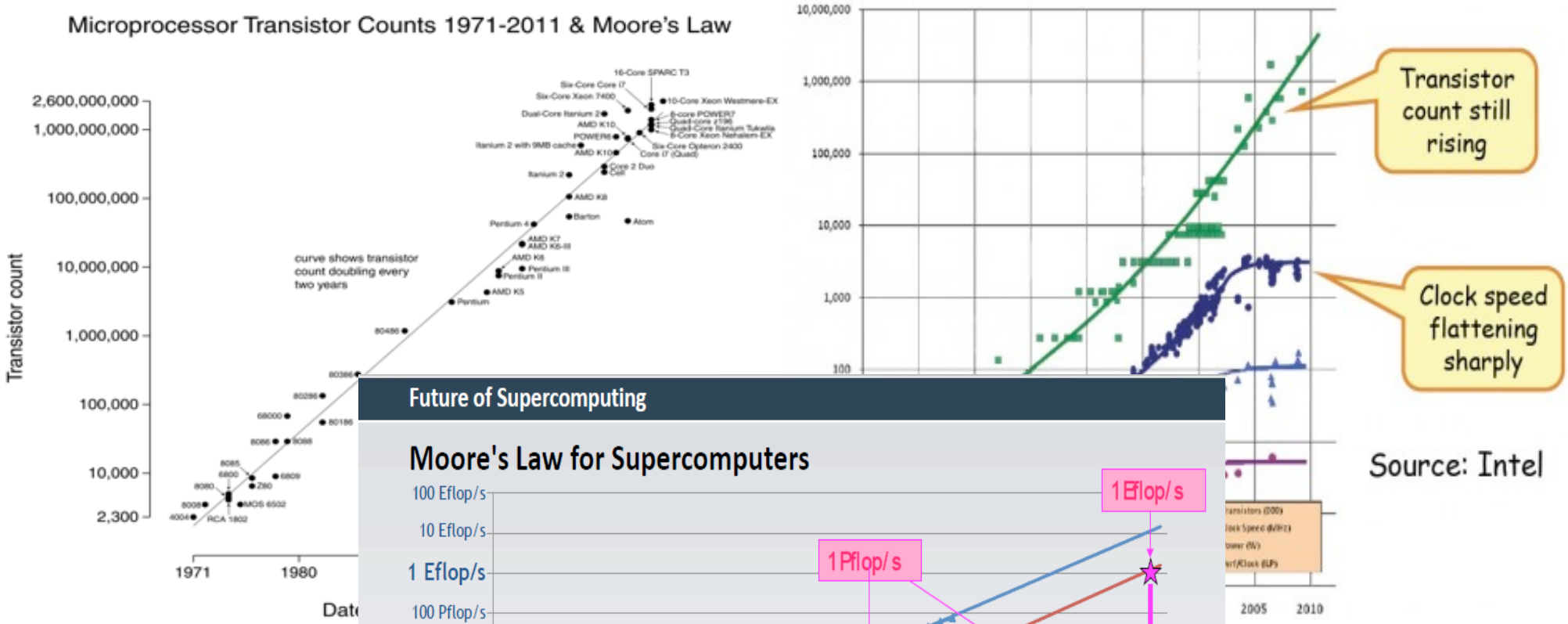
Completed

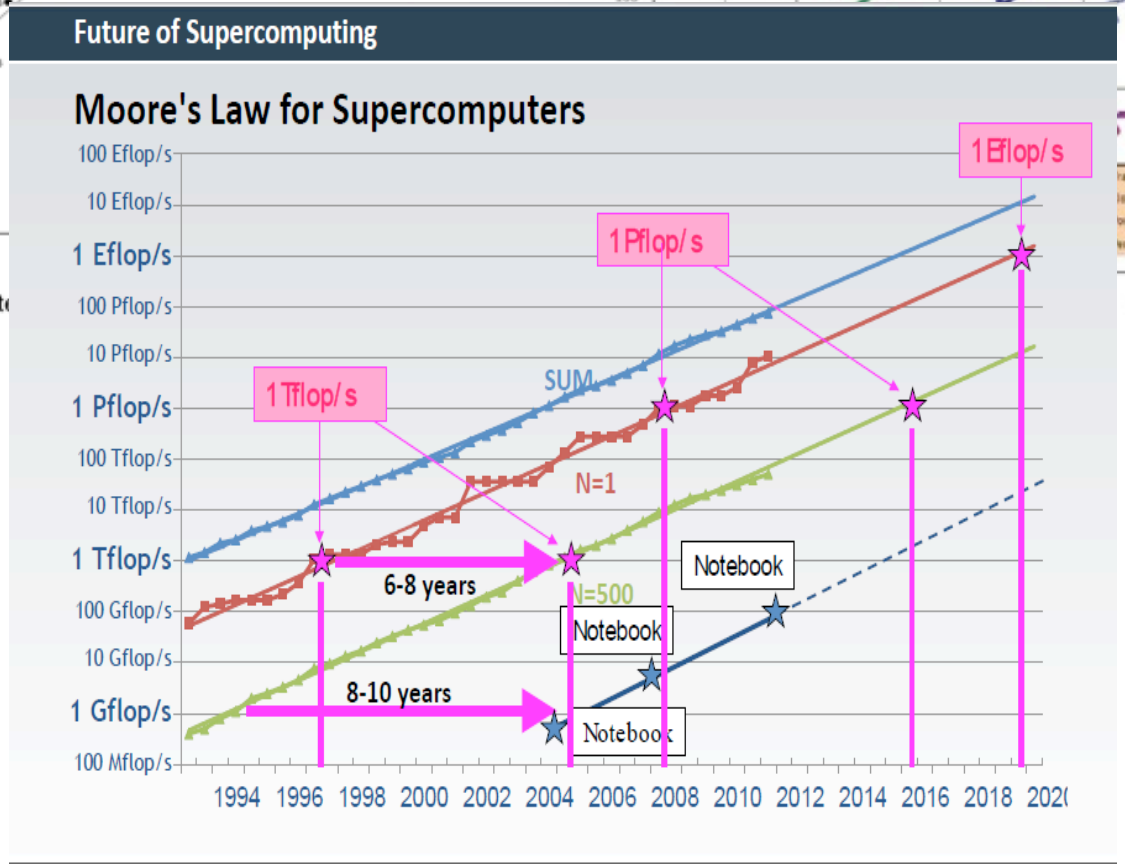
- National e-Science Centre (I, II, III)
- Dynamic Virtual Organisations for e-Science Education
- Biomedical Research Informatics Delivered by Grid Enabled Services
- GridNet, GridNet 2
- Grid Enabled Microarray Expression Profile Search
- Glasgow early adoption of Shibboleth
- Joint Data Standards 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

Subset of On-Going

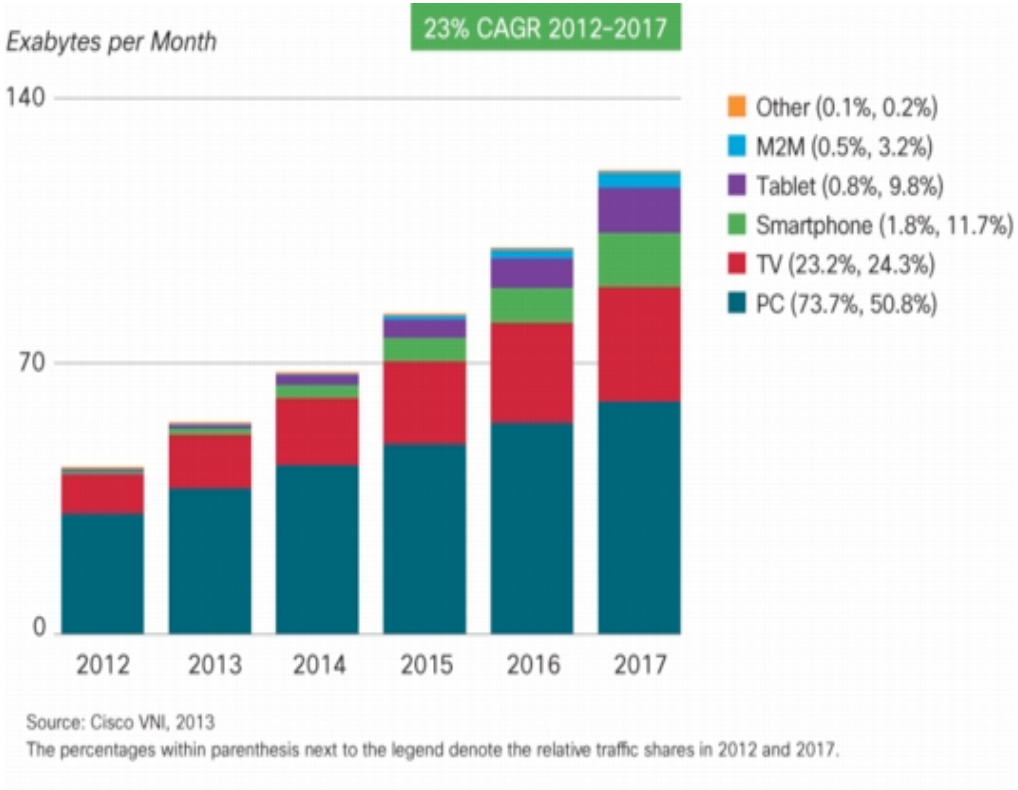
- EU European Platform for Study of Wolfram, Alstrom, Bardet Biedl (EuroWABB)
- Multicenter prospective study of biochemical profiles of monoamine-producing tumors (PMT Study)
- European Society of Hypertension Study on Pheo/PGL
- International DSD
- EU FW7 European Network for Study of Adrenal Tumors 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
- Hort-eye Cloud analytics
- 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
- 88days Backpacker app
- VicSpin Victoria-wide Flu Surveillance System
- ElectraNetLIDAR/VectorNZ Lidar
- Growing Landscape Carbon
- Replicats
- Bushfire data management platform

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

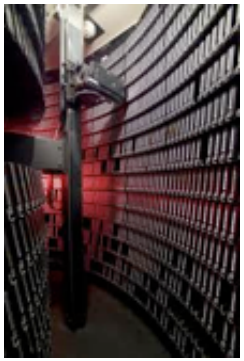
- From tablets, to papyrus, to books
  - (quite adequate for several thousand years)
- Enter silicon transistors circa 1960
  - punch cards,
  - punched streamer tape,
  - magnetic tape,
  - floppies,
  - …
    - ~RIP!

Data Present

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






Data Deluge

- The combi… at approxima…

- The total a… 2012. This is 48…

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

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.

(March 2007)
igital Universe
ta Corporation
h_zettabyte/

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

Online System

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

1 TIPS is approximately 25,000 SpecInt95 equivalents

Offline Processor Farm ~20 TIPS

Tier 0

CERN Computer Centre

Tier 1

France Regional Centre

Germany Regional Centre

Italy Regional Centre

FermiLab ~4 TIPS

ScotGrid

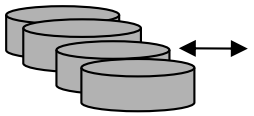
~622 Mbits/sec

Tier 2

Caltech ~1 TIPS

Tier2 Centre ~1 TIPS

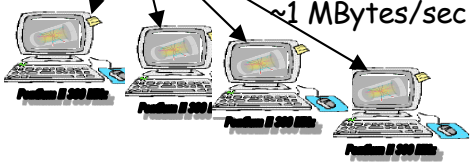
Physicists work on analysis "channels".
Each institute will have ~10 physicists working on one or more channels; data for these channels should be cached by the institute server



Physics data cache

Institute ~0.25TIPS

~1 MBytes/sec



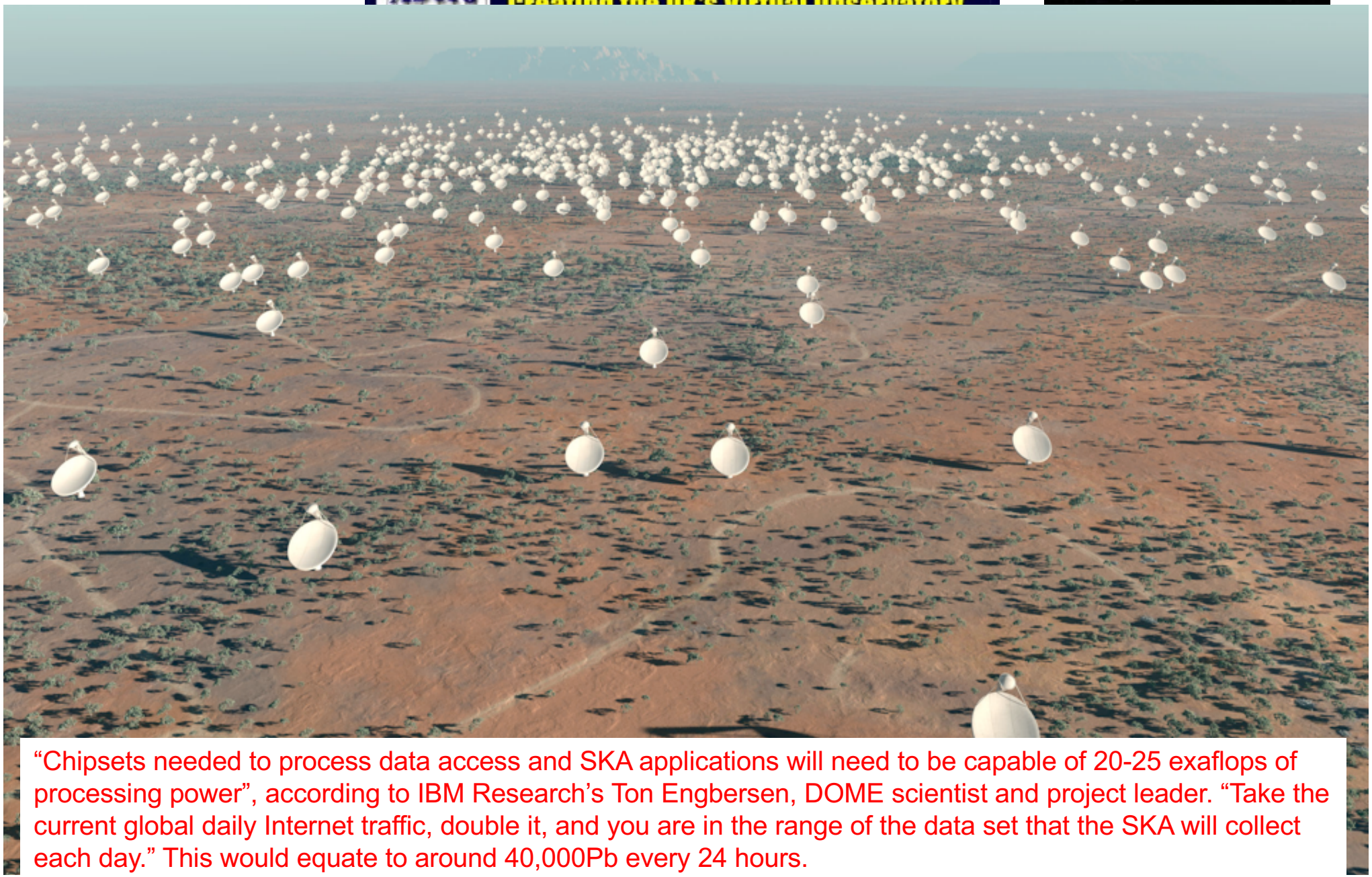
Tier 4

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.

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

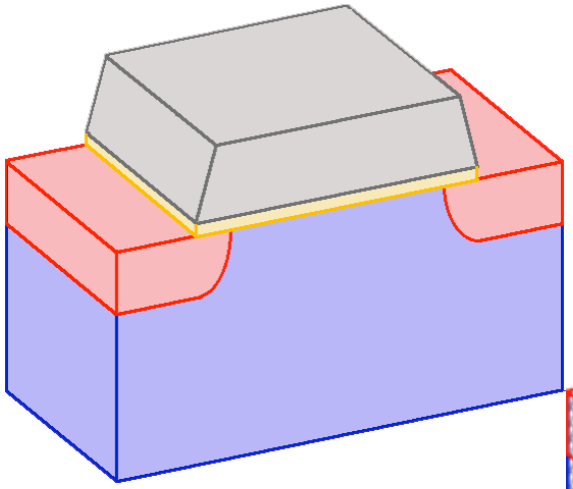
11 PDRAs (7 + 4)
7 PhD

Industrial partners

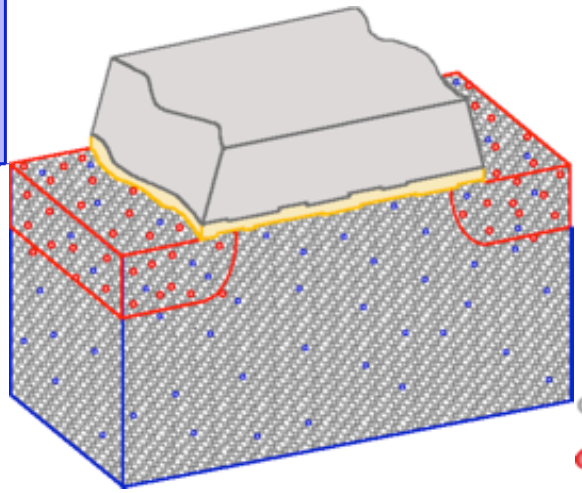
ARM, wolfson microelectronics, synopsys, freescale semiconductor, National Semiconductor, FUJITSU

University partners

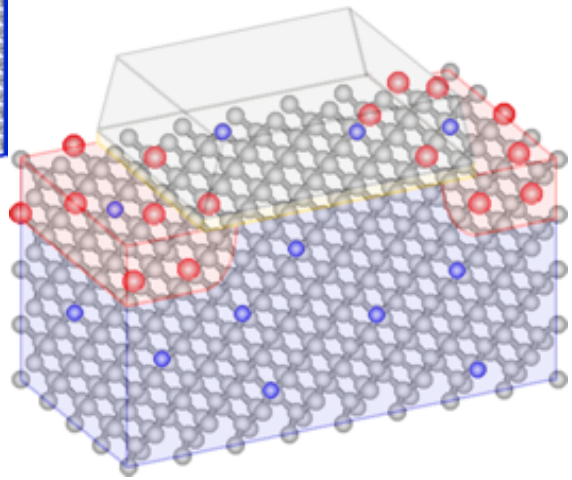
Semiconductor device variability



Historic simulation paradigm

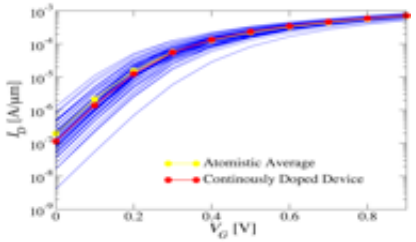


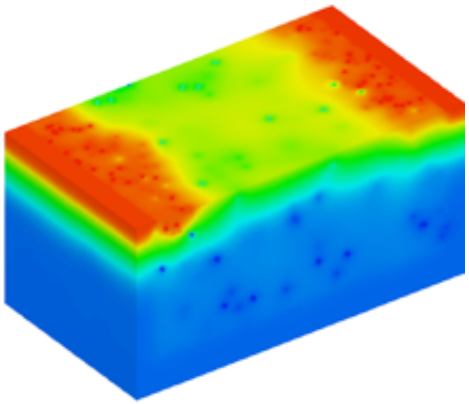
A 22 nm MOSFET
In production 2011

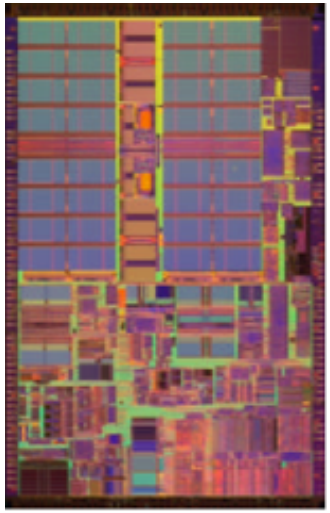


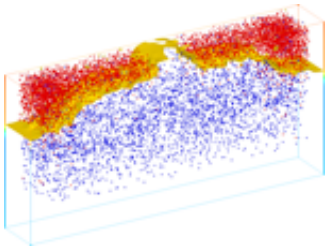
A 4.2 nm MOSFET
In production 2023

Challenges of NanoCMOS Design





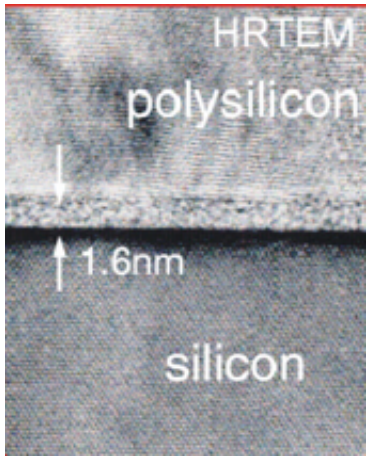


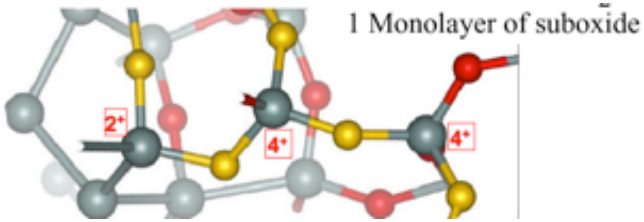


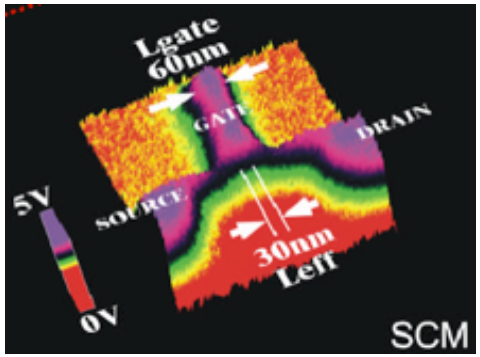
3D

+

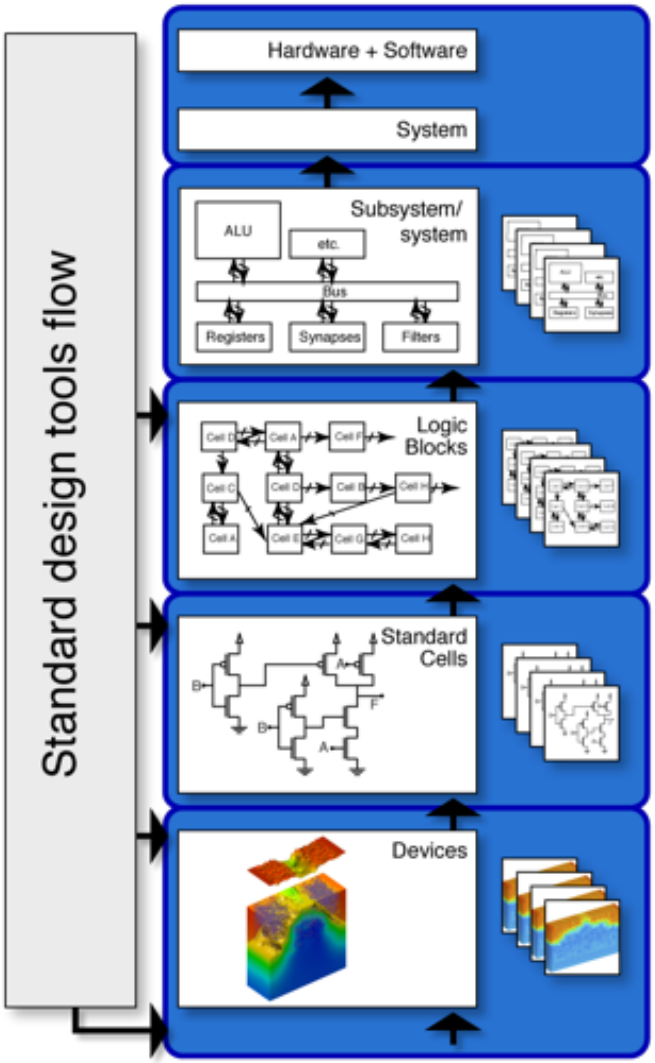
Statistical







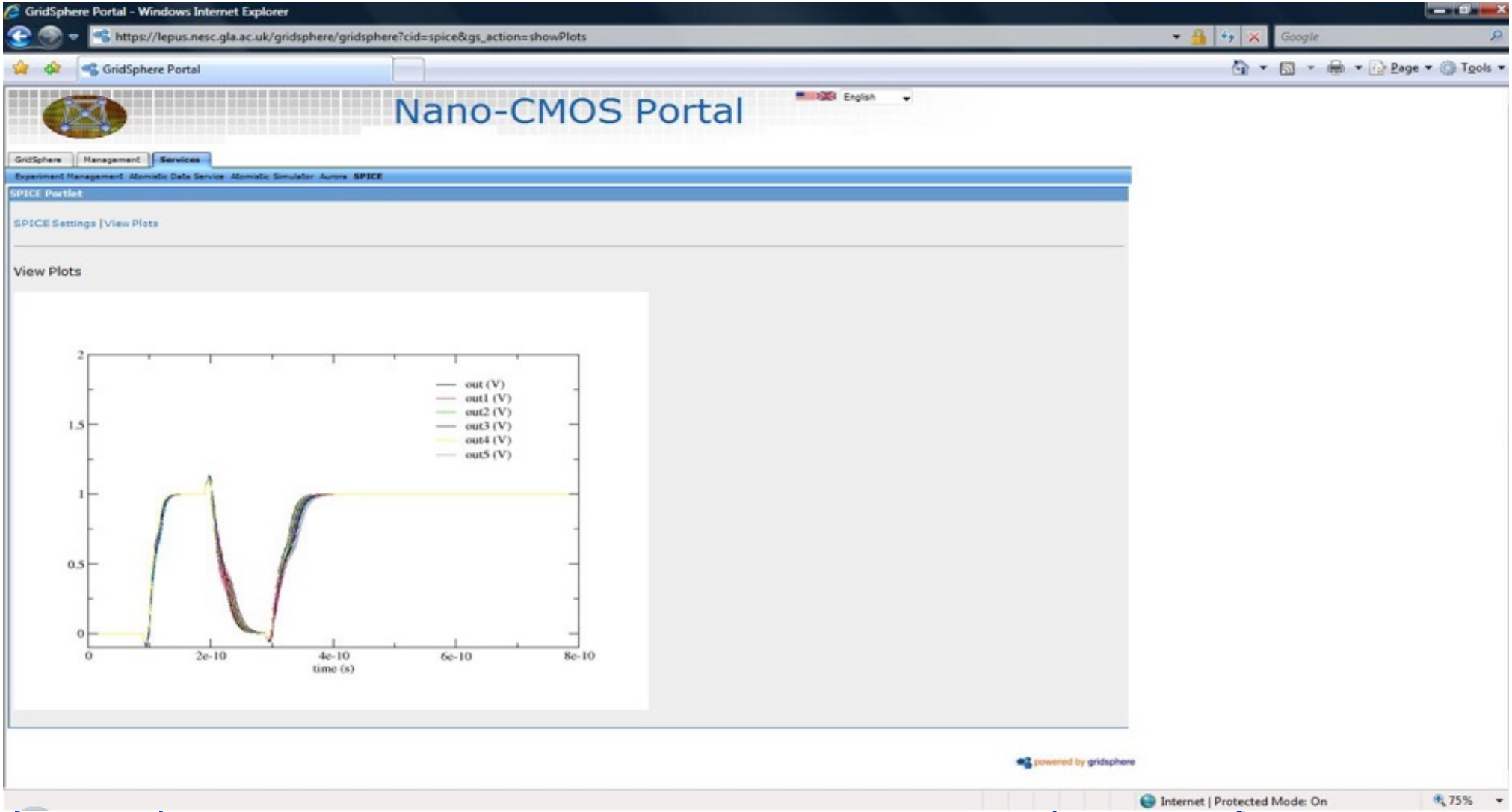
Challenges Hierarchical statistical system simulations



- 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

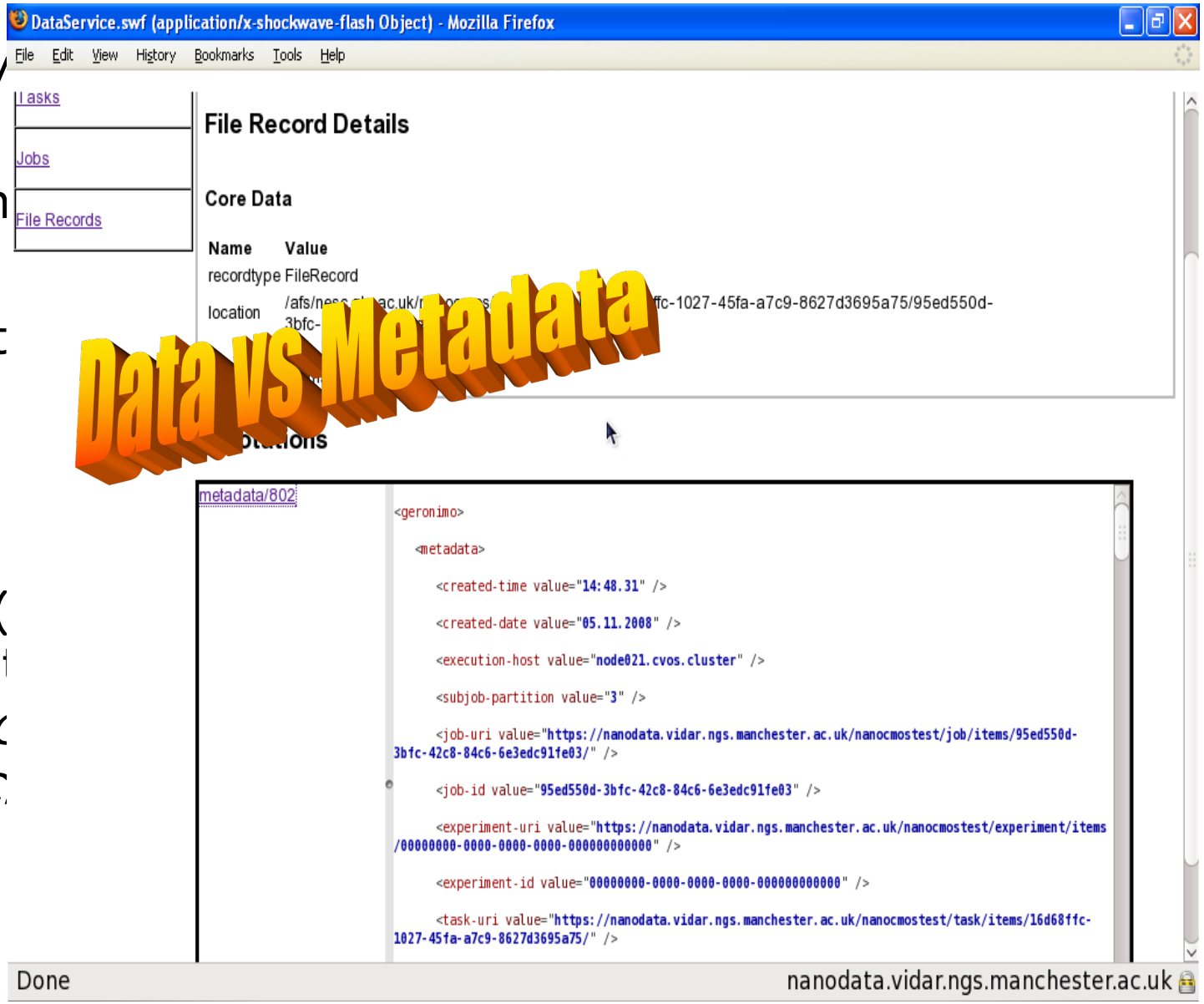
Secure, distribut

Meta-data captu

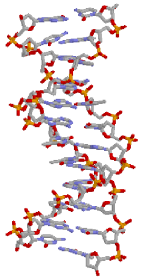
Job submission (
massive (at the

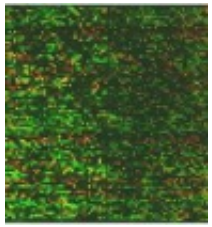
- *ScotGrid, NC*
- *Millions of C*

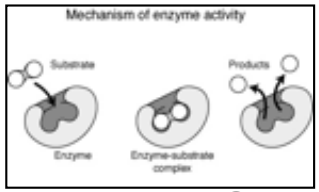
The –g flag!!!

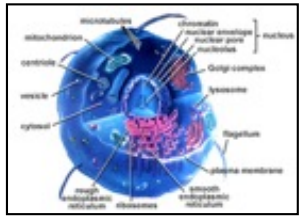


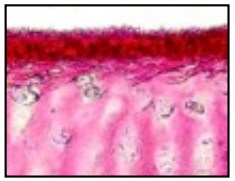
The e-Health Future…

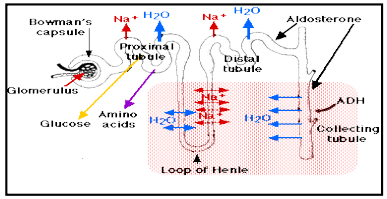






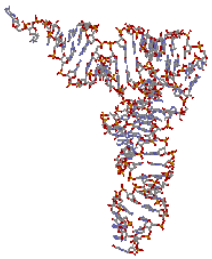


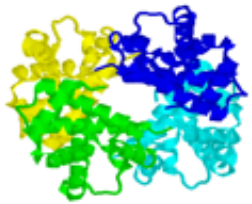


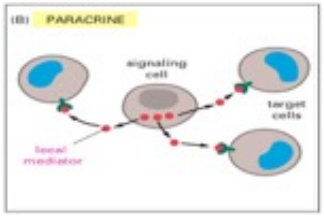


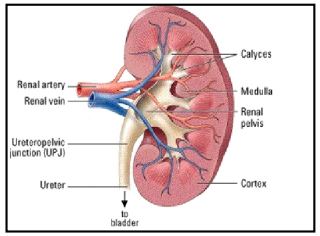
Nucleotide sequences
Nucleotide structures
Gene expressions
Protein Structures
Protein functions
Protein-protein interaction (pathways)
Cell
Cell signalling
Tissues
Organs
Physiology
Organisms
Populations

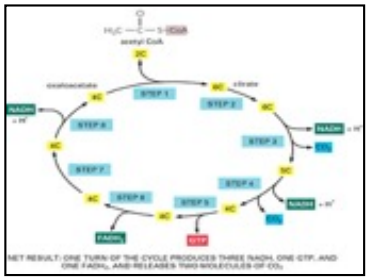
Security











+ 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?
  - …

More (and more and more) genomes…

Distributed, completely heterogeneous data

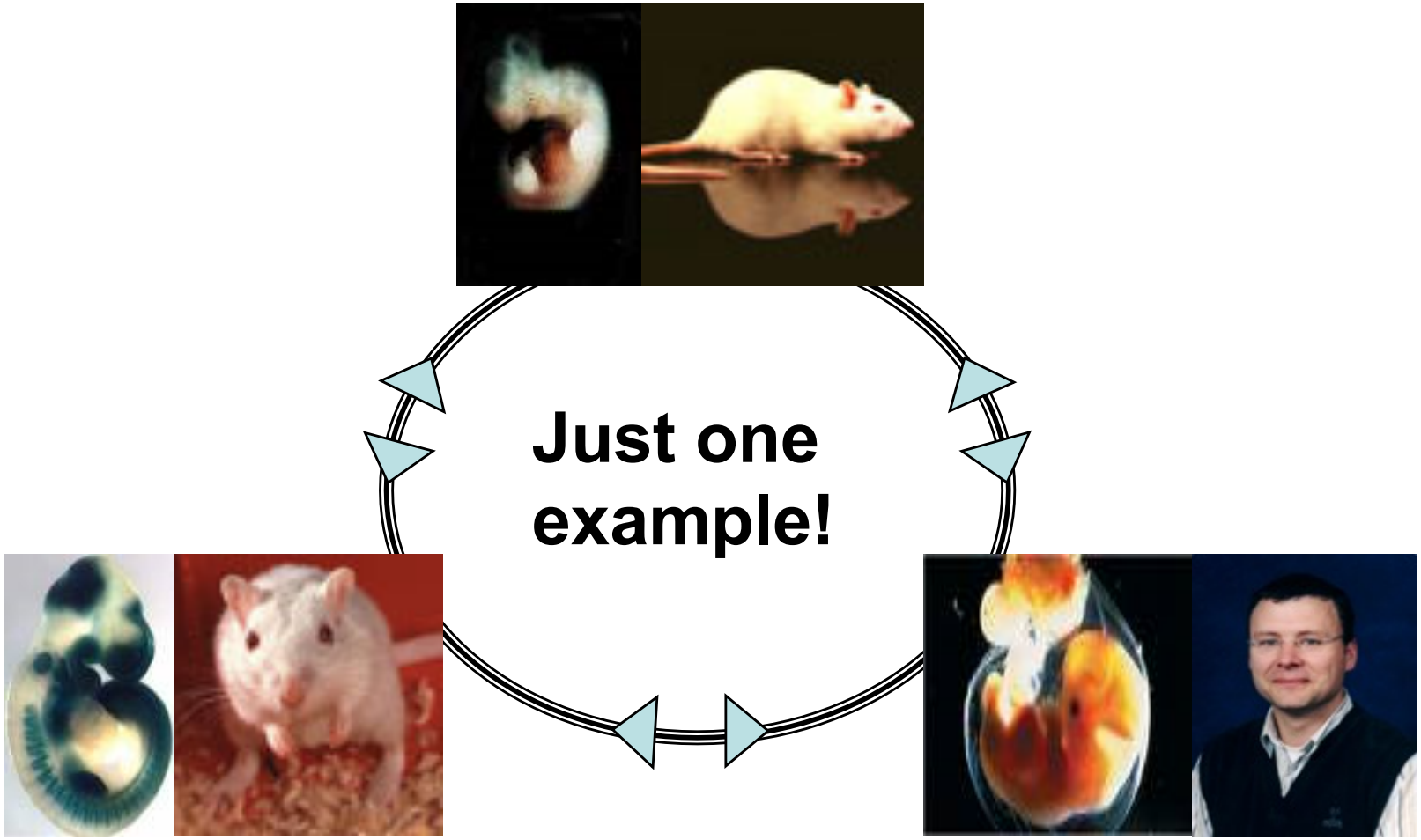
Messy

GPL96 - Notepad

File Edit Format View Help

```
^Annotation!Annotation_date = 09/20/2006 15:35:01!Annotation_platform = GPL96!Annotation_platform_title = Affymetrix GeneChip Human Genome U
n replication factor C, 40-kDa subunit (A1) mRNA, complete cds  1590810 M87338  RFC2                    117_at heat shock 70kDa protein 6 (HSP
5                    1431_at        cytochrome P450, family 2, subfamily E, polypeptide 1          Human cytochrome P450IIE1 (ethanol-indu
an farnesyl-protein transferase beta-subunit mRNA, complete cds 292032  L00635  FNTB                    177_at phospholipase D1, phosphatidylc
86095   NM_000991       RPL28                   200004_at       eukaryotic translation initiation factor 4 gamma, 2             Homo sapiens eu
                200010_at       ribosomal protein L11           Homo sapiens ribosomal protein L11 (RPL11), mRNA        15431289        NM_0009
_002954 RPS27A                  200018_at       ribosomal protein S13           Homo sapiens ribosomal protein S13 (RPS13), mRNA        1459191
S5 (RPS5), mRNA 71164878        NM_001009       RPS5                    200025_s_at     ribosomal protein L27           Homo sapiens ribosomal
(RPS11), mRNA   34335149        NM_001015       RPS11                   200032_s_at     ribosomal protein L9            Homo sapiens ribosomal
200039_s_at     proteasome (prosome, macropain) subunit, beta type, 2           Homo sapiens proteasome (prosome, macropain) subunit, beta type
member 1                Homo sapiens ATP-binding cassette, sub-family F (GCN20), member 1 (ABCF1), transcript variant 1, mRNA   69354731
ing factor 2, 45kDa (ILF2), mRNA        24234746        NM_004515       ILF2                    200053_at                       ated antigen 7
SCC3L1                  200059_s_at     ras homolog gene family, member A               Homo sapiens                    member A, mRNA (c
2288    AF275719        HSP90AB1                        200065_s_at     ADP-ribosyl                             clone 24537 ADP-ri
eterogeneous nuclear ribonucleoprotein M                Homo sapiens M                  3126877 AF061832
RNA (cDNA clone MGC:4498 IMAGE:2964510), complete cds   3387258                                 200079_s_at     lysyl-tRNA synt
014267  C11orf58                200085_s                                        ptide 2 (18kDa, elongin B)
rase, CAAX box, alpha (FNTA), transcript vari                                   lpha (FNTA), transcript variant 3, mRNA
omo sapiens ATPase, H+ transporting,                                    NM_003945       ATP6V0E
U                       200594_x_at                             ment factor A)                  Homo sapiens he
, mRNA  4507676 NM_003299       HSP                             Da beta (Grp94), member 1       Homo sapiens he
ein kinase, cAMP-dependent, regulat                             A), transcript variant 3, mRNA  47132579///4713
006265  RAD21                  200                     037     D repeat domain 1 (WDR1), transcript variant 1, m
rotein complex 2, beta 1 subunit (AP2B1), tra                   037     NM_001030006///NM_001282        AP2B1
n 3 (phosphorylase kinase, delta)                               sphorylase kinase, delta) (CALM3), mRNA 58218967        NM_0051
mRNA    47419913        NM_004184       WARS                    200             at      SET translocation (myeloid leukemia-associated)         Homo sa
Homo sapiens protein tyrosine phosphatase, receptor type, F (PTPRF), transcript variant 1, mRNA 109633040      NM_002840       PTPRF
yptophan 5-monooxygenase activation protein, zeta polypeptide           Homo sapiens tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activa
0646_s_at       nucleobindin 1          Homo sapiens nucleobindin 1 (NUCB1), mRNA       39725676        NM_006184       NUCB1
d protein beta)         Homo sapiens signal sequence receptor, beta (translocon-associated protein beta) (SSR2), mRNA   6552341 NM_003145
ide translocator), member 5 (SLC25A5), mRNA     4502098 NM_001152       SLC25A5                 200658_s_at     prohibitin              Homo sa
NM_006145       DNAJB1                  200665_s_at     secreted protein, acidic, cysteine-rich (osteonectin)           Homo sapiens secreted p
g protein 1 (XBP1), mRNA        14110394        NM_005080       XBP1                    200671_s_at     spectrin, beta, non-erythrocytic 1
7977    NM_003347       UBE2L3                  200677_at       pituitary tumor-transforming 1 interacting protein              Homo sapiens pi
nt 1, mRNA///Homo sapiens ubiquitin-conjugating enzyme E2L 3 (UBE2L3), transcript variant 2, mRNA       38157977///38157975     NM_003347///NM_
on factor 1 gamma               Homo sapiens eukaryotic translation elongation factor 1 gamma (EEF1G), mRNA     83656774        NM_001404
D (Asp-Glu-Ala-Asp) box polypeptide 24 (DDX24), mRNA    14251213        NM_020414       DDX24                   200695_at       protein phospha
mRNA    8051609 NM_006854       KDELR2                  200700_s_at     KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor
6127    NM_001959       EEF1B2                  200706_s_at     lipopolysaccharide-induced TNF factor           Homo sapiens lipopolysaccharide
bule-associated protein, RP/EB family, member 1         Homo sapiens microtubule-associated protein, RP/EB family, member 1 (MAPRE1), mRNA
phase kinase-associated protein 1A (p19A) (SKP1A), transcript variant 2, mRNA   25777710///25777712     NM_006930///NM_170679   SKP1A
t 2, mRNA       61676201///61676202     NM_005898///NM_203364   GPIAP1                  200723_s_at     GPI-anchored membrane protein 1
apiens ARP2 actin-related protein 2 homolog (yeast) (ACTR2), transcript variant 1, mRNA///Homo sapiens ARP2 actin-related protein 2 homolog (ye
AX1 (hPTPCAAX1) mRNA, complete cds      1777754 U48296  PTP4A1                   200734_s_at     ADP-ribosylation factor 3               Homo sa
3 (S. cerevisiae) (SUMO3), mRNA 48928057        NM_006936       SUMO3                   200741_s_at     ribosomal protein S27 (metallopanstimul
6       NM_002074       GNB1                    200747_s_at     nuclear mitotic apparatus protein 1             Homo sapiens nuclear mitotic ap
e/serine-rich 2         Homo sapiens splicing factor, arginine/serine-rich 2 (SFRS2), mRNA      47271442        NM_003016       SFRS2
P-ribosylation-like factor 6 interacting protein 5              Homo sapiens ADP-ribosylation-like factor 6 interacting protein 5 (ARL6IP5), mR
nce similarity 120A (FAM120A), mRNA     68299753        NM_014612       FAM120A                 200768_s_at     methionine adenosyltransferase
rity 120A (FAM120A), mRNA       68299753        NM_014612       FAM120A                 200775_s_at     heterogeneous nuclear ribonucleoprotein
491///NM_001008492///NM_004404///NM_006155      SEPT2                   200779_at       activating transcription factor 4 (tax-responsive enhan
n-related protein 1 (alpha-2-macroglobulin receptor)            Homo sapiens low density lipoprotein-related protein 1 (alpha-2-macroglobulin r
sapiens IQ motif containing GTPase activating protein 1 (IQGAP1), mRNA  57242794        NM_003870       IQGAP1                  200792_at
```

Translational Research



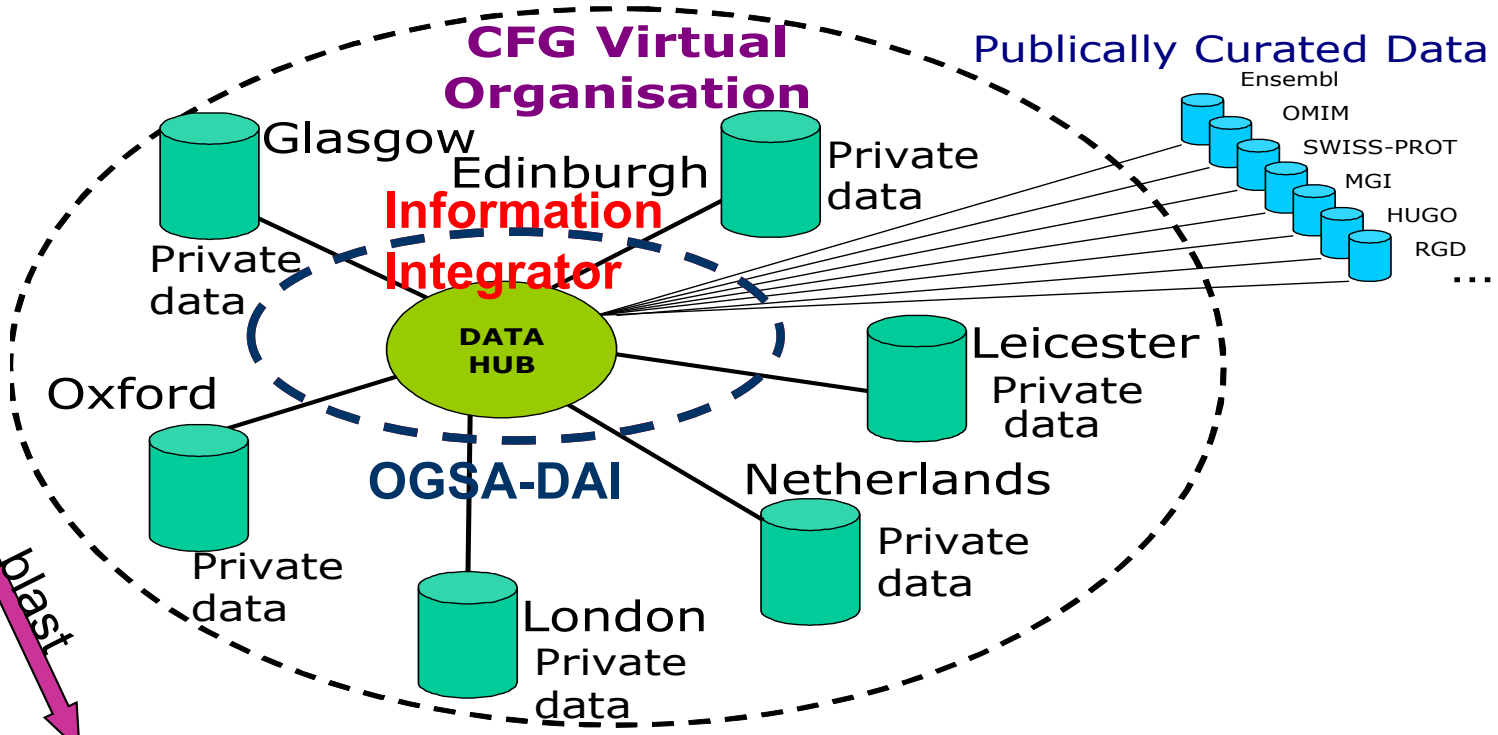
BRIDGES Project

VO Authorisation

Synteny Service

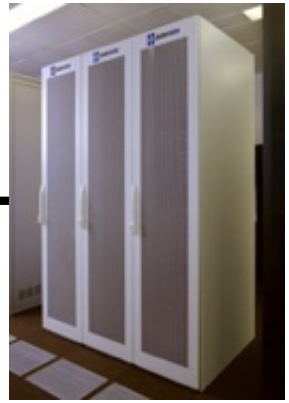
Magna Vista Service

blast

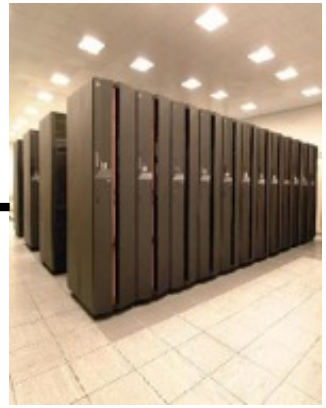


+

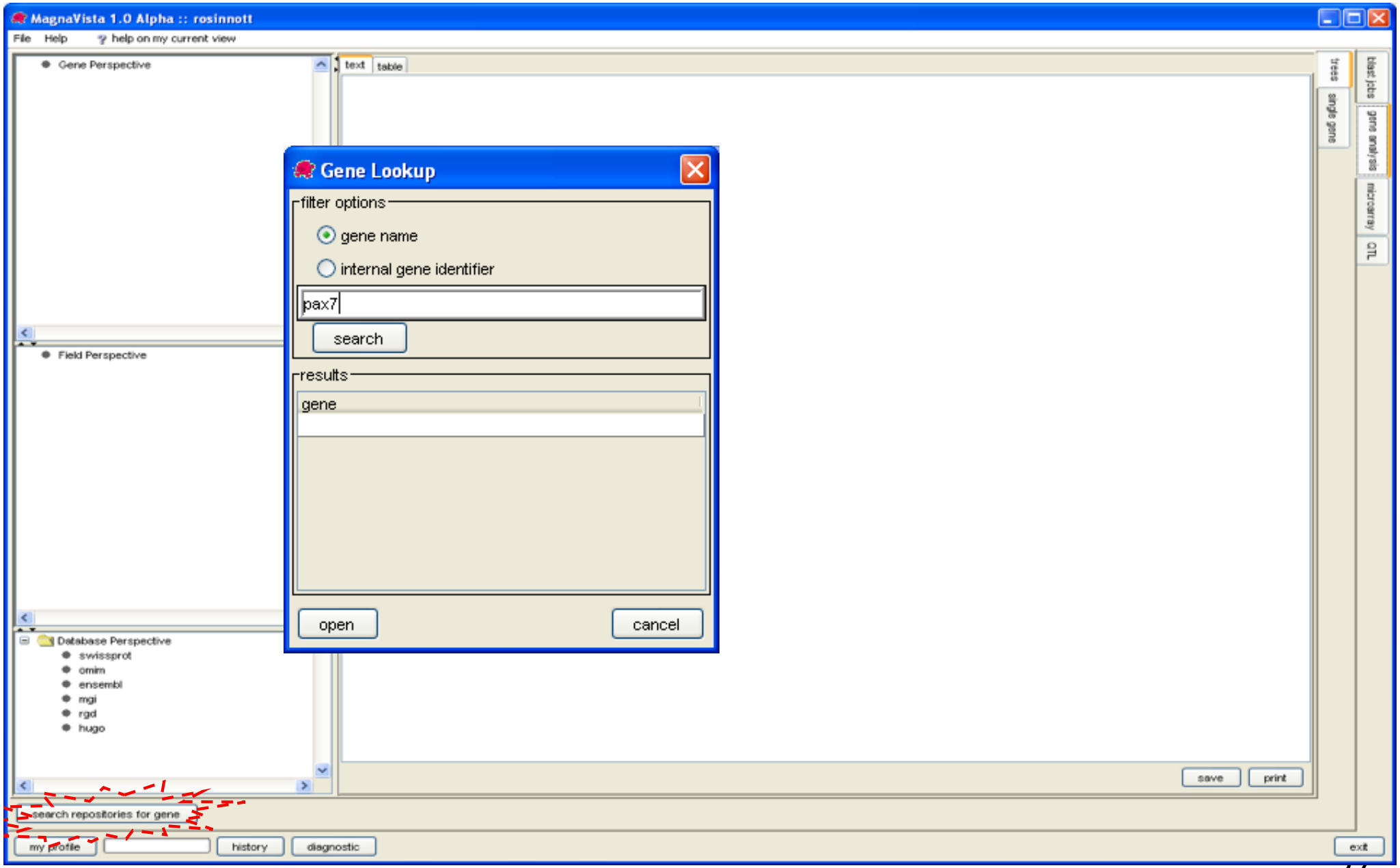
+



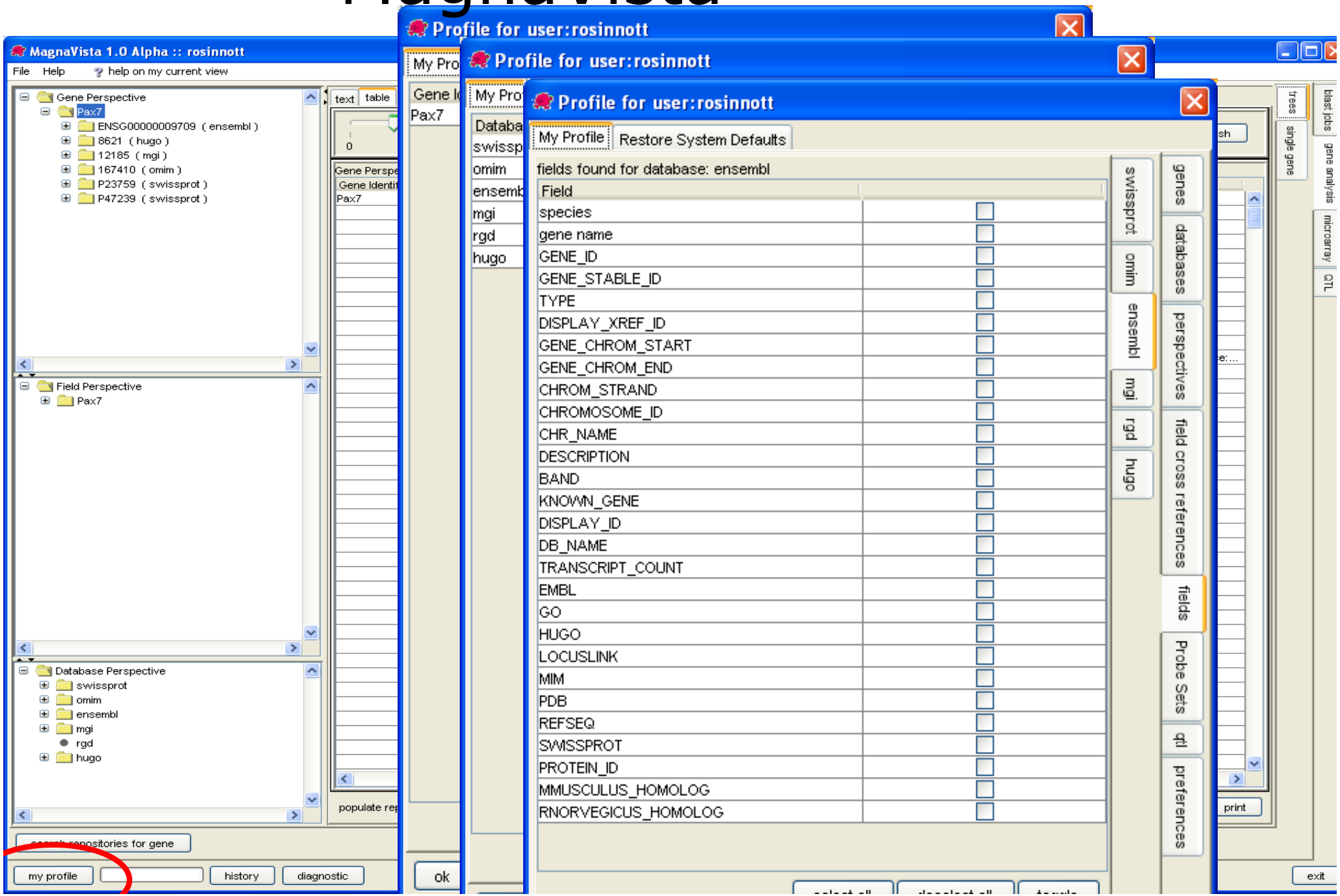
+



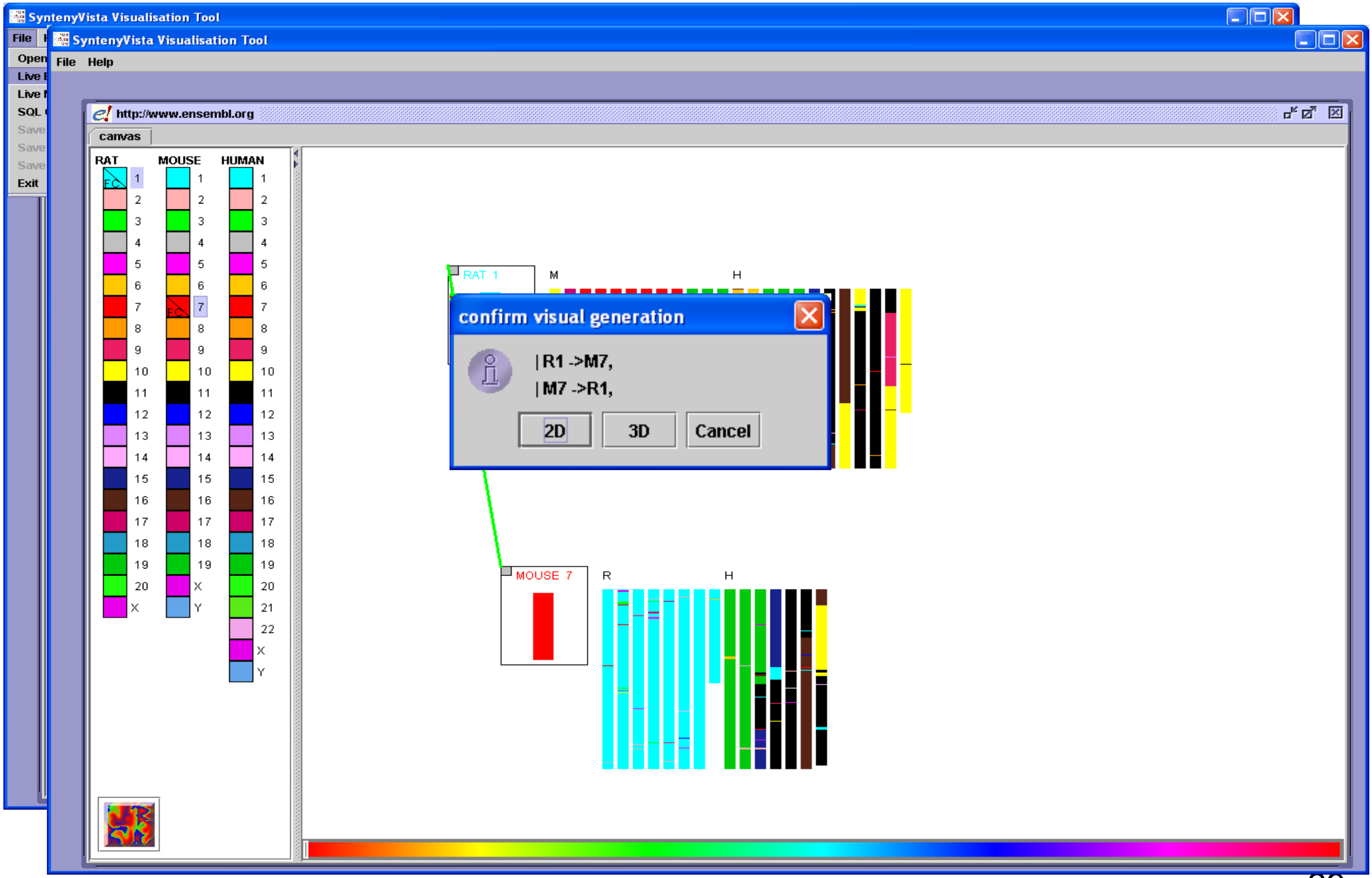
MagnaVista

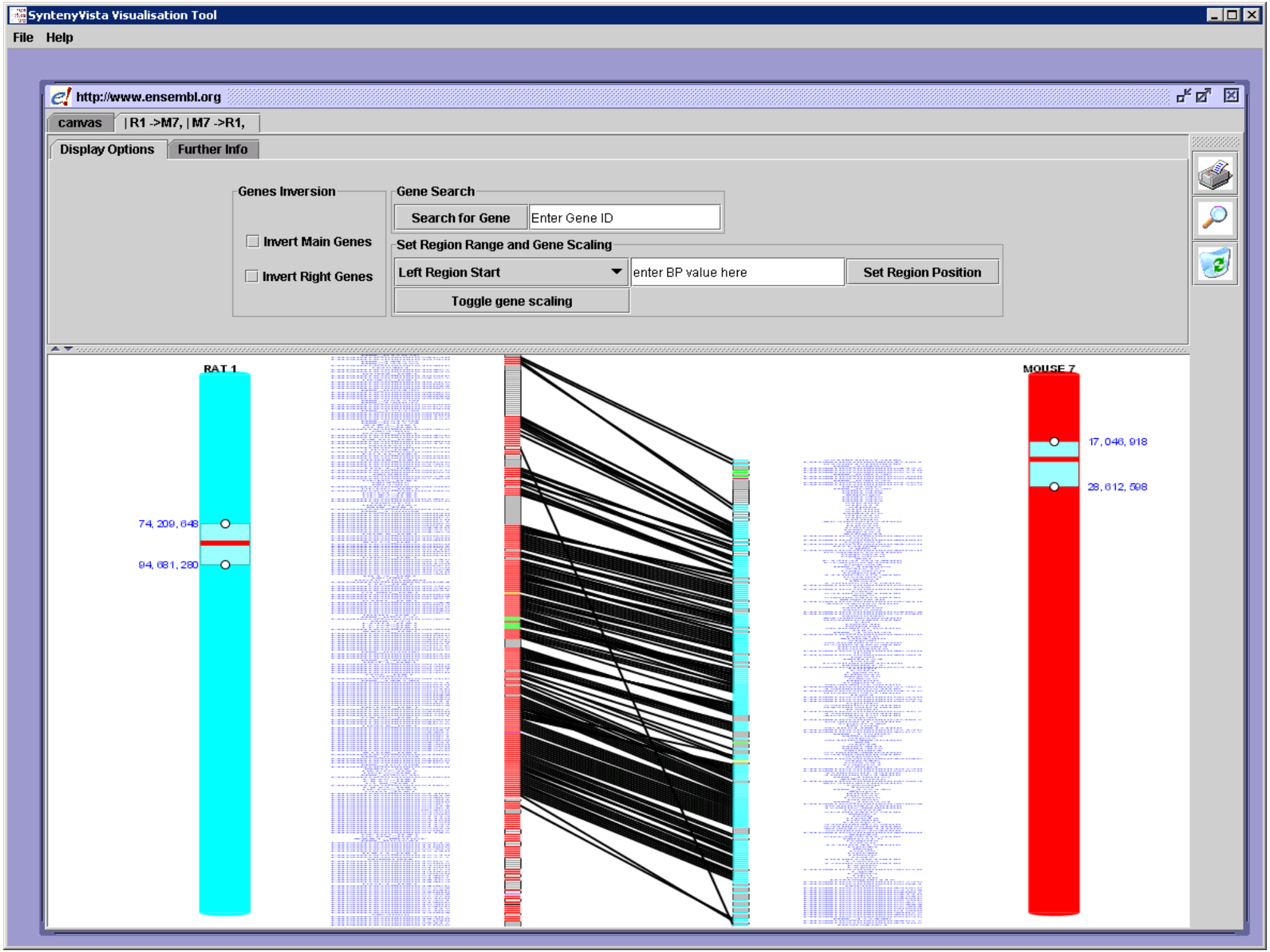

MagnaVista 1.0 Alpha :: rosinnott
File
Help
help on my current view
Gene Perspective
text
table
trees
single gene
blast jobs
gene analysis
microarray
QTL
Gene Lookup
filter options
gene name
internal gene identifier
pax7
search
results
gene
open
cancel
Field Perspective
Database Perspective
swissprot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

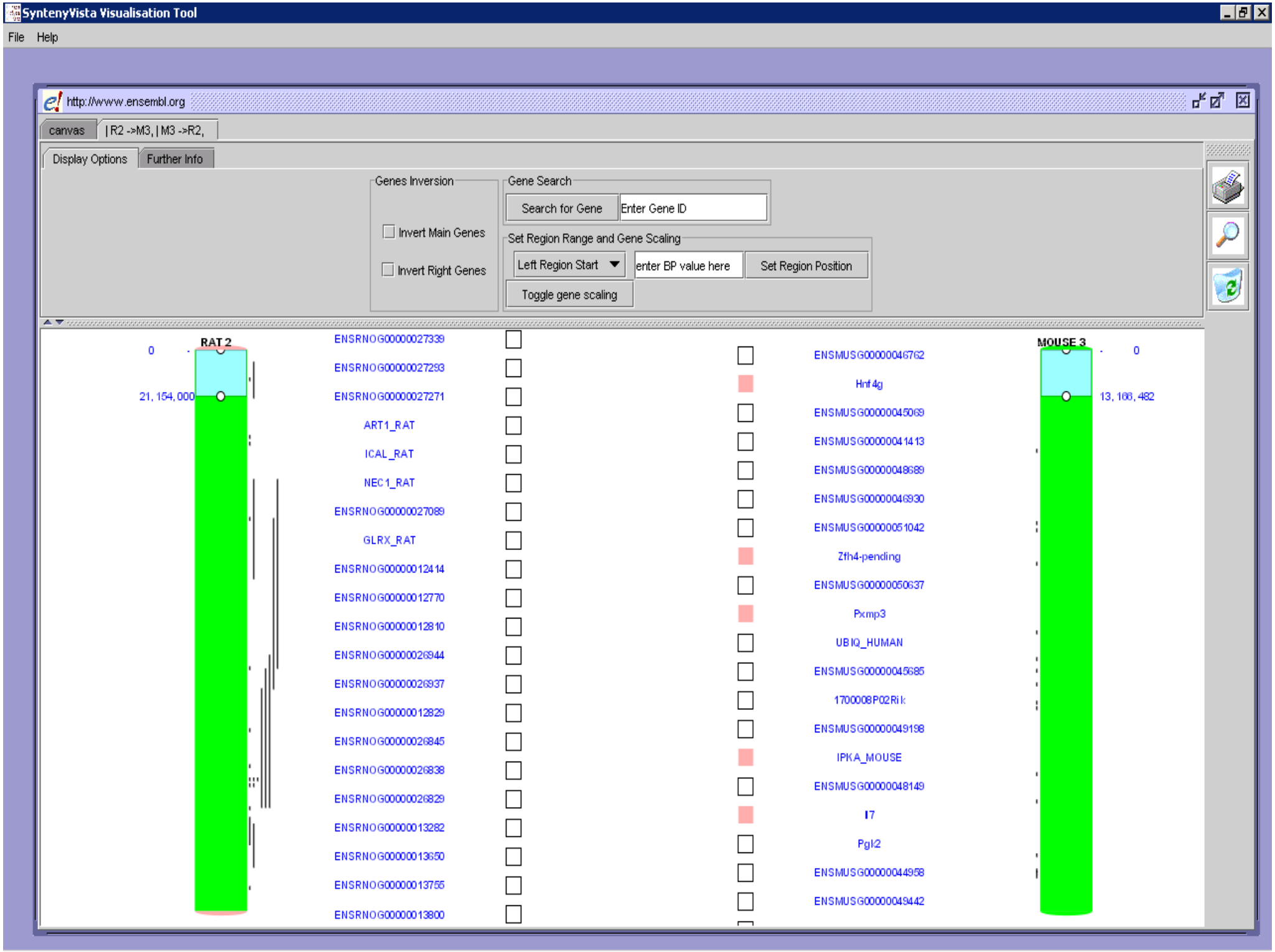
MagnaVista


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



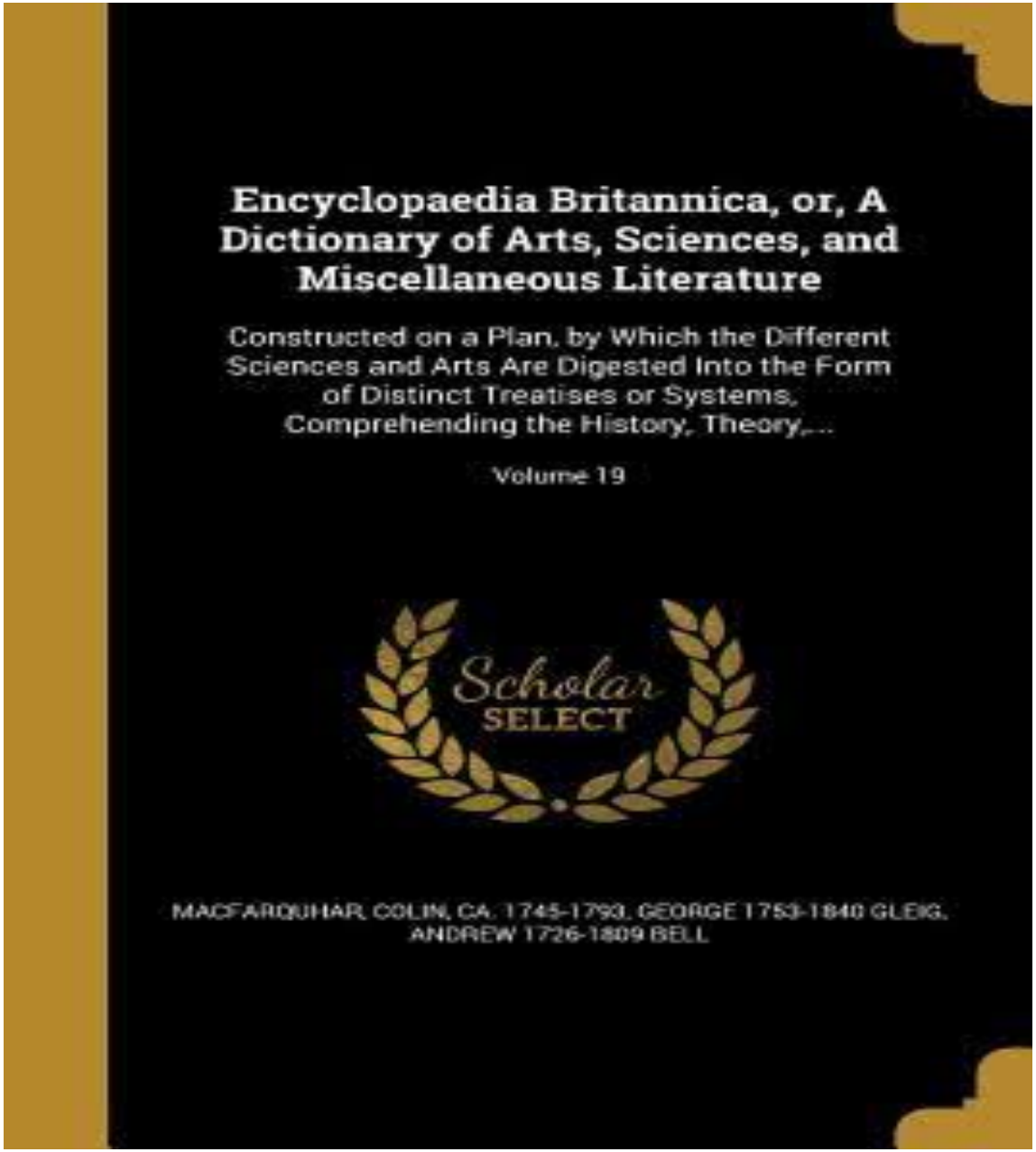

SyntenyVista Visualisation Tool
File
Help
http://www.ensembl.org
canvas
| R1 ->M7, | M7 ->R1,
Display Options
Further Info
Genes Inversion
Invert Main Genes
Invert Right Genes
Gene Search
Search for Gene
Enter Gene ID
Set Region Range and Gene Scaling
Left Region Start
enter BP value here
Set Region 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, 166, 482

BREAK

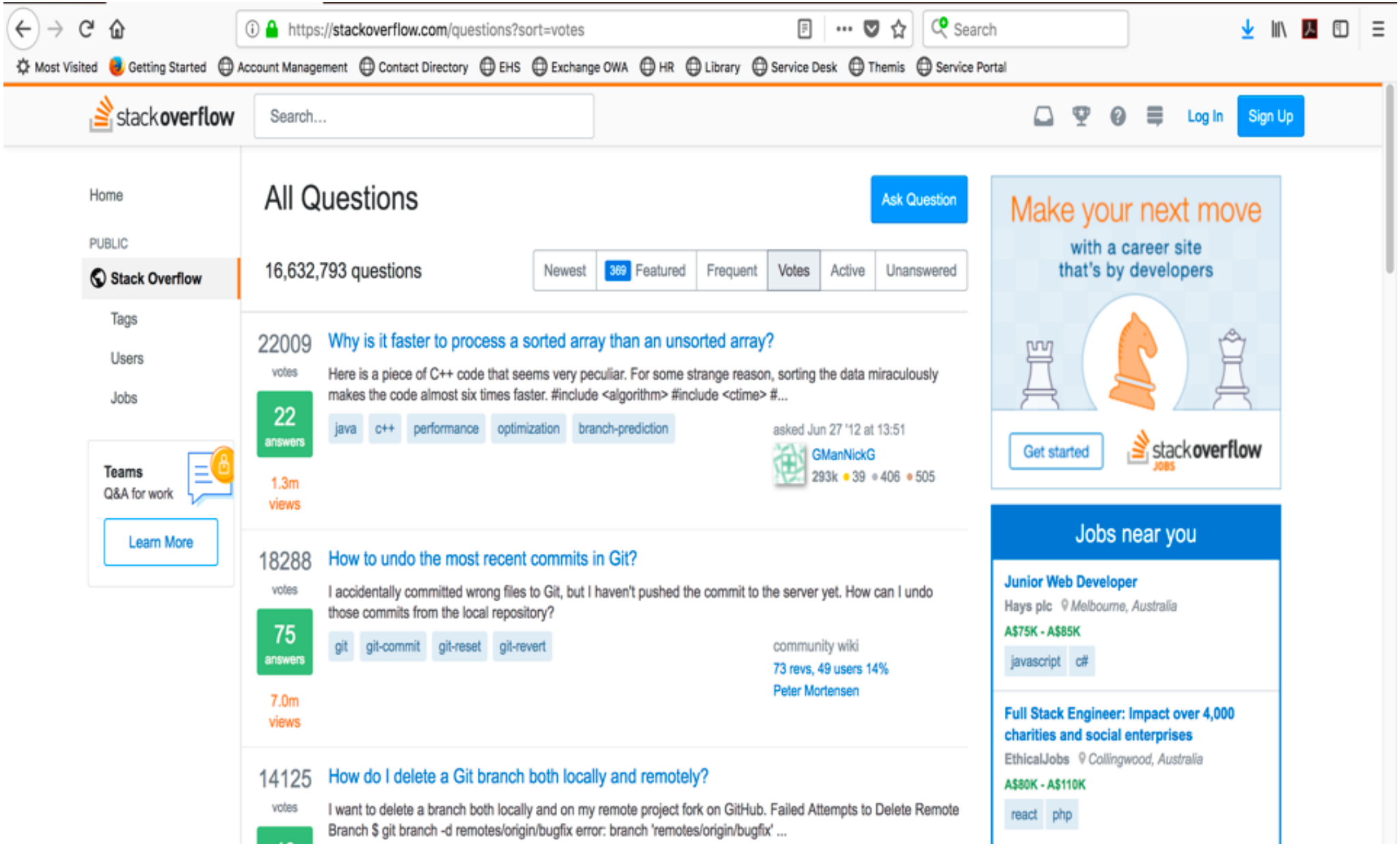
Data –> Knowledge?

- Once upon a time…



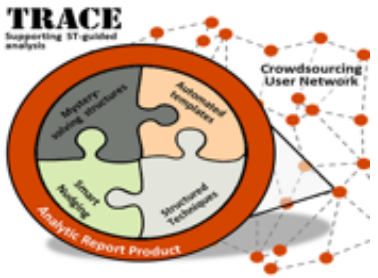
Crowdsourcing Knowledge & Reasoning

- Many approaches that *work (?)*…



CREATE Program

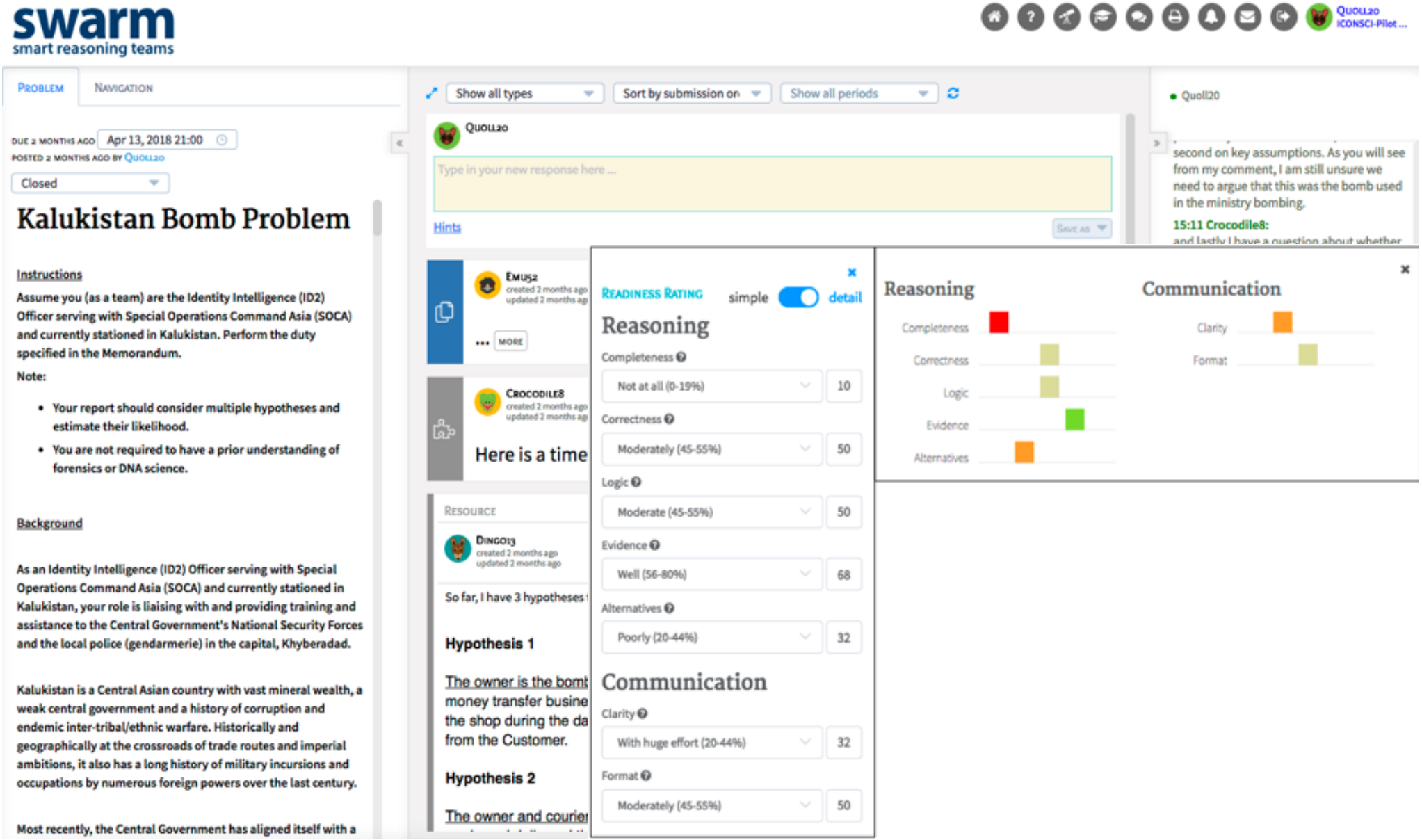
- CREATE
  - Commenced in 2017
  - Involve(d) four teams*
    - TRACE – Trackable Reasoning and Analysis for Collaboration and Evaluation (Syracuse)
    - Co-Arg – Cogent Argumentation System with Crowd Elicitation (George Mason University)
    - BARD – Bayesian Argumentation via Delphi (Monash)
    - SWARM – Smartly-assembled Wiki-style Argument Marshalling (UniMelb)
      - www.swarmproject.info



* Interesting paradigm of funding…

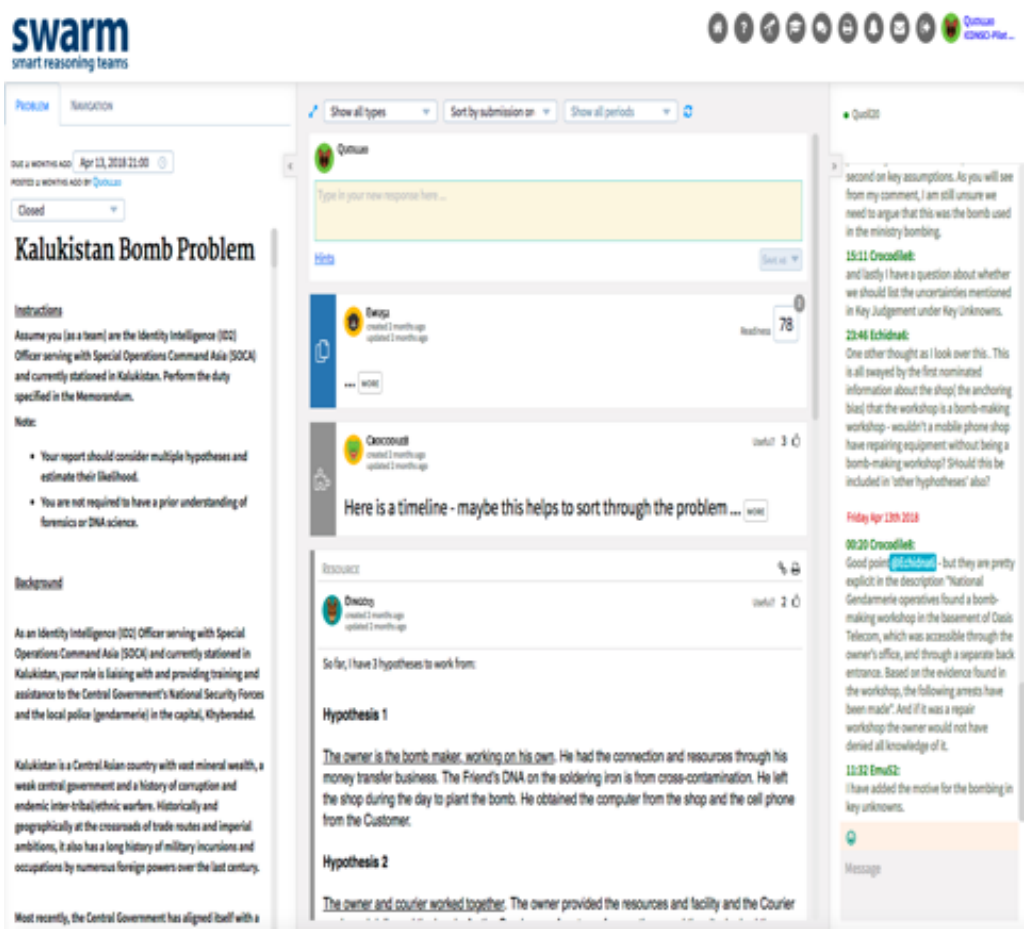
SWARM Overview

- Solution overview



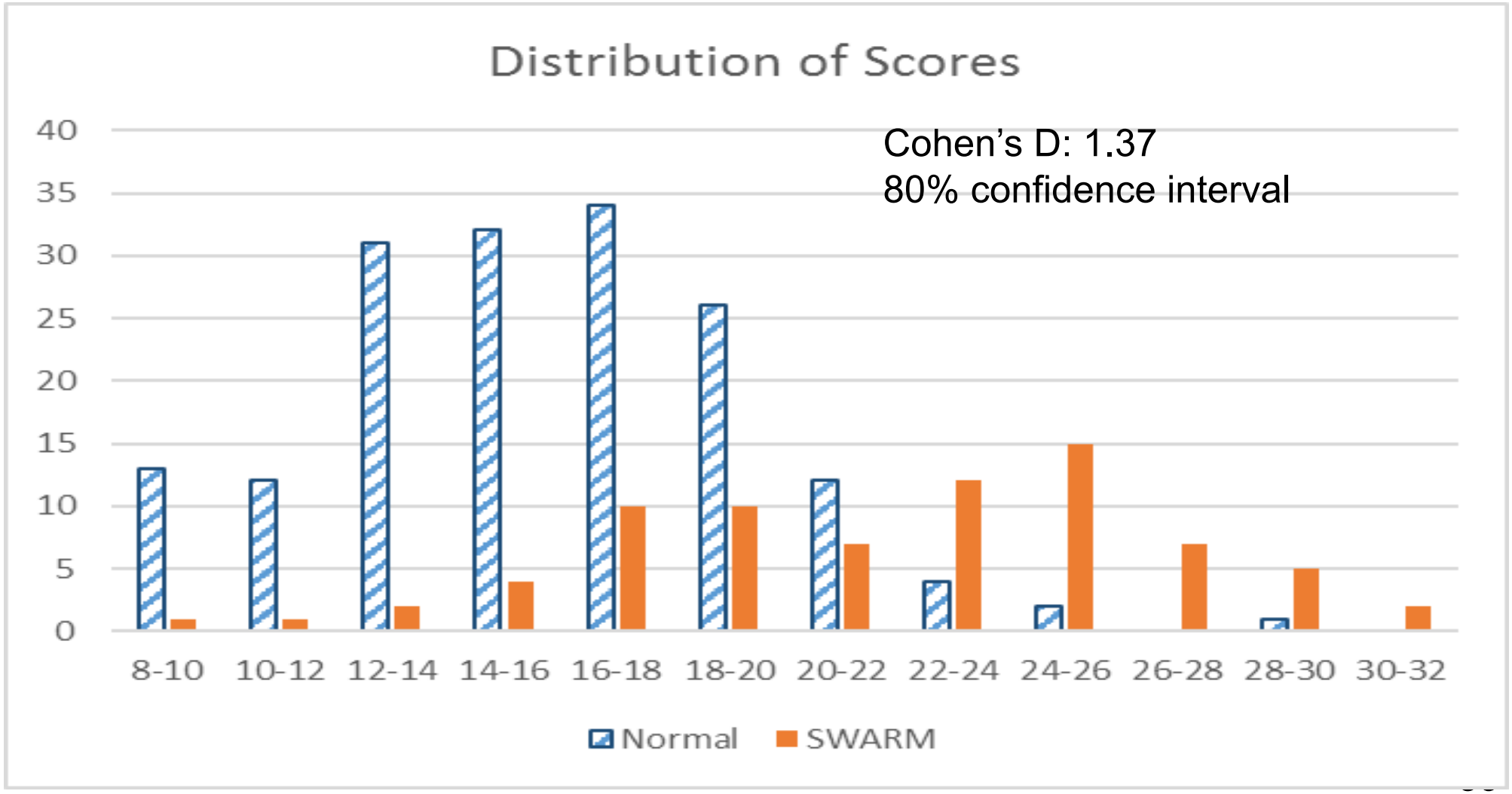
SWARM Overview

- Key features
  - Anonymous
    - Avatars and rating
  - No leader
  - Team size flexibility
    - Dockerized AWS/NeCTAR
    - Kubernetes/Docker SWARM
  - Public vs Private
    - Off platform work supported
  - Arbitrary contributions
    - No payments/obligations
  - Social interactions (chat)
    - Social warmth
  - Ad hoc use encouraged



Proof of the Pudding

- Does SWARM help to reason better?
  - ASIO, VicPolice, …
    - 81 reports on platform; 167 off platform (normal)



SWARM Clouds

- Developed on openStack (NeCTAR) 

  - Scripted solution using Docker & Kubernetes
- Deployed to AWS (US) 

  - $1000+ / month for basic use
    - (costs ramp up a LOT!)
- Benefits of scripted solution
  - Developed/test/trialled on free Cloud (NeCTAR)
  - Deployed to AWS when ready
  - Not possible to do if would have used AWS Elastic Container Solution for Kubernetes (EKS)
    - (…or Azure AKS or Google GKE etc etc)

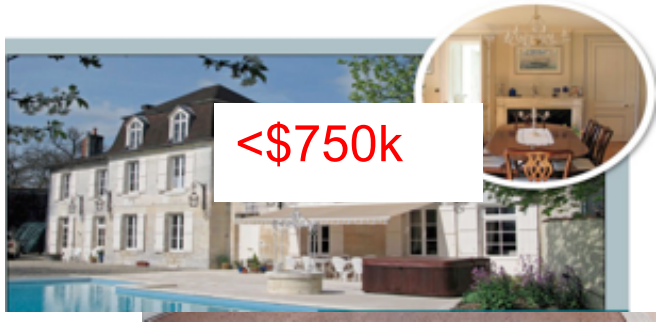
Social Sciences

- Data, data everywhere
  - Researchers exploring all kinds of areas with societal impact
  - Many major resources distributed around the globe
    - Data Archives
      - 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
  - $40m+ project (+$18.9m)
  - 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



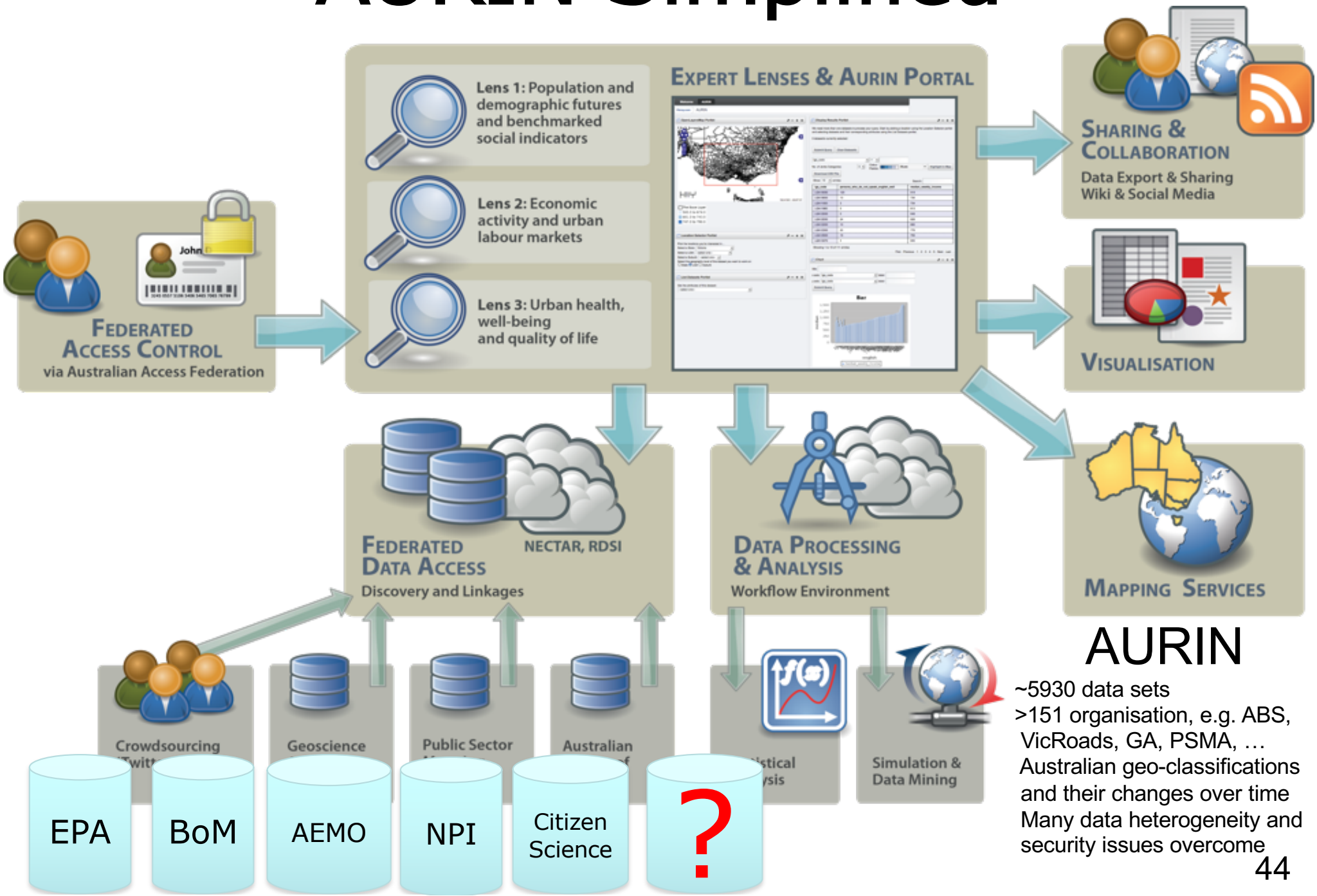
- Urban and build environment is extremely broad
  - health,
  - transport,
  - future population,
  - liveability,
  - crime,
  - housing,
  - design,
  - …




- 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)

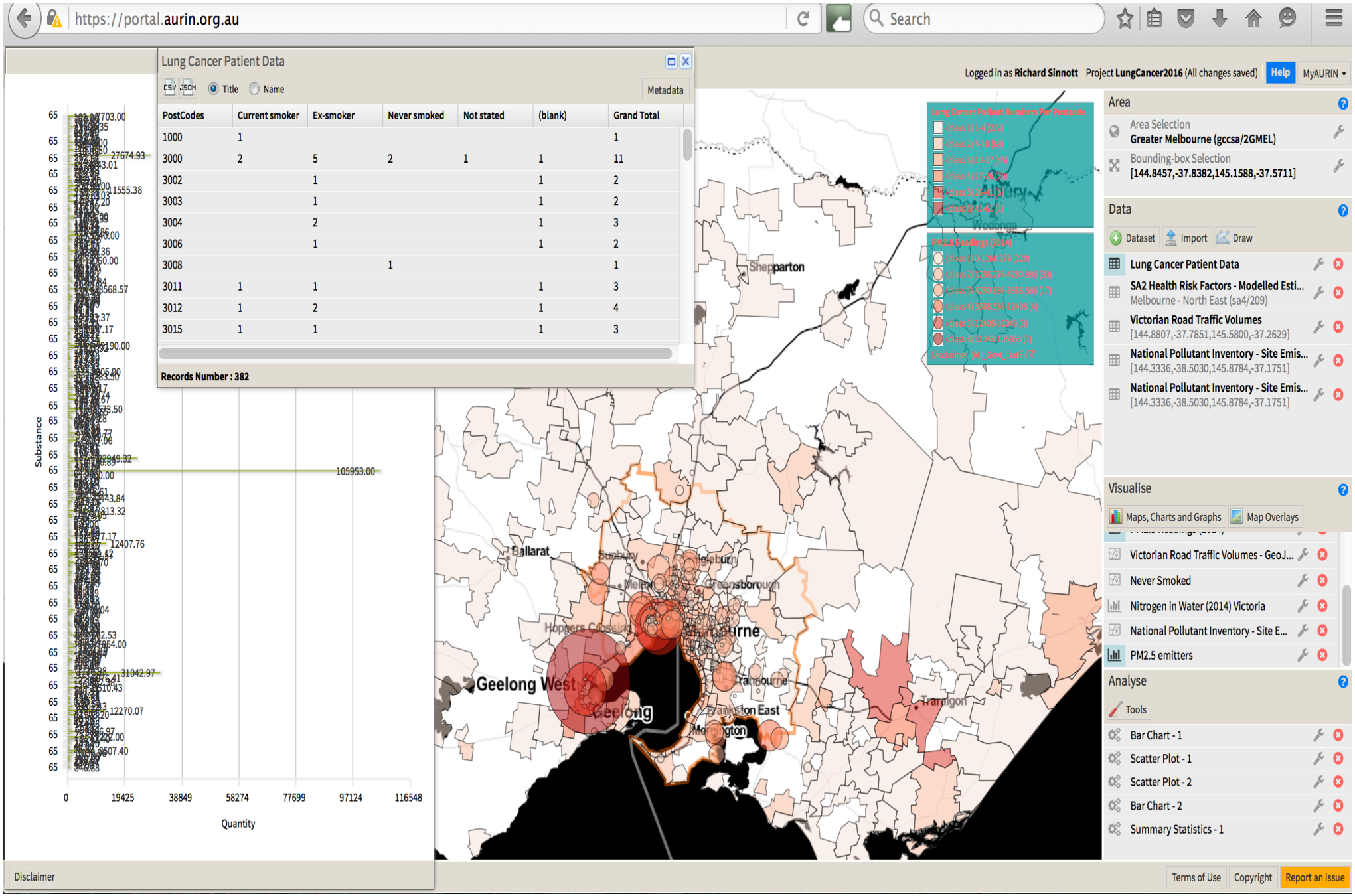
AURIN Simplified



AURIN

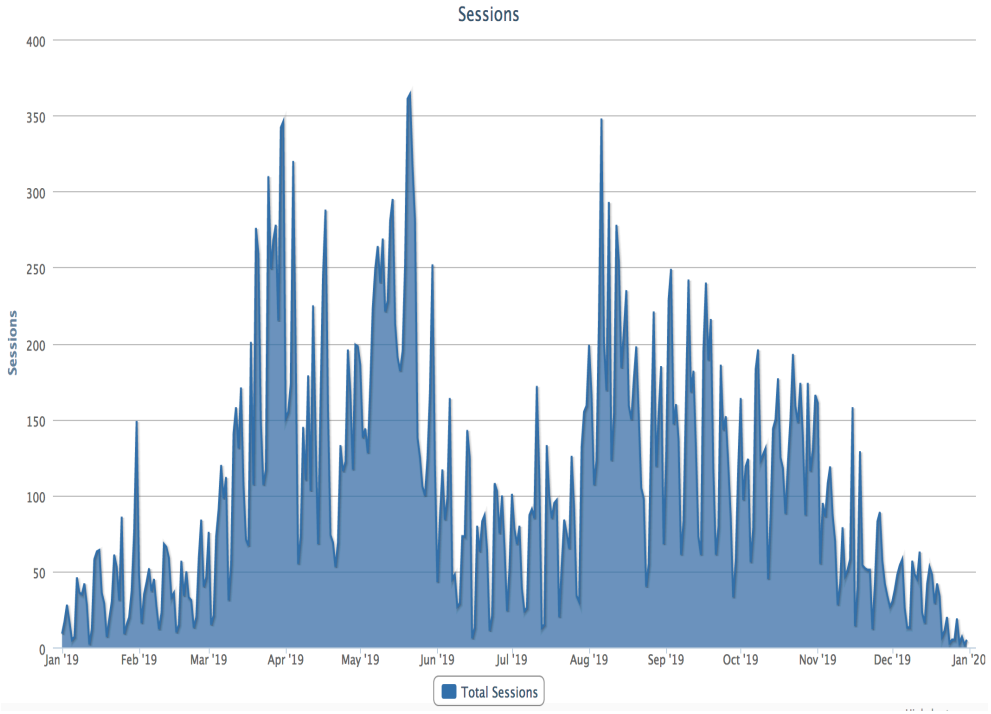
~5930 data sets
>151 organisation, e.g. ABS, VicRoads, GA, PSMA, …
Australian geo-classifications and their changes over time
Many data heterogeneity and security issues overcome

AURIN Example (in one slide!)



AURIN Clouds

- Original plan to use NeCTAR
  - Early reliability issues
- Actual plan
  - Servers purchased (VMware)
  - Used for production system
    - 8 years old/partial refresh 2016
    - 3 days outage in 8 years!!!
  - Failover system on NeCTAR
- Dev, Staging, Production
  - 20,000+ users
  - ~4million lines of code
    - CouchDB, web services (ReST, …), geoJSON, …
  - Much of the code not written by my team!!!
  - Move to using Docker containers/Kubernetes ongoing
    - e.g. L. Chen, Y. Pan, R.O. Sinnott, *Auto-Scaling a Cloud-based Walkability Tool through Kubernetes and Docker Swarm*, CLOSER 2020, Prague, Czech Republic, May 2020.



Demonstration in Workshops

(note – Assignment II)

AURIN Homework

(https://portal.aurin.org.au)

Find the suburb (SA2) in Greater Melbourne that had the highest number of Jobseeker recipients in June 2020.

(not assessed!)

Questions ...?