Domain Drivers – Lecture 2

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
  - 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
    - eHealth & Bioinformatics

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

Project Portfolio

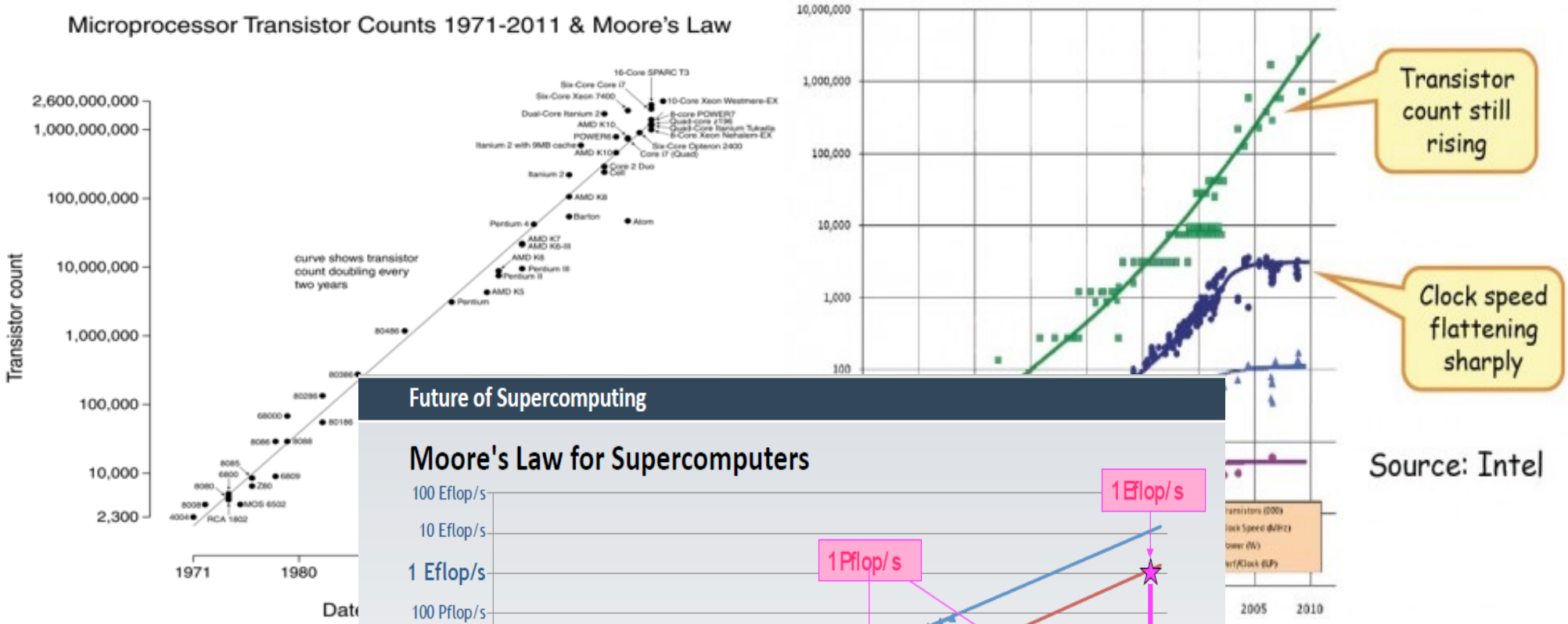
Completed

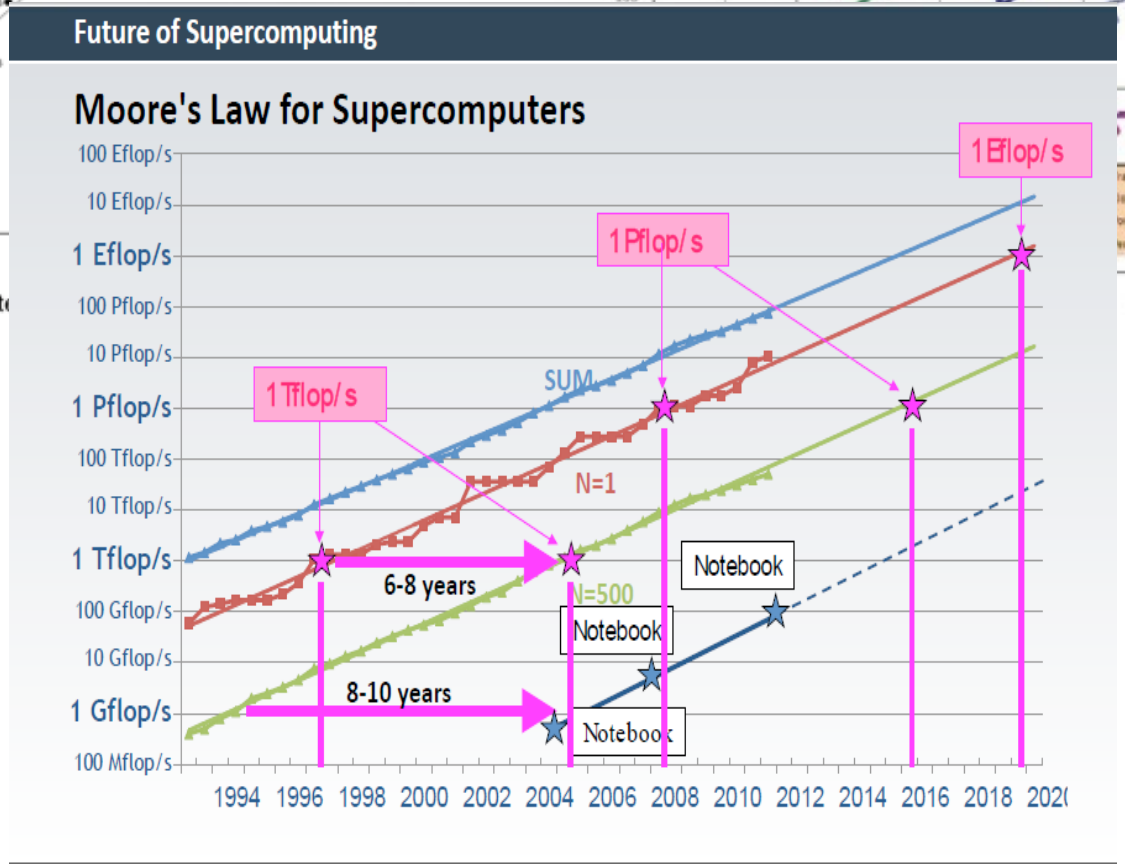
- National e-Science Centre (I, II, III)
- Dynamic Virtual Organisations for e-Science Education
- Biomedical Research Informatics Delivered by Grid Enabled Services
- GridNet, GridNet 2
- Grid Enabled Microarray Expression Profile Search
- Glasgow early adoption of Shibboleth
- Joint Data Standards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

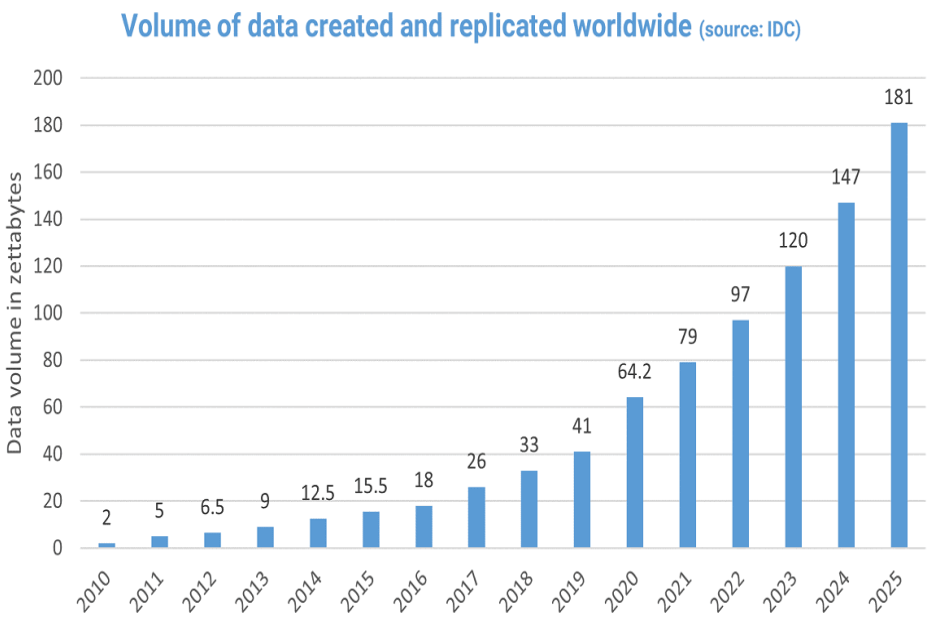
- EU European Platform for Study of Wolfram, Alstrom, Bardet Biedl (EuroWABB)
- Multicenter prospective study of biochemical profiles of monoamine-producing tumors (PMT Study)
- European Society of Hypertension Study on Pheo/PGL
- International DSD
- EU FW7 European Network for Study of Adrenal Tumors Cancer Research Platform (ENSAT-CANCER)
- VicHealth Health Indicators and Spatial Objective Data
- National Spinal Injury Research Platform
- Australian Urban Research Infrastructure Network (AURIN)/Spatial Urban Data Observatory (SUDO)
- 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
- 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



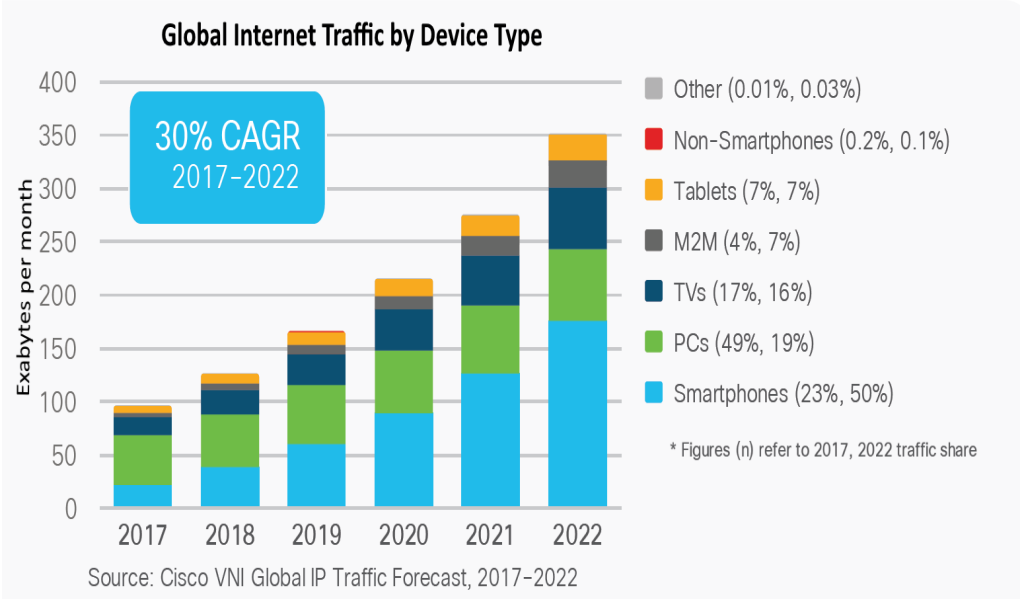


Table 1. The VNI Forecast Within Historical Context

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

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 Scaling::The Deluge

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

MG SIEGLER

Wednesday, August 4th, 2010 2 Comments

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.

Science News ... from universities, journals, and other research organizations

Big Data, for Better or Worse: 90% of World's Data Generated Over Last Two Years

May 22, 2013 — A full 90% of all the data in the world has been generated over the last two years. The internet companies are awash with data that can be grouped and utilised. Is this a good thing?

An increasing amount of data is becoming available on the internet. Each and every one of us is constantly producing and releasing data about ourselves. We do this either by moving around passively -- our behaviour being registered by cameras or card usage -- or by logging onto our PCs and surfing the net.

The volumes of data make up what has been designated 'Big Data' -- where data about individuals, groups and periods of time are combined into bigger groups or longer periods of time.

Research advantages

Petter Bae Brandtzæg of SINTEF ICT points to the huge research centres now developed at internet companies such as Facebook and Google.

'The advantage they have is the enormous volume of data that other social researchers can only dream of,' he says. However, it has also changed the way SINTEF researchers work. Even those not working in the major internet companies can still access Big Data.

Brandtzæg has investigated a tool called Wisdom developed by the American-based company MicroStrategy, and has started applying it in the deITA-project which addresses young people's social activity on the internet.

'This gives me access to data about over 20 million people -- without making a single inquiry. I can analyse different preferences on Facebook and look at age and gender differences between various groups and nations across the world. So far I have compared gender differences in social activity on Facebook between people in Norway, Spain, England, USA, Russia, Egypt, India and China.'

Data protection is a problem we often associate with Big Data, but according to Brandtzæg, data from Wisdom is restricted to large groups and does not go down to 'individual level'. This makes it possible for him to compare large groups without any data protection problems.

Big Data makes it possible to achieve research results that cover a wide range of issues, and can tell us a great deal about developments in the world in many different areas. (Credit: © zothen / Fotolia)

Related Topics

Computers & Math
- Computers and Internet
- Information Technology
- Encryption

Science & Society
- Security and Defense
- Privacy Issues
- Popular Culture

Articles
- List of data structures
- Computing
- Streaming media
- Identity theft
- Data mining
- Wi-Fi

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

1 TIPS is approximately 25,000 SpecInt95 equivalents

Offline Processor Farm ~20 TIPS

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

Tier 0

CERN Computer Centre

Tier 1

France Regional Centre

Germany Regional Centre

Italy Regional Centre

FermiLab ~4 TIPS

• • •

ScotGrid

Tier 2

Caltech ~1 TIPS

Tier2 Centre ~1 TIPS

~622 Mbits/sec

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

Institute ~0.25TIPS

Physics data cache

~1 MBytes/sec

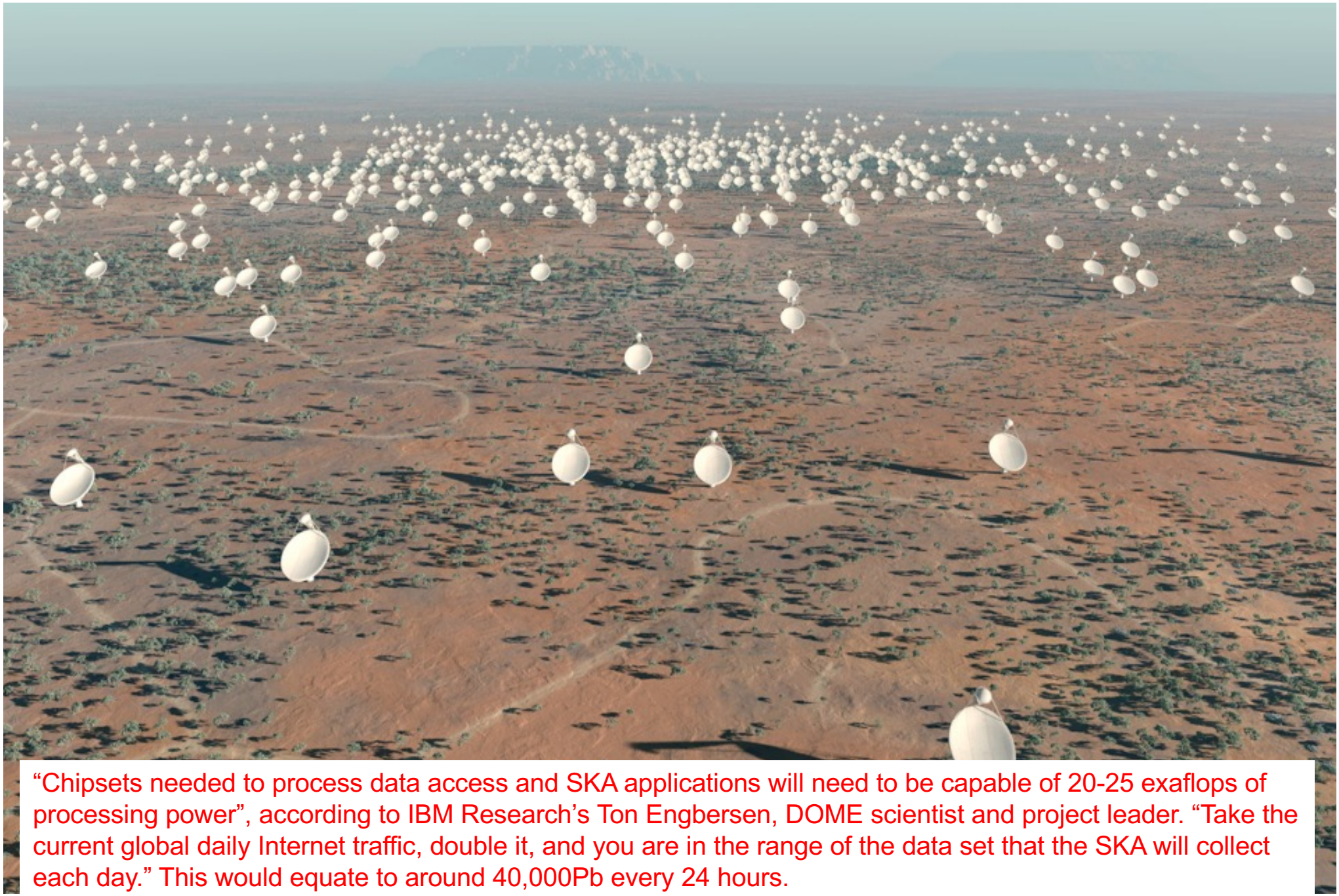
Tier 4

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

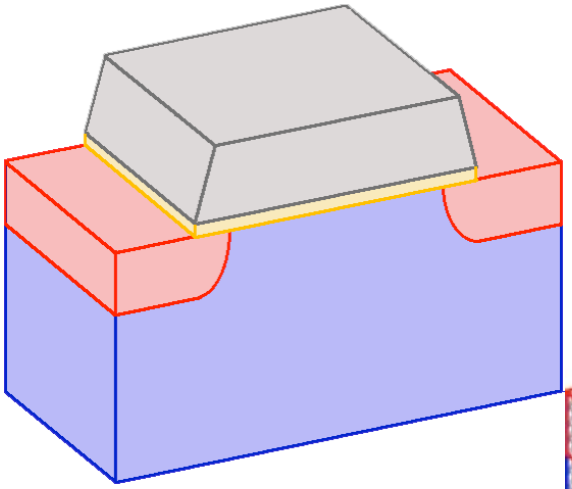
11 PDRAs (7 + 4)
7 PhD

Industrial partners

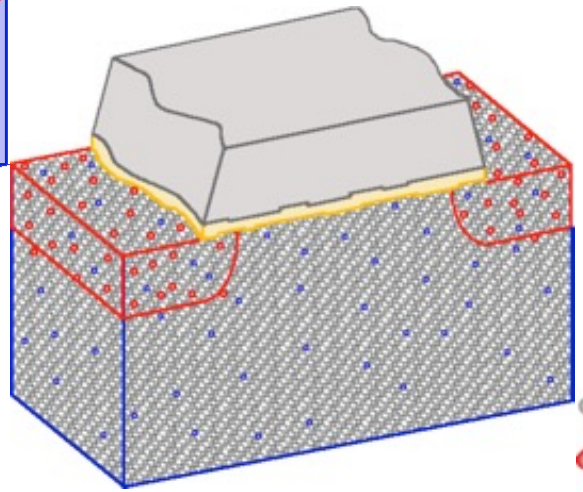
ARM, wolfson microelectronics, synopsys, freescale semiconductor, National Semiconductor, FUJITSU

University partners

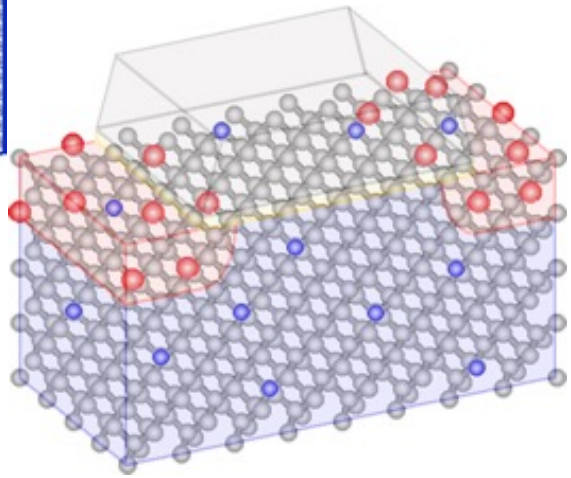
Meeting the Design Challenges of Nano-CMOS Electronics



Historic simulation paradigm

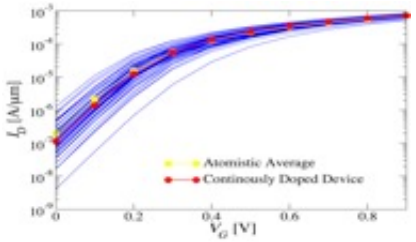


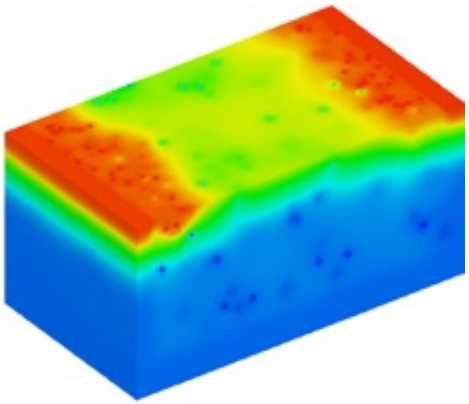
A 22 nm MOSFET
In production 2011

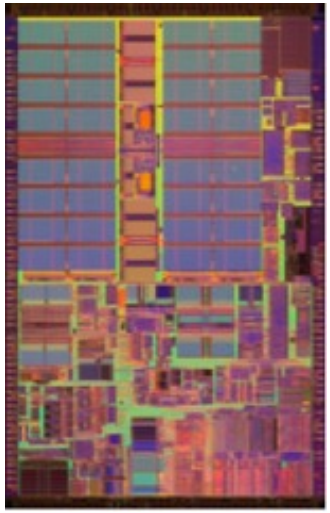


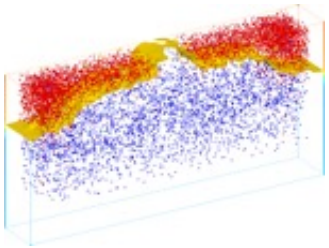
A 4.2 nm MOSFET
In production 2023

Challenges of NanoCMOS Design





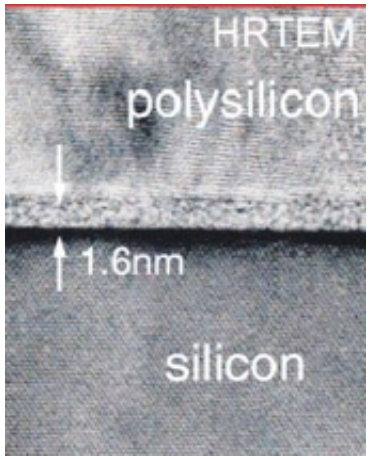


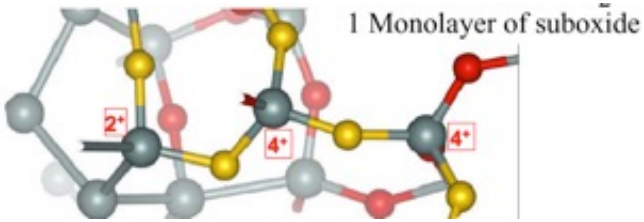


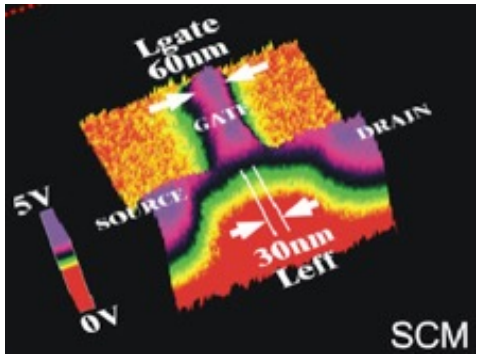
3D

+

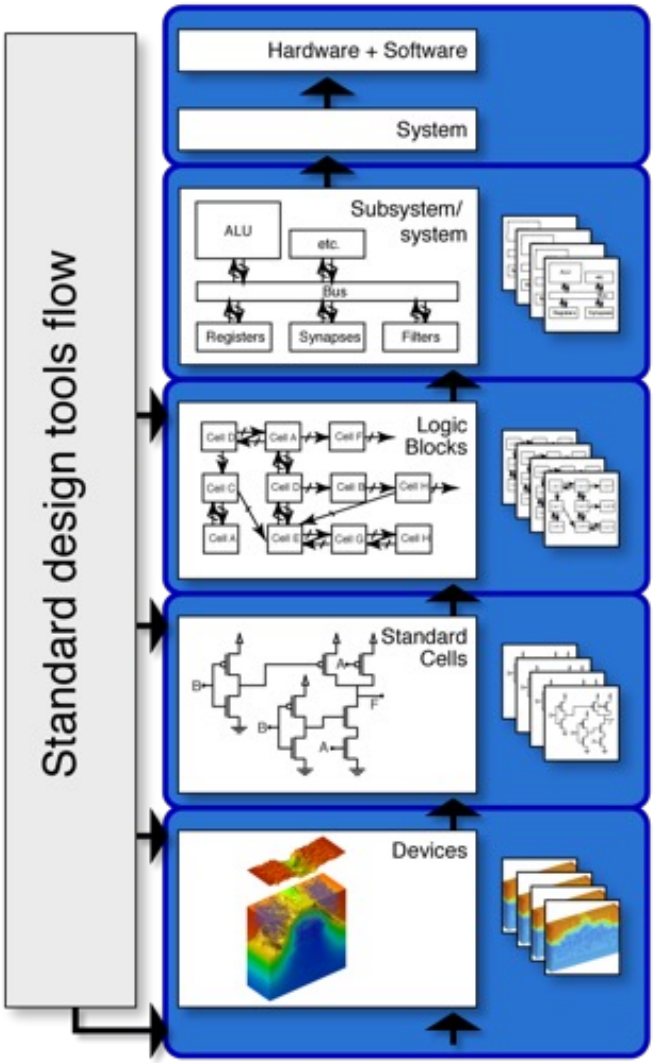
Statistical







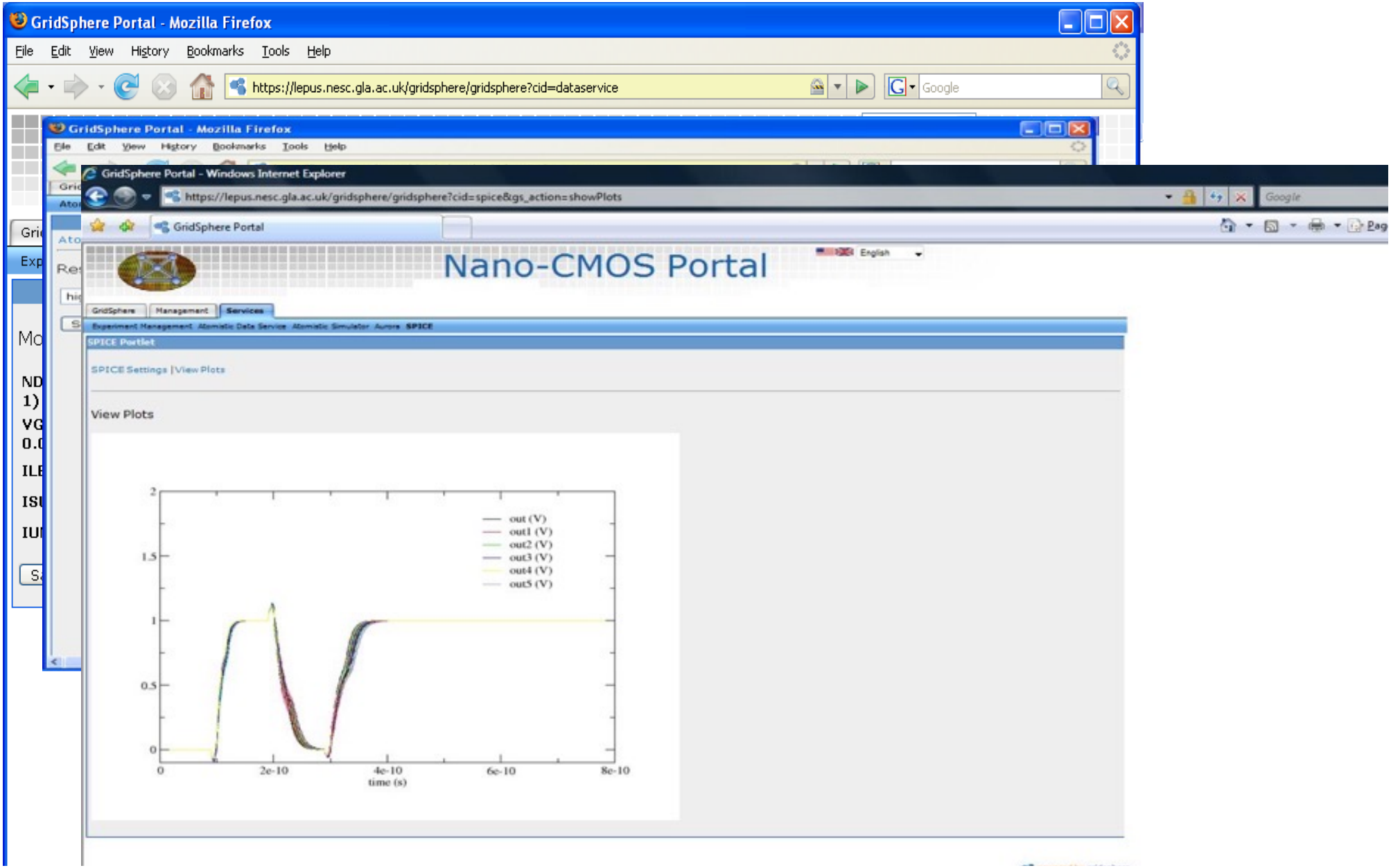
Challenges of Hierarchical statistical system simulations



- 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-Xperiences

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



But ended up with…

… a command-line based solution

This community

Security solution

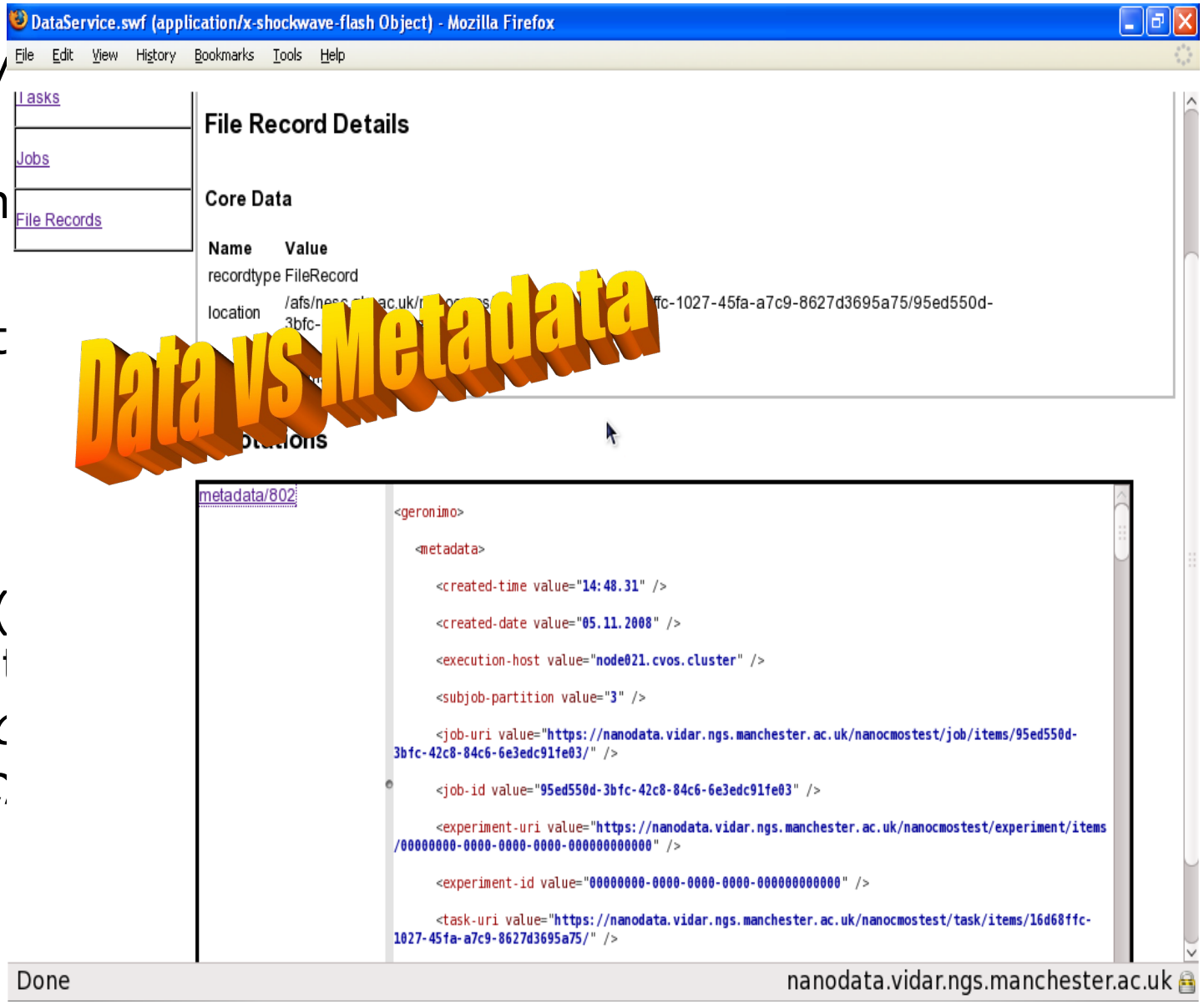
Secure, distribut

Meta-data captu

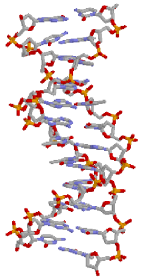
Job submission (
massive (at the

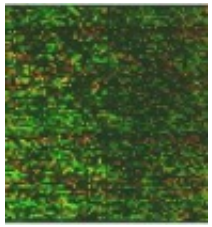
- *ScotGrid, NC*
- *Millions of C*

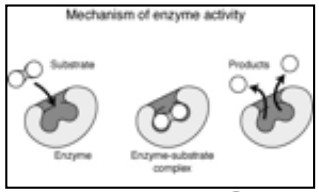
The –g flag!!!

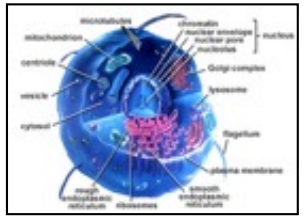


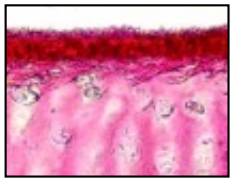
The e-Health Future…

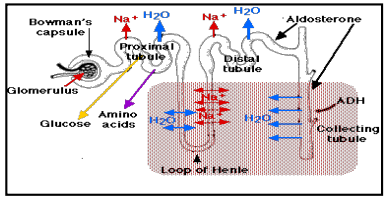












Nucleotide sequences

Nucleotide structures

Gene expressions

Protein Structures

Protein functions

Protein-protein interaction (pathways)

Cell

Cell signalling

Tissues

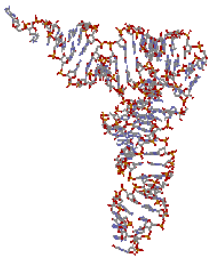
Organs

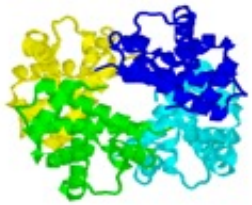
Physiology

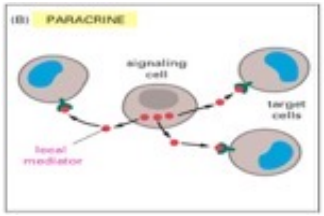
Organisms

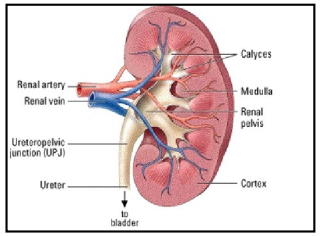
Populations

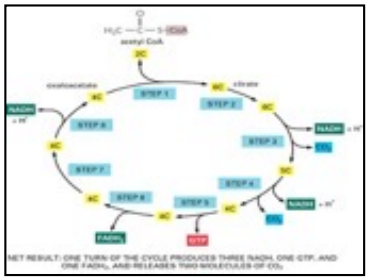
Security











+ 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

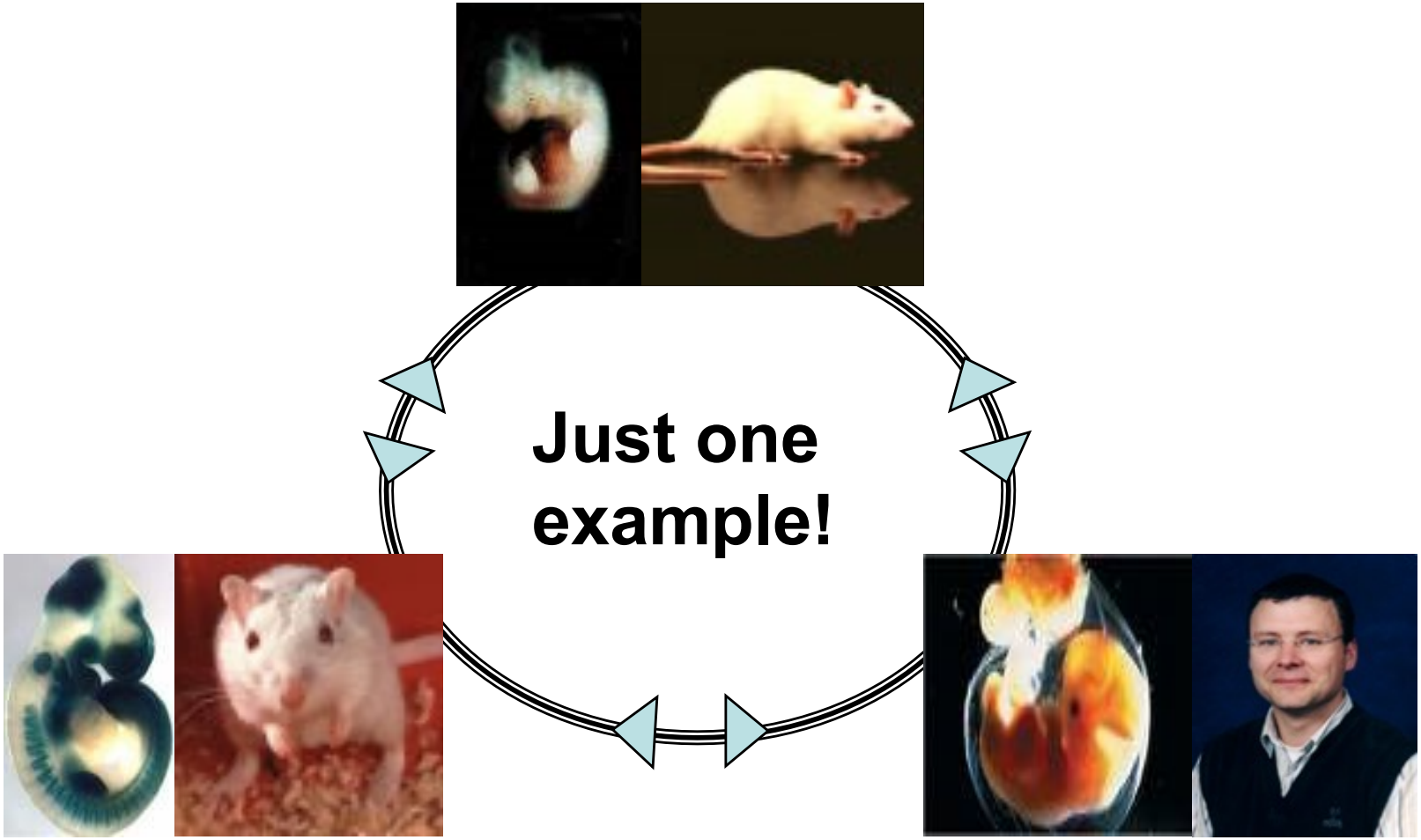
Messy

GPL96 - Notepad

File Edit Format View Help

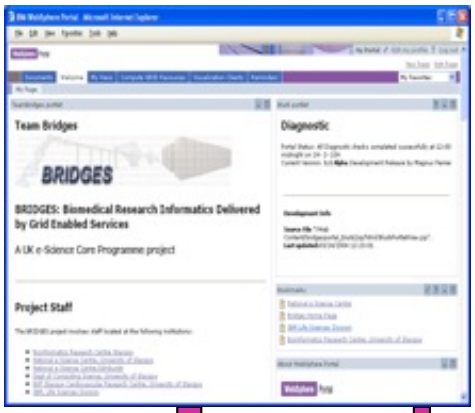

```
!Annotation!Annotation_date = 09/20/2006 15:35:01!Annotation_platform = GPL96!Annotation_platform_title = Affymetrix GeneChip Human Genome U
n replication factor C, 40-kDa subunit (A1) mRNA, complete cds  1590810 M87338  RFC2                    117_at heat shock 70kDa protein 6 (HSP
5                    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   3387325                         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                             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

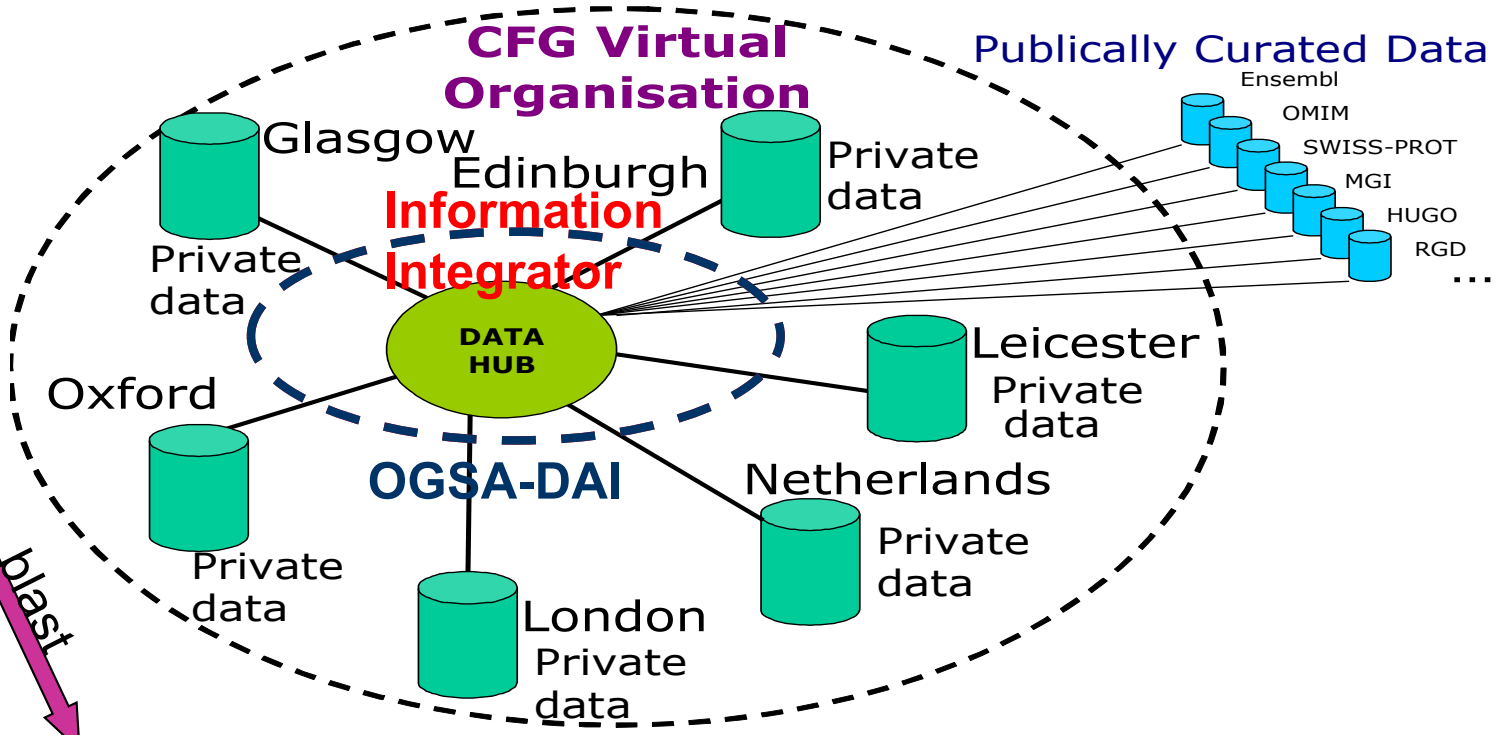
VO Authorisation



Synteny Service

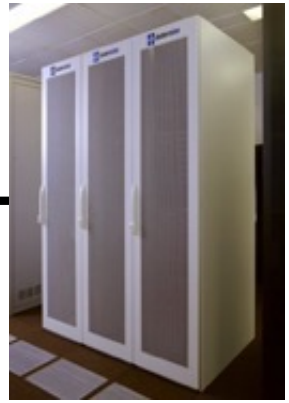
Magna Vista Service

blast

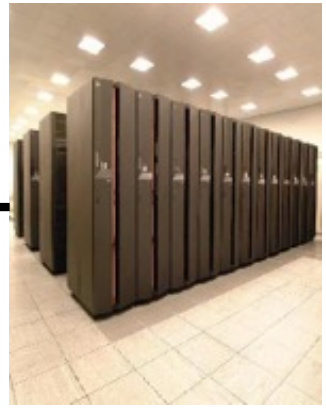


+

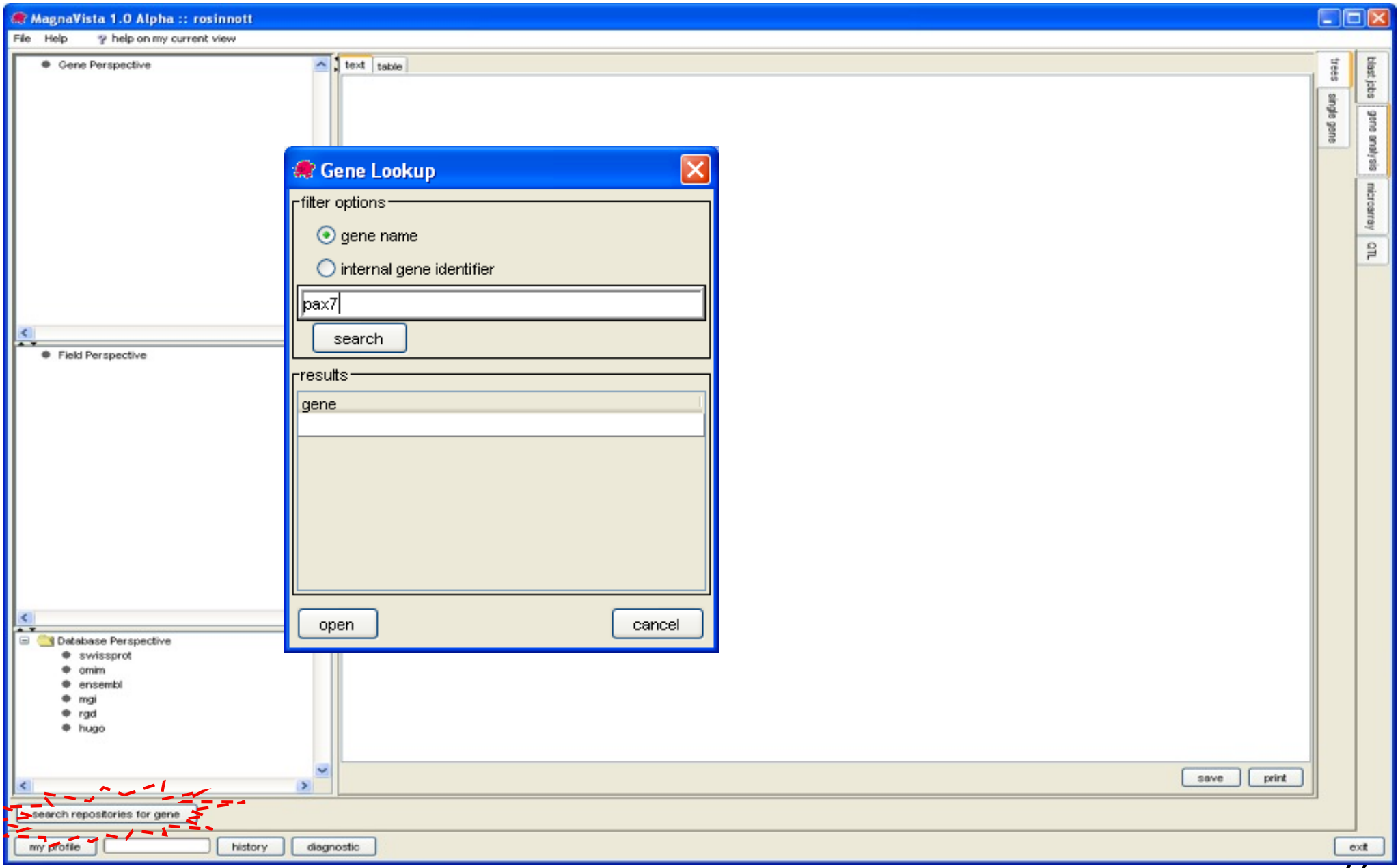
+



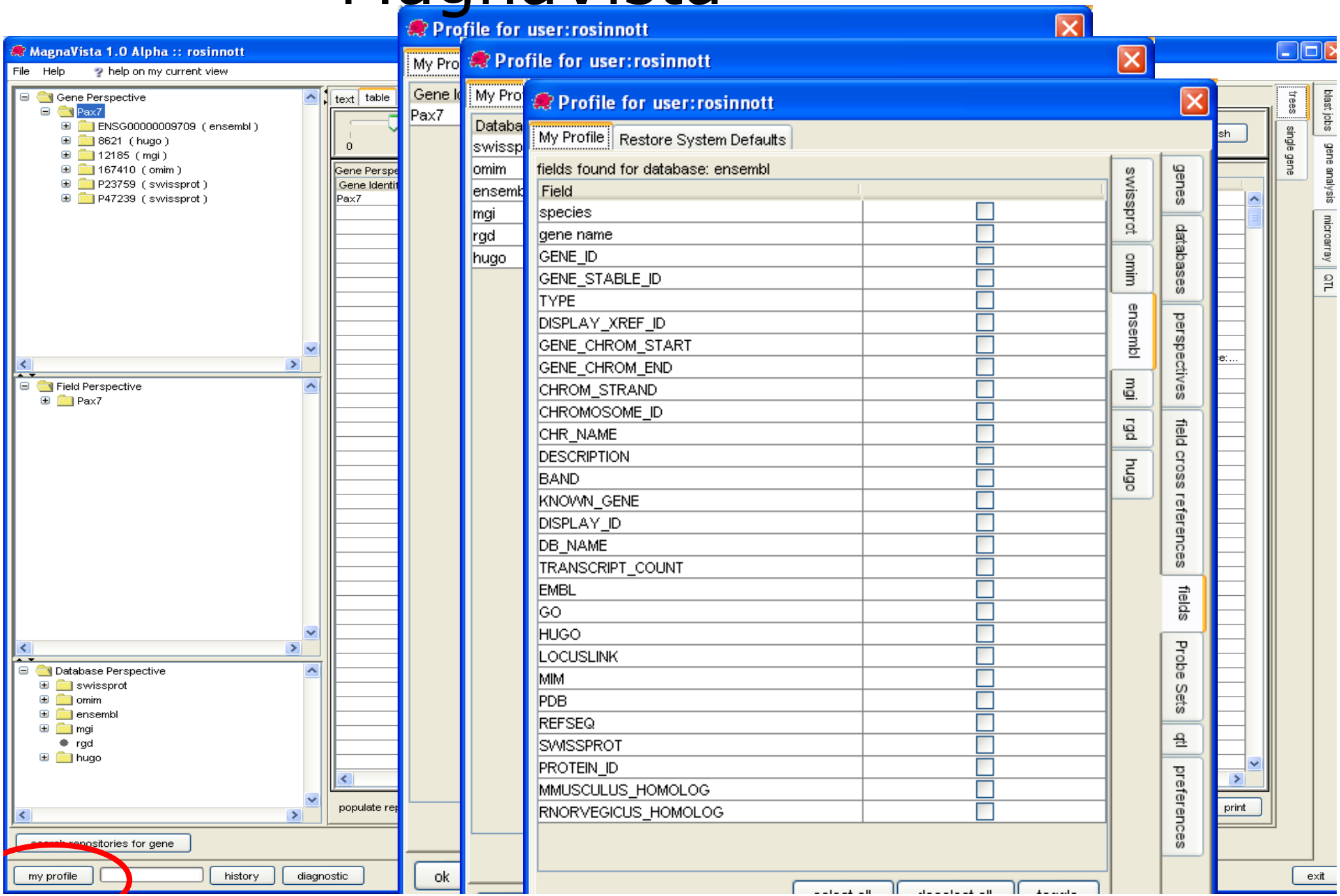
+



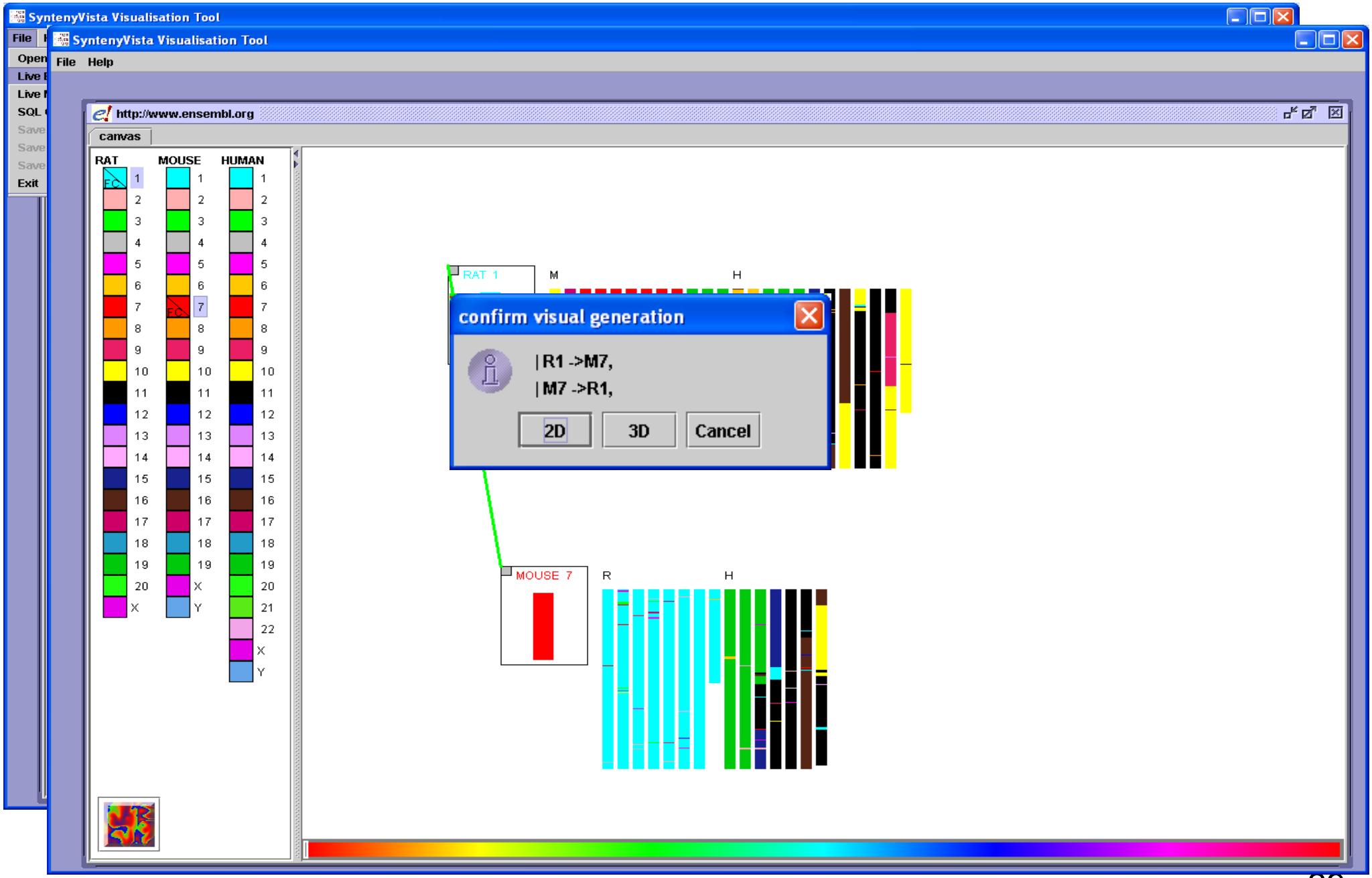
MagnaVista

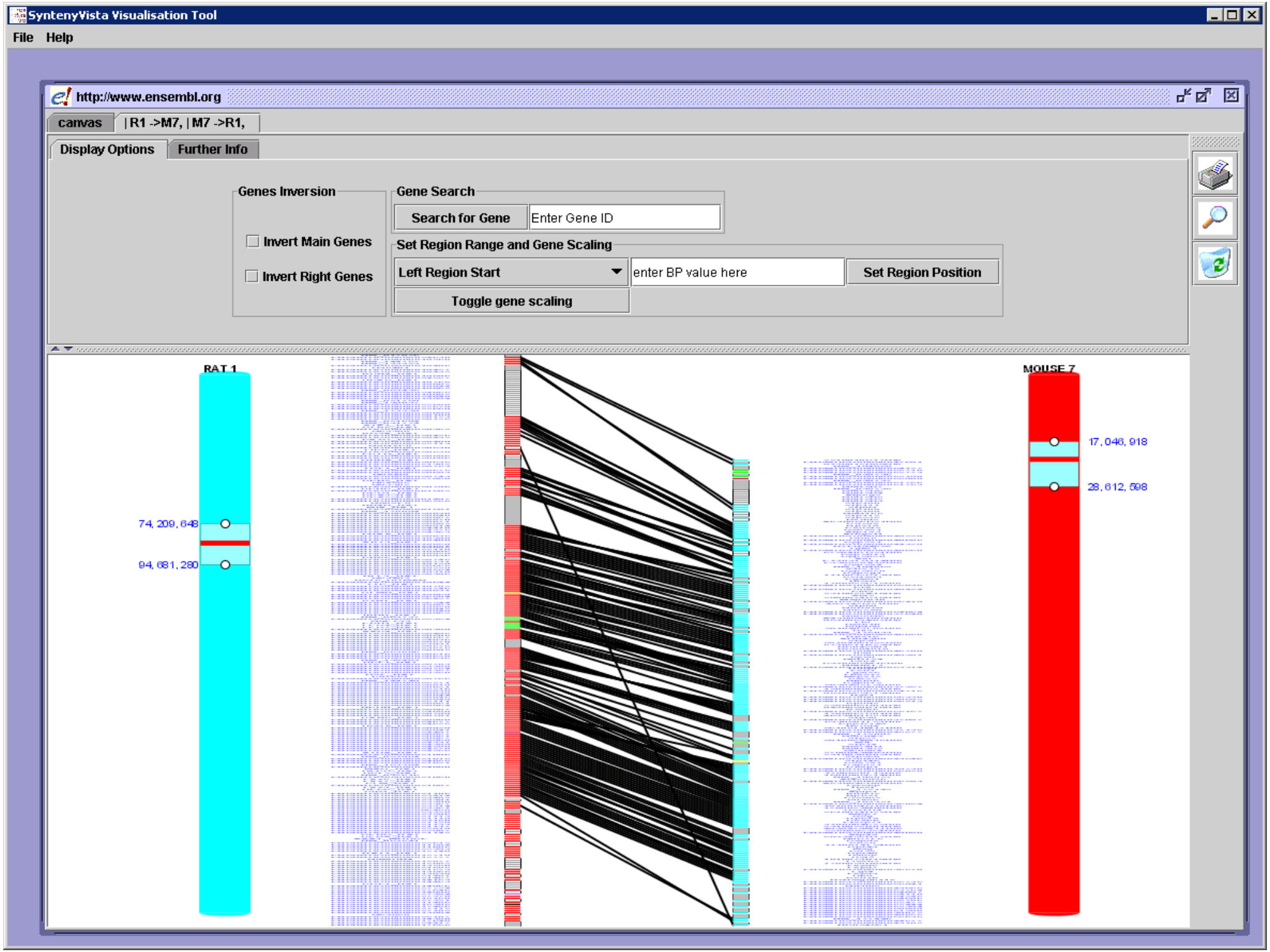

MagnaVista 1.0 Alpha :: rosinnott
File
Help
help on my current view
Gene Perspective
text
table
trees
single gene
blast jobs
gene analysis
microarray
QTL
Gene Lookup
filter options
gene name
internal gene identifier
pax7
search
results
gene
open
cancel
Field Perspective
Database Perspective
swissprot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

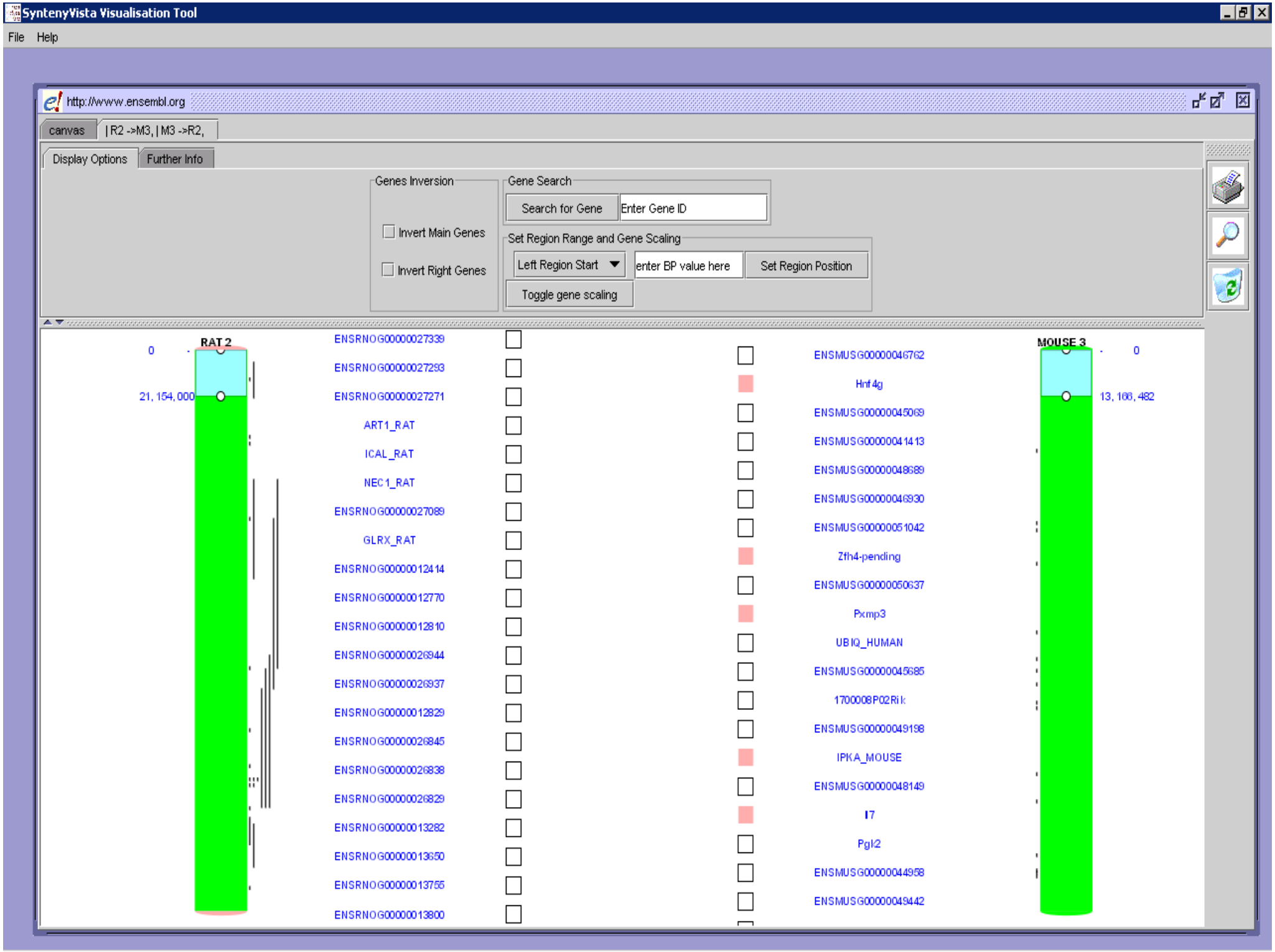
MagnaVista


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

Supporting Research in the Clinical/Biomedical Domain

- Different researchers/communities have different research computing needs
  - Based on my experiences
    - Data and data management
    - Inter-organisational collaborations
    - Security
    - Analytics, e.g. machine learning
    - Visualisation
      - *approximately in that order!*
  - Key to many research endeavours is Ethics
    - Governance, Trust, Policy, Procedures, GDPR, …
      - All essential before any tech solutions applied!!!
  - Importance of relationships!!!
    - Trust, engagement, delivery, …

A word about data…

- Clinical/biomedical data is…
  - almost always different across sites,
  - inconsistent across sites,
  - captured at different times/for different reasons,
  - hosted on different systems,
  - uses different databases/data technologies,
  - uses different coding schemes (or none!),
  - has different end users in mind,
  - has different security/ethics/governance arrangements…

Such challenges must be overcome!
Rapid prototyping, community data agreements
=> gradual hardening of secure systems
for use in clinical/biomedical **research**

Biomedical Data Sharing Models

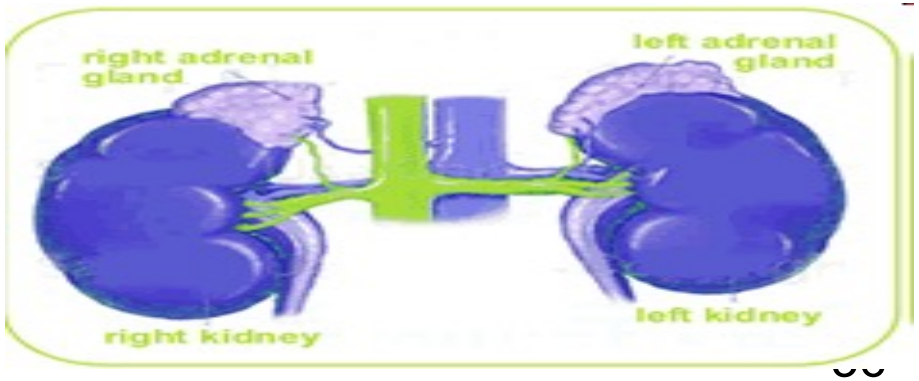
- Many models exist in the biomedical domain
  - There is no single solution that always applies
- Numerous models defined and supported
  - High-level architectural classification
    - *Dependent*, *open*-distributed, data sharing models;
      - direct distributed/federated querying (BRIDGES);
    - *Independent* centralised data sharing models;
      - Independent of existing hospital systems
      - Easiest to achieve (ENSAT-CANCER, …)
    - *Dependent push-based* (out-going) data sharing models;
      - Re-use existing medical data
      - Achievable but sustainable? (ADDN, …)
    - *Dependent* security-oriented *pull-based* data sharing models;
      - (asynchronous) data access
    - *Dependent* security-oriented *push-based* (incoming) clinical data sharing models
      - The major challenge (on many fronts, e.g. Optus/Medibank!!!)

Case Study Adrenal Cancer

(*Independent centralised data sharing model*)

- **European Network for Study of Adrenal Tumours (ENSAT – www.ensat.org)**
  - 6m Euro; 5-year project;
    - Started 2011; officially ended 2016, but…
    - Funded one developer
      - Initial prototype in 6 weeks!
  - Focus on 4 types of adrenal 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.)
  - *Imaging…*
  - *Biobanking…*
  - *Clinical Trials…*



(Clinical) Data Aggregation

Aka how to improve data quality…

	ACC	Pheo	NAPACA	APA	Total
Records	4638	4986	7949	3983	**22046**
Biosamples	8562	6108	11009	13270	**38949**
Clinical Annotations	58760	27367	22169	13930	**122226**
Annotations Per Patient (Mean)	12.66	5.48	2.78	3.49	**5.54**
Biosamples Per Patient (Mean)	1.84	1.22	1.38	3.33	**1.76**
Active Centers	72	82	62	41	**121**

Associated Study and Registry Distribution

	ACC	Pheo	NAPACA	APA	Total	Principal Investigator	Study Protocols	Study sites/e
EURINE-ACT	569	8	2130	237	2945	Wiebke Arlt	EURINE-ACT	
Ki-67	51	0	1	0	52	Martin Fassnacht		
Stage III/IV ACC	41	0	0	0	41	Eric Baudin		
PMT	0	259	162	2	423	Graeme Eisenhofer	PM	PMT
TMA	8	188	22	0	218	Ronald Krijger		
Long-term PHPGL	0	282	0	0	282	Pierre-Francois Plouin		
AVIS-2	0	0	0	2	2			
PMT3	0	12	0	0	12	Graeme Eisenhofer		
ADIUVO	64	0	0	0	64	Massimo Terzolo		ADIUVO
ADIUVO Observational	111	0	0	0	111	Massimo Terzolo		
HairCo-2	9	0	1	0	10	Marcus Quinkler		
FIRST-MAPPP	0	53	0	0	53	Eric Baudin		FIRST-MAPPP
German Cushing Registry	0	1	104	0	105	Martin Reincke		
German Conn Registry	0	1	0	1151	1154	Martin Reincke		German Conn Registry

Pheo PMT Study

	Full Study	Phase 1 (Screening)	Phase 2 (Clonidine)	Phase 3 (Pheo Characterization)	Phase 4 A (Excluded Follow-Up)	Phase 4 B (Pheo Follow-Up)
Patients	2238	454	17	6	1520	241

ENSAT-CANCER

ENS@T Registry

ENSAT Home | ACC Home | ACC Search | ACC Exported Data |

ACC Search Results

There are **60** records matching the following query:

Parameter	Condition	
year_of_birth	year_of_birth >= 1940 AND year_of_birth <= 1950	☑
sex	M	☑

Repeat Search

Export these results

Export all your patient data

Run a new search

ENSAT ID	Referral Doctor	Record Date	Date of First Registration	Sex	Year of Birth	Consent Level Obtained
FRPA2-4	Jerome Bertherat (jerome.bertherat@cch.aphp.fr)	16 Apr 2007	16 Apr 2007	M	1944	National (France)
FRPA2-51	Jerome Bertherat (jerome.bertherat@cch.aphp.fr)	09 Apr 2003	09 Apr 2003	M	1950	National (France)
FRPA3-20	Eric Baudin (eric.baudin@igr.fr)	14 Dec 1989	14 Dec 1989	M	1946	National (France)
FRPA3-24	Eric Baudin (eric.baudin@igr.fr)	03 May 2000	03 May 2000	M	1947	National (France)
GYWU-671	Martin Fassnacht (fassnacht_m@medizin.uni-wuerzburg.de)	19 Apr 2011	26 Jan 2011	M	1940	National (Germany)
GYWU-662	Martin Fassnacht (fassnacht_m@medizin.uni-wuerzburg.de)	19 Apr 2011	19 Jan 2011	M	1948	National (Germany)
ITFL-3	Massimo Mannelli (m.mannelli@DFC.UNIFI.IT)	13 Dec 2010	14 Mar 2006	M	1947	Local

Associated Record Search

Surgery

Pathology

Biomaterial

Mitotane

Chemotherapy

Radiofrequency

Radiotherapy

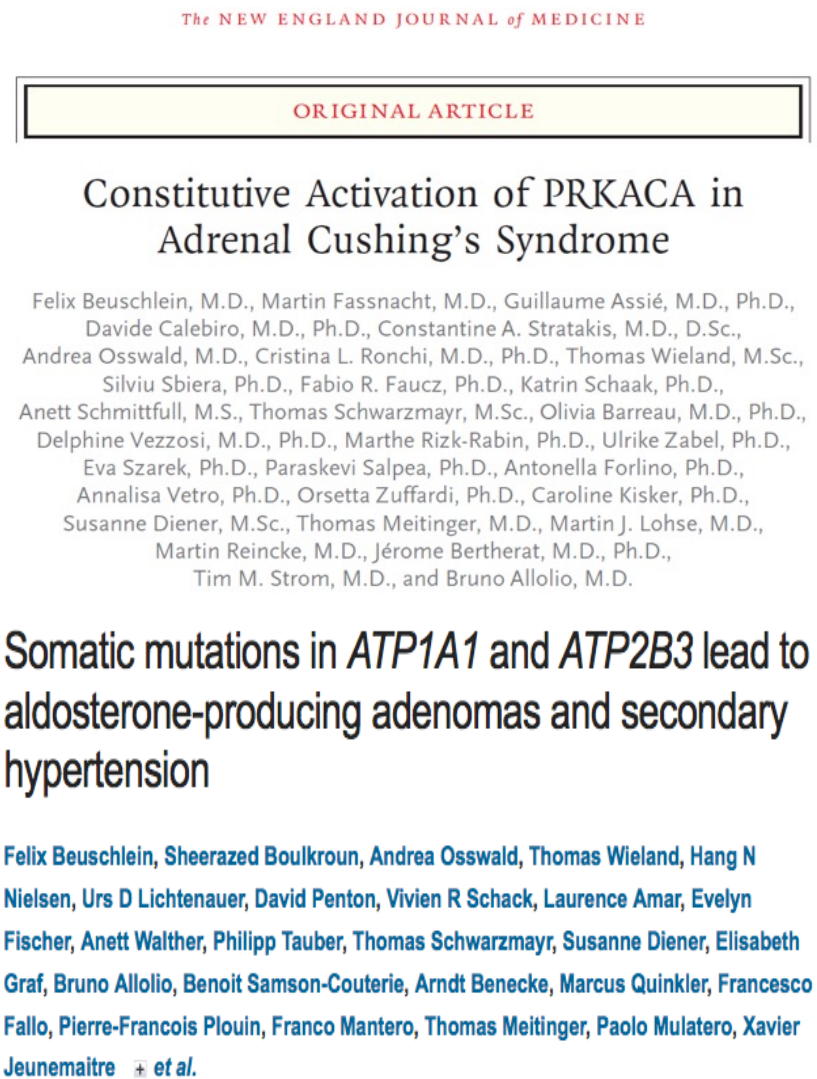
Chemoembolisation

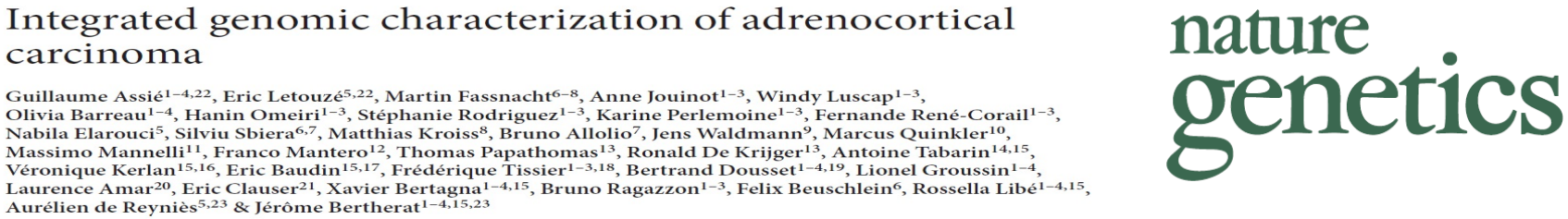
Follow-Up

ACC Home

(Research) Impact

- Publications (100+)
  - Nature Genetics (multiple)
  - New England Journal of Medicine (multiple)
  - Journal of Clinical Endocrinology and Metabolism
  - Journal of Nuclear Medicine
  - Journal of Endocrinology
  - Journal of Human Molecular Genetics
  - International Journal of Cancer
  - Journal of Molecular and Cellular Endocrinology
  - European Journal of Endocrinology
  - Journal of Cancer Cell
  - Journal of Clinical Cancer Research
  - International Journal of Clinical Chemistry
  - Journal of Hormone and Metabolic Research
  - Journal of Modern Pathology
  - Journal of Hypertension
  - European Journal of Cancer
  - Journal of Annals of Clinical Biochemistry
  - Journal of Endocrine-related cancer
  - Journal of the National Cancer Institute
  - American Journal of Cancer Research
  - ...





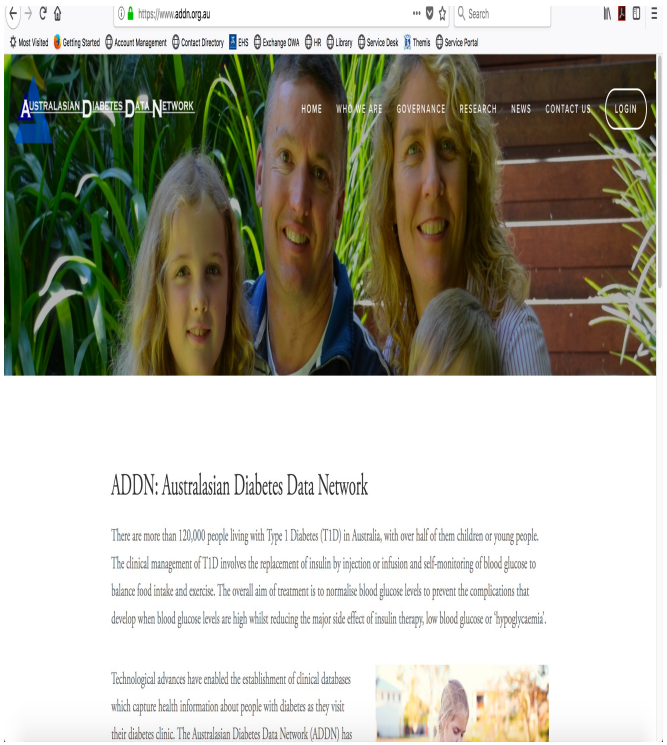
Biomedical Data Sharing Models

- Many models exist in the biomedical domain
  - There is no single solution that always applies
- Numerous models defined and supported
  - High-level architectural classification
    - *Dependent*, *open*-distributed, data sharing models;
      - direct distributed/federated querying (BRIDGES);
    - *Independent* centralised data sharing models;
      - Independent of existing hospital systems
      - Easiest to achieve (ENSAT-CANCER, …)
    - *Dependent push-based* (out-going) data sharing models;
      - Re-use existing medical data
      - Achievable but sustainable? (ADDN, …)
    - *Dependent* security-oriented *pull-based* data sharing models;
      - (asynchronous) data access
    - *Dependent* security-oriented *push-based* (incoming) clinical data sharing models
      - The major challenge (on many fronts!)

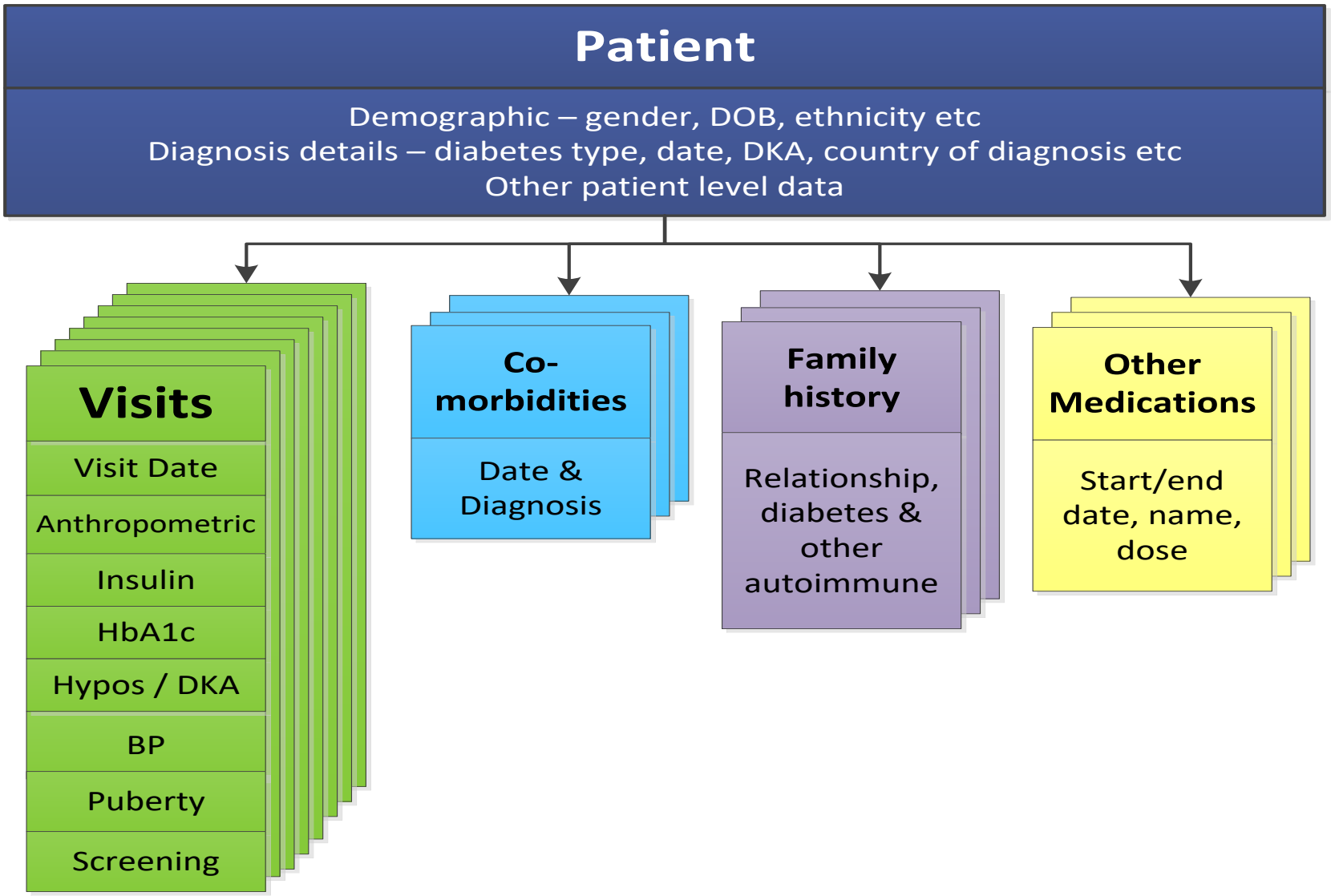
Case Study: Australasian Diabetes Data Network (ADDN)

(Dependent push-based (out-going) data sharing model)

- National, longitudinal database for adult/child patients with T1D
- Used to
  - Increase patient centred research
  - Support Australasia-wide collaboration
  - Benchmarking across diabetes centres to drive quality improvement
  - Inform policy development for diabetes management
- Direct data entry not feasible/viable
  - Currently
    - 12 adult sites, 23 pediatric sites
    - many other sites under discussion
      - Extremely heterogeneous systems

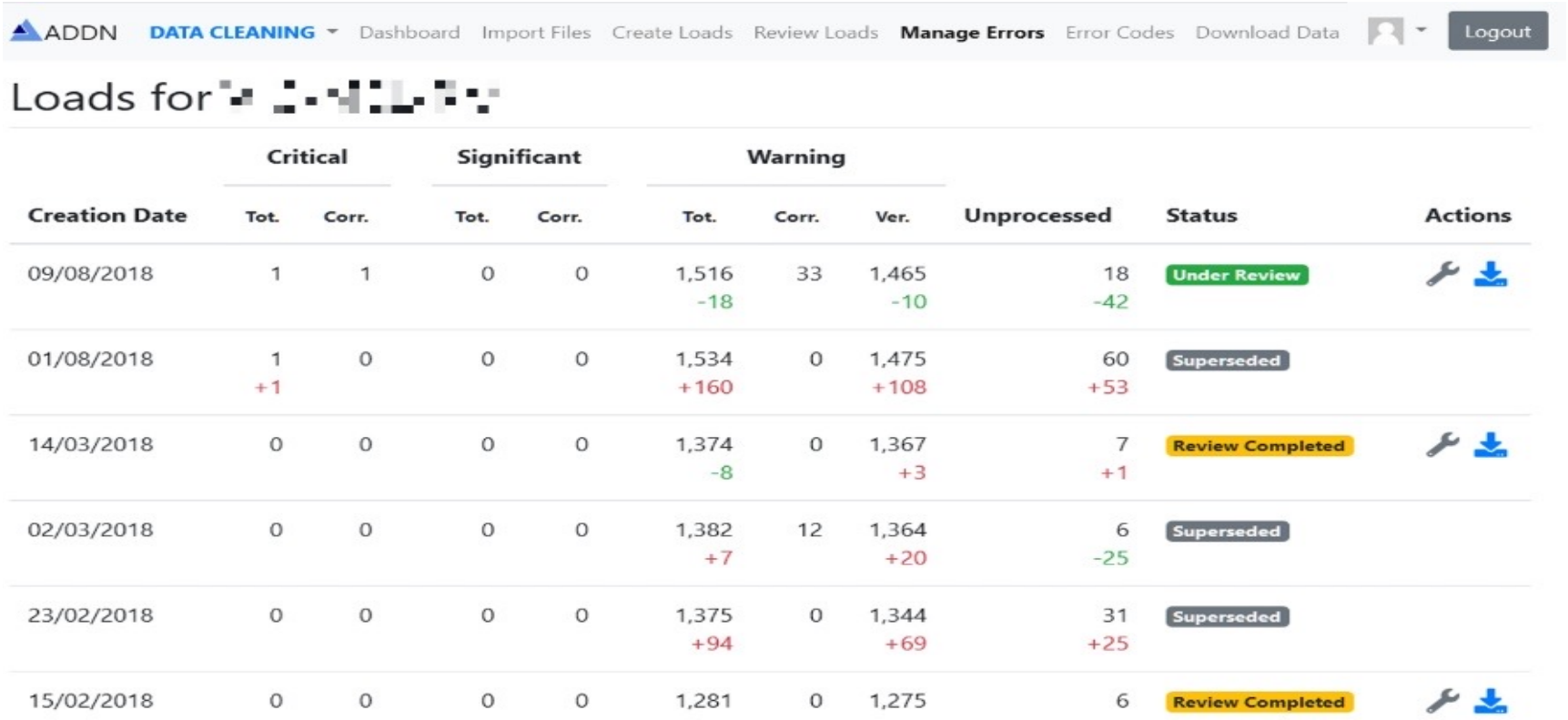


ADDN Dataset



ADDN Data Quality

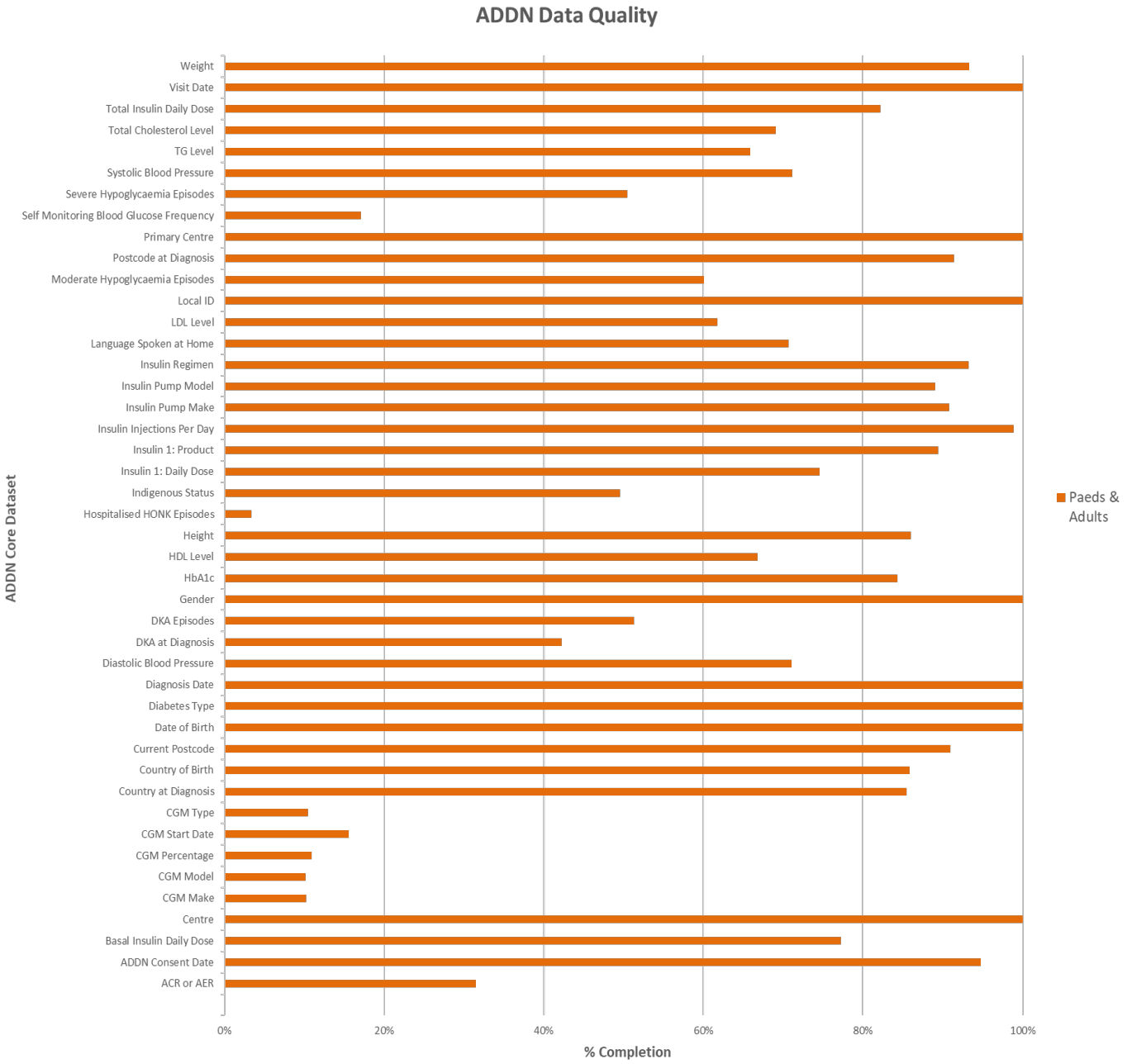
- Centres submit data twice yearly



	Critical		Significant		Warning					
Creation Date	Tot.	Corr.	Tot.	Corr.	Tot.	Corr.	Ver.	Unprocessed	Status	Actions
09/08/2018	1	1	0	0	1,516 -18	33	1,465 -10	18 -42	Under Review	
01/08/2018	1 +1	0	0	0	1,534 +160	0	1,475 +108	60 +53	Superseded	
14/03/2018	0	0	0	0	1,374 -8	0	1,367 +3	7 +1	Review Completed	
02/03/2018	0	0	0	0	1,382 +7	12	1,364 +20	6 -25	Superseded	
23/02/2018	0	0	0	0	1,375 +94	0	1,344 +69	31 +25	Superseded	
15/02/2018	0	0	0	0	1,281	0	1,275	6	Review Completed	

- *Clean* their data using on-line data cleaning tool
- Final dataset created – used for benchmarking reports, analysis and research data requests

Data Completeness

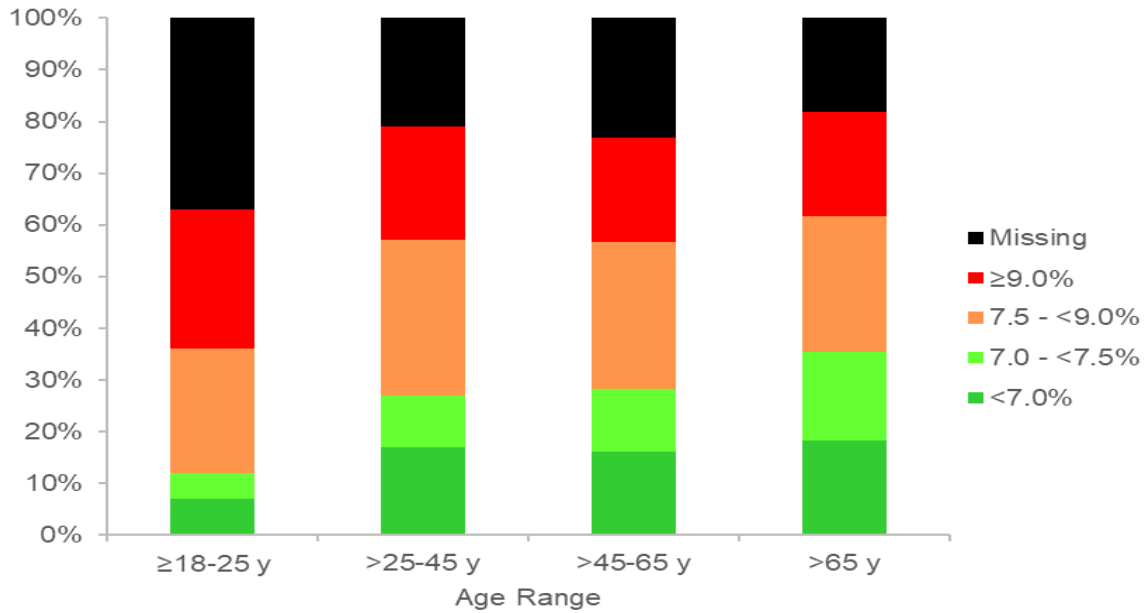


ADDN dataset =
131 data items

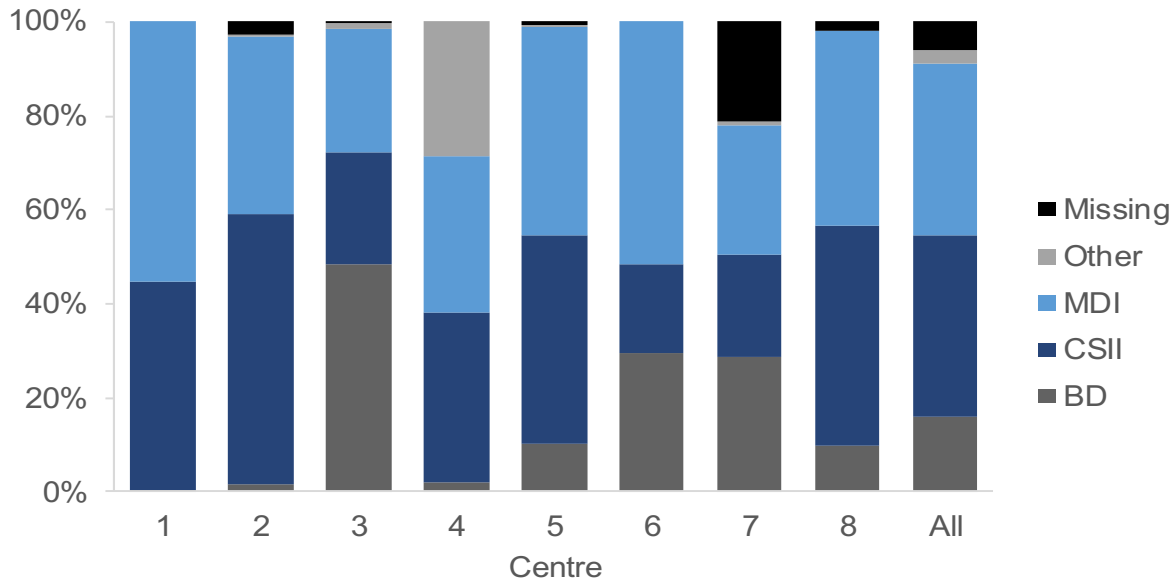
Core dataset =
44 data items

ADDN Example Analytics

Adult HBA1c
(average glycaemia over the preceding 6–8 weeks)



Paed Insulin Regimen
(Multiple Daily Injections, Continuous Subcutaneous Insulin Infusion, Bi-daily)



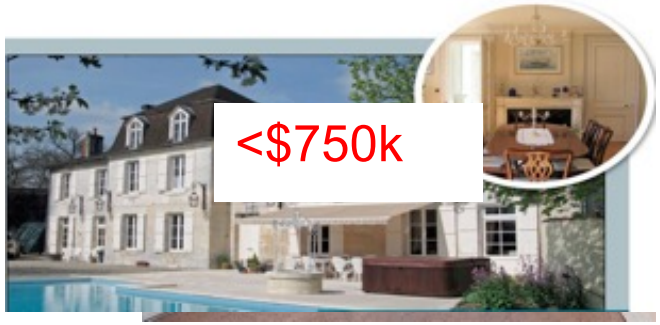
Supporting Urban Research

- Australian Urban Research Infrastructure Network (AURIN) federally funded project
  - ~$70m funding
  - www.aurin.org.au
    - University of Melbourne are lead agent
      - 12+ years of my life!

Australian Government
Department of Education and Training

- Establishing an e-Infrastructure for Urban and Built Environment Researchers
  - Distributed, (completely!) heterogeneous datasets
  - Data interrogation services
  - Security (unit level data, health data, com
  - Online analysis tools
  - Collaboration!!!

The Urban Context



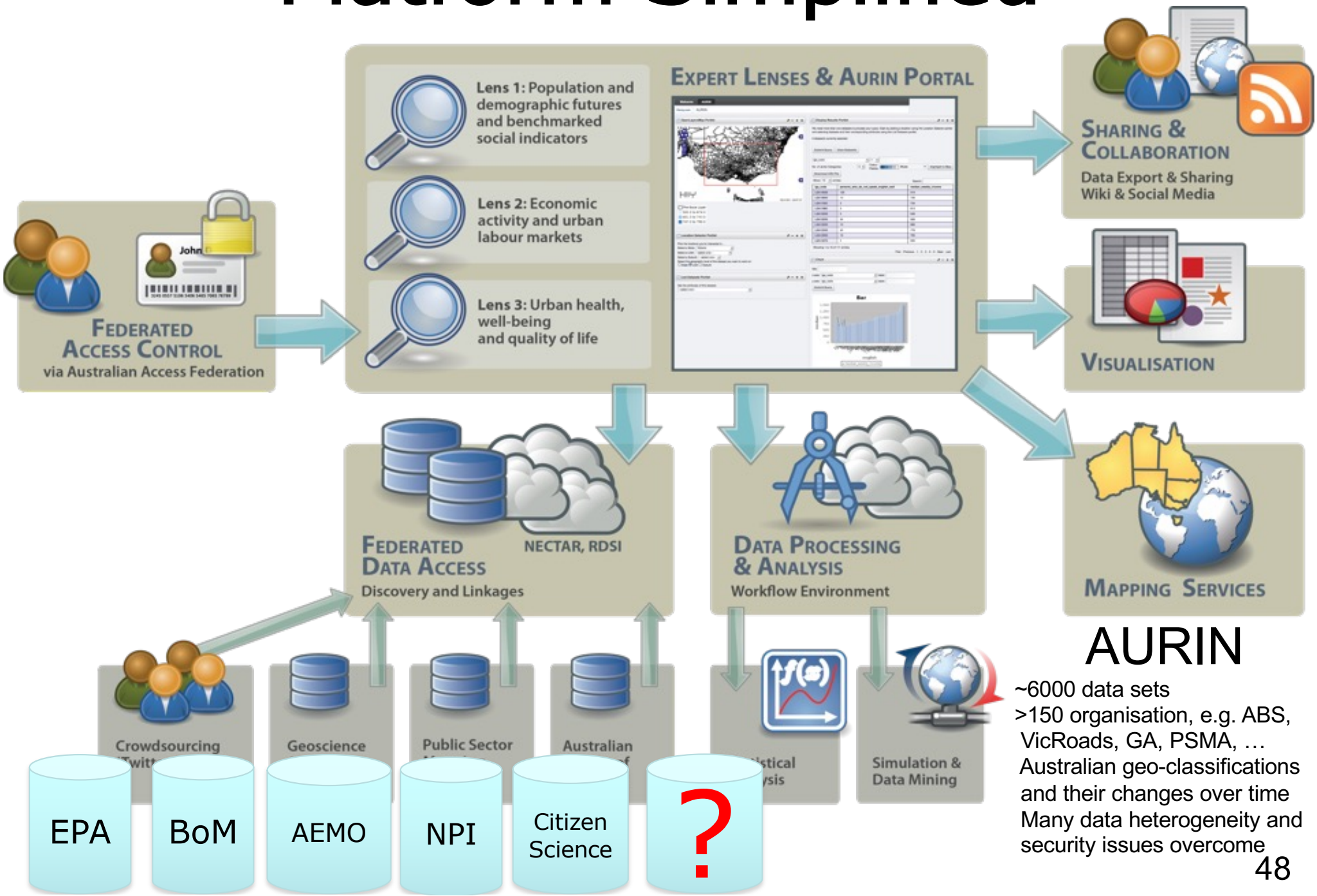
- Urban and build environment is extremely broad
  - health,
  - transport,
  - future population,
  - liveability,
  - crime,
  - housing,
  - design,
  - …




- 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)

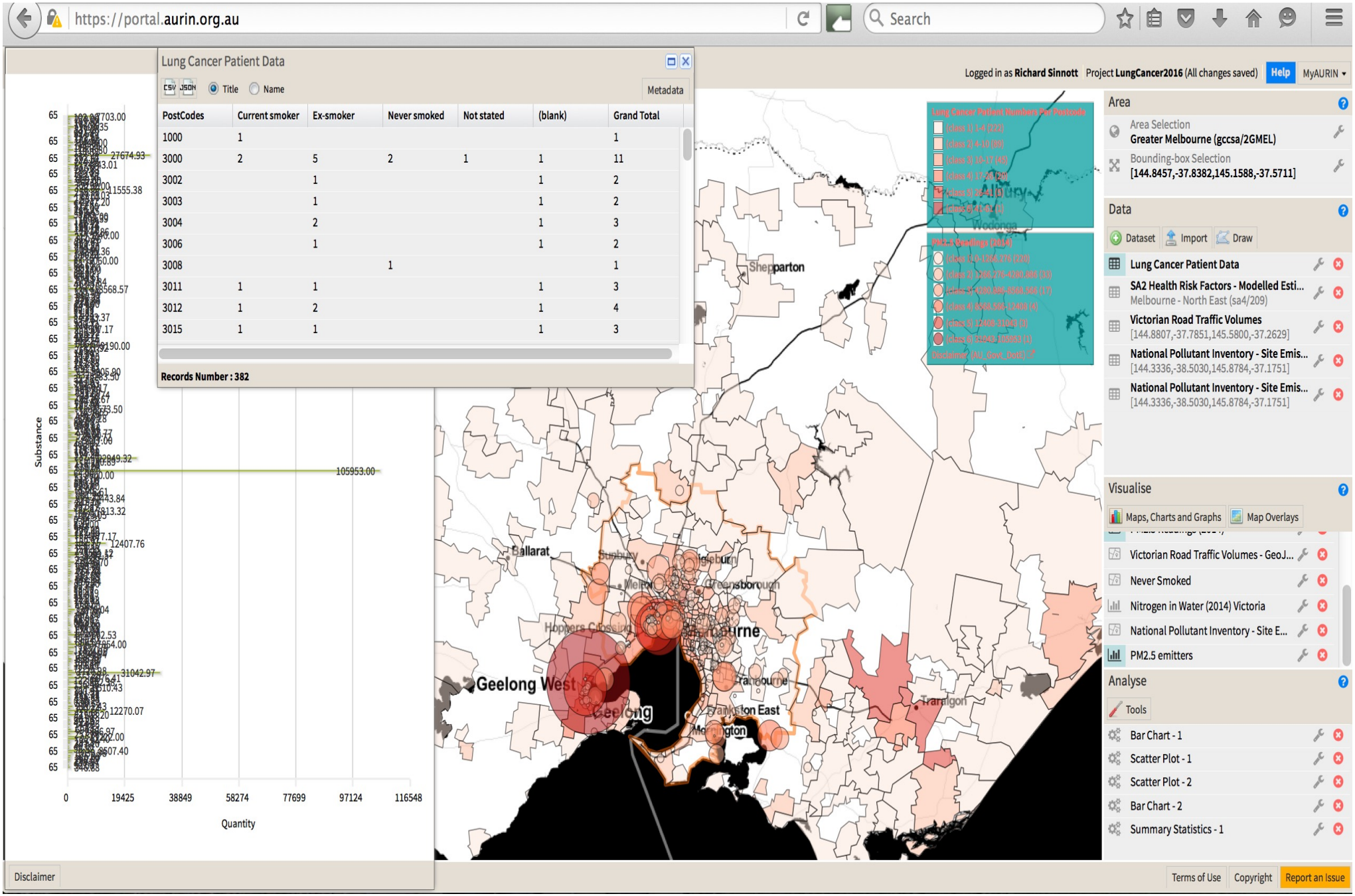
Platform Simplified



AURIN

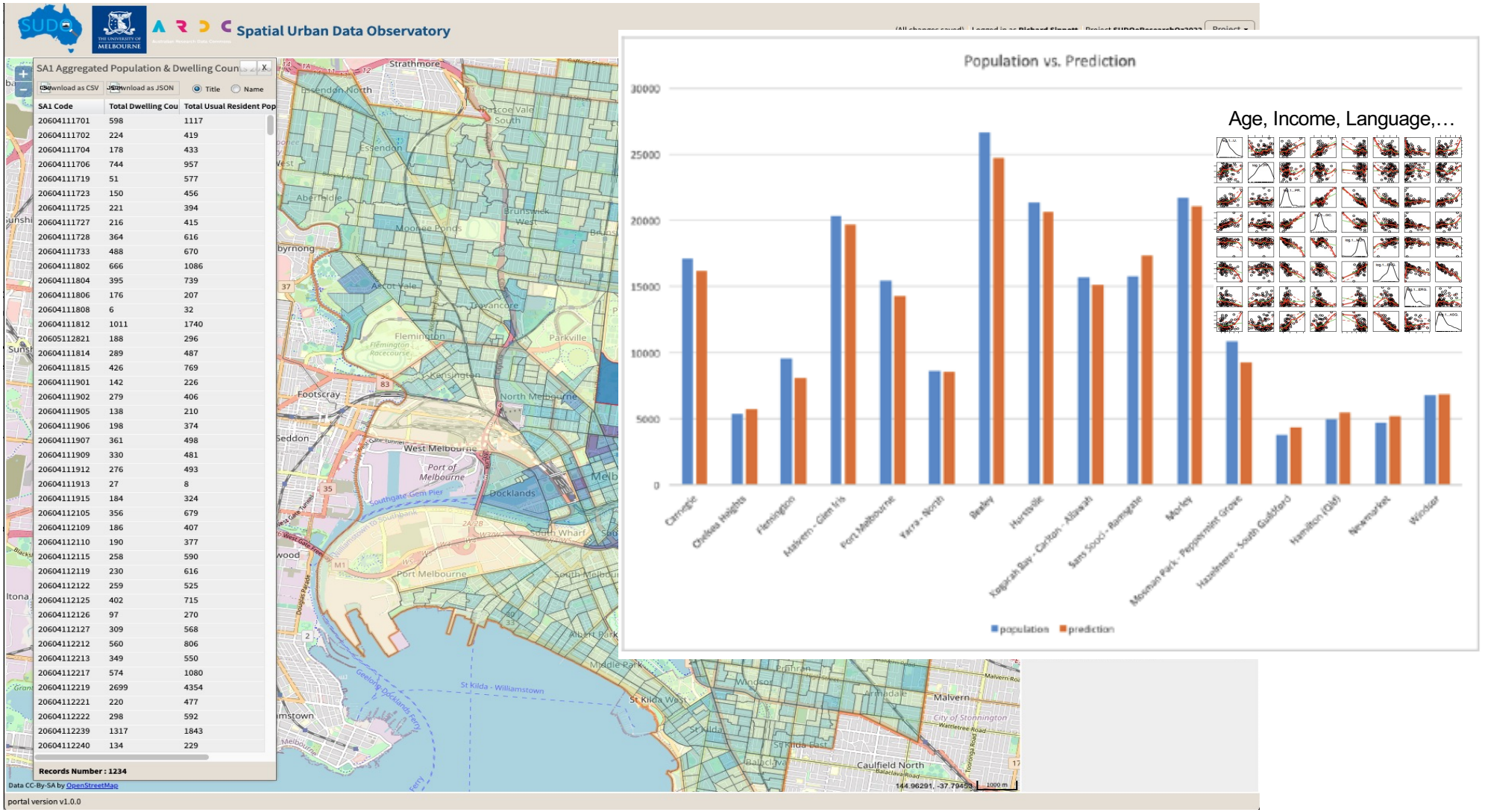
~6000 data sets
>150 organisation, e.g. ABS, VicRoads, GA, PSMA, …
Australian geo-classifications and their changes over time
Many data heterogeneity and security issues overcome

AURIN Example (in one slide!)



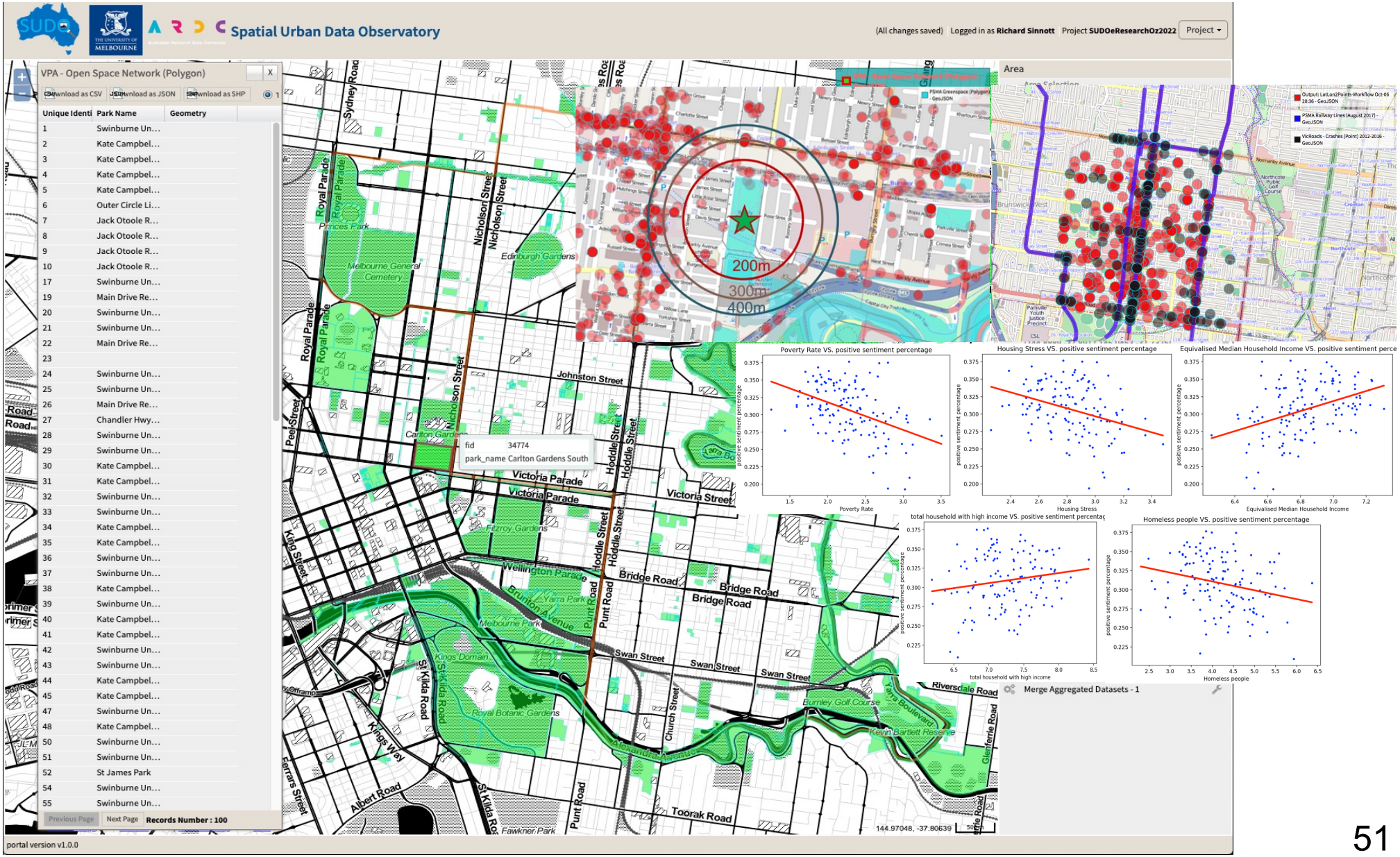
Population Estimation

(*student exemplars for ass2 inspiration*)



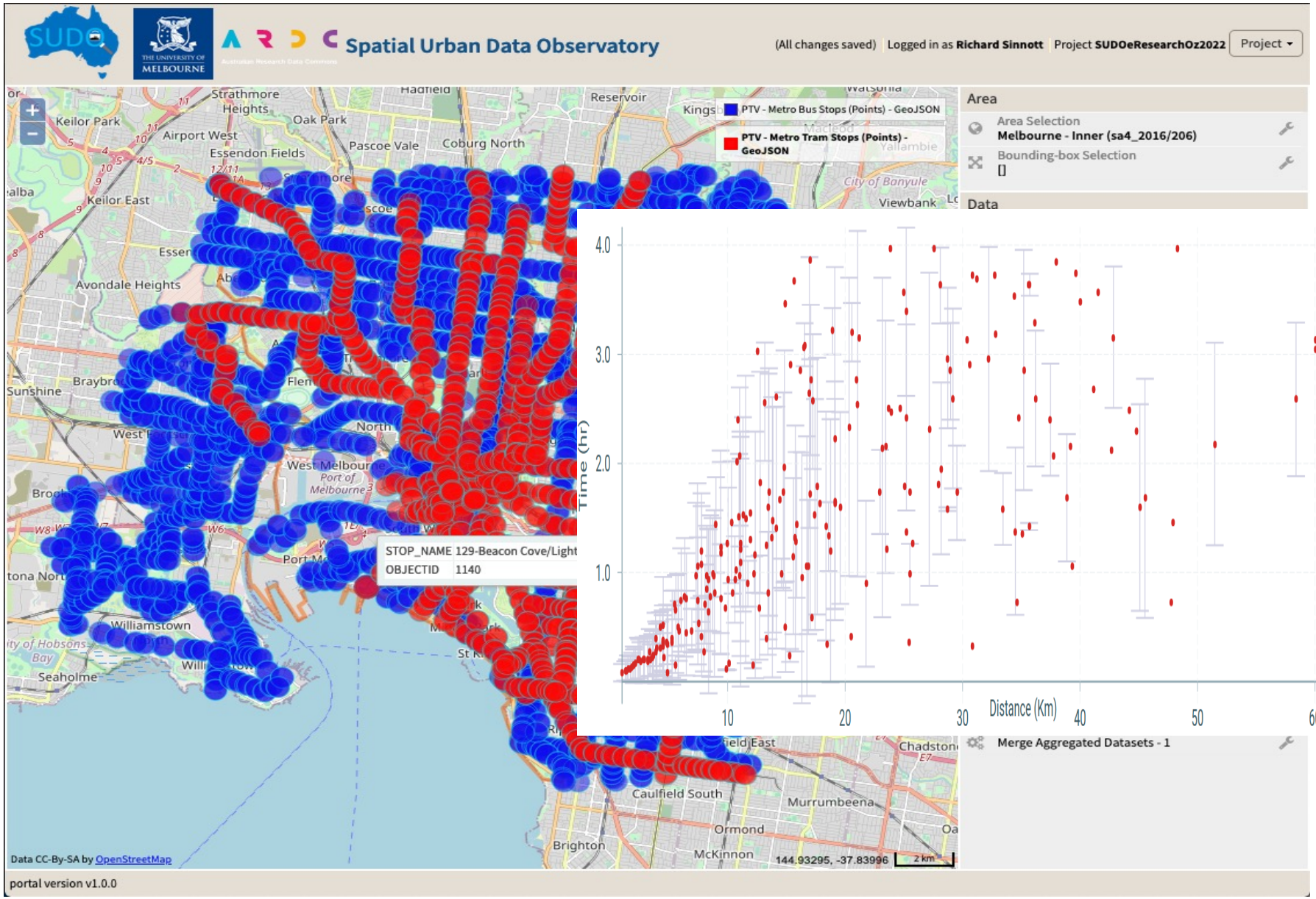
Green Space and Happiness

(*student exemplars for ass2 inspiration*)



Commuting Around Melbourne

(*student exemplars for ass2 inspiration*)



Demonstration of SUDO in Workshop

(Essential for Assignment II)

Homework

(https://sudo.eresearch.unimelb.edu.au)

Find out how many public toilets there are/were in Carlton
(not assessed, but useful for assignment 2!)

1. Log in to SUDO (select UniMelb)
2. Navigate to Australia::Victoria::SA4::SA3
3. Browse for Public Toilets (e.g., UQ_ERG)
4. Select any/all attributes
5. Select some *other* random data set at SA2 level, e.g., from the ABS
6. Spatialise the *other* data set
(Tools::Spatial Tools::Spatialise Aggregated Data Set and select the *other* data set) – adds the polygon to the data
7. Count toilets
(Tools::Spatial Tools::Count Points in Polygons using the output Spatialised Data Set and the data points from the Toilet Data Set)

Questions ...?