GeneLab Strategic Plan: Perspectives and Thoughts

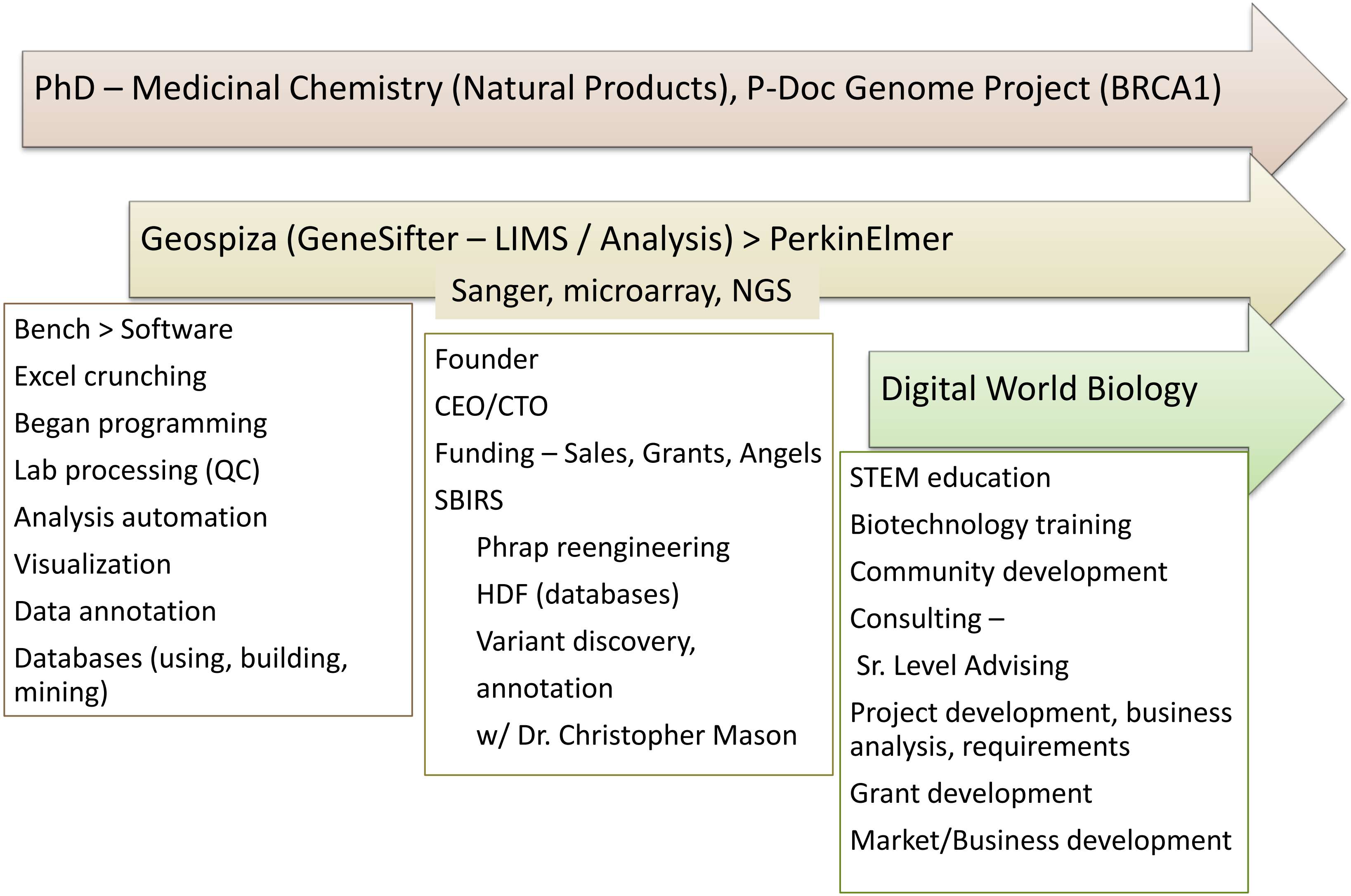
Todd M. Smith Ph.D. Digital World Biology

Committee on Biological and Physical Sciences in Space, Washington DC, April 1, 2015

Agenda

- All about me
- Software experiences
- Initial thoughts
- Perspectives as a provider and user

Profile



PhD – Medicinal Chemistry (Natural Products), P-Doc Genome Project (BRCA1)

Geospiza (GeneSifter – LIMS / Analysis) > PerkinElmer

Sanger, microarray, NGS

Bench > Software

Excel crunching

Began programming

Lab processing (QC)

Analysis automation

Visualization

Data annotation

Databases (using, building, mining)

Founder

CEO/CTO

Funding – Sales, Grants, Angels

SBIRS

- Phrap reengineering
- HDF (databases)
- Variant discovery,
- annotation
- w/ Dr. Christopher Mason

Digital World Biology

STEM education

Biotechnology training

Community development

Consulting –

Sr. Level Advising

Project development, business analysis, requirements

Grant development

Market/Business development

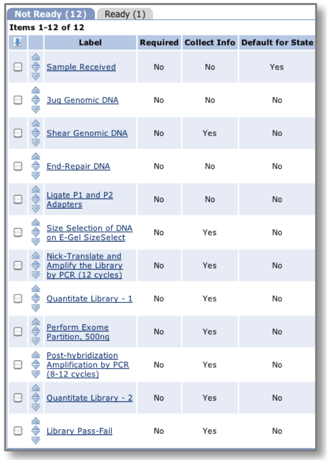
Software Experience

GeneSifter Lab Edition | GeneSifter Analysis Edition

RNA-Seq ChIP-Seq Me-Seq Var-Seq My-Seq

Application-specific Lab Workflows

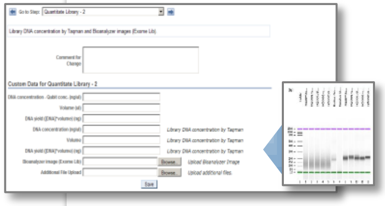
Kits
Best Practices

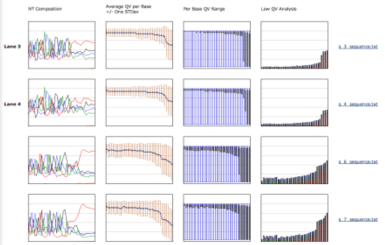


Not Ready (12) | Ready (1)

Items 1-12 of 12

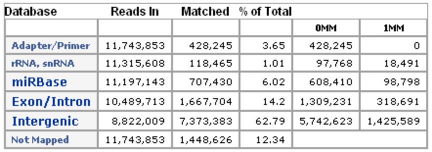
□		Label	Required	Collect Info	Default for State
□		Sample Received	No	No	Yes
□		3ug Genomic DNA	No	No	No
□		Shear Genomic DNA	No	Yes	No
□		End-Repair DNA	No	No	No
□		Ligate P1 and P2 Adapters	No	No	No
□		Size Selection of DNA on E-Gel SizeSelect	No	Yes	No
□		Nick-Translate and Amplify the Library by PCR (12 cycles)	No	Yes	No
□		Quantitate Library - 1	No	Yes	No
□		Perform Exome Partition, 500ng	No	Yes	No
□		Post-hybridization Amplification by PCR (8-12 cycles)	No	Yes	No
□		Quantitate Library - 2	No	Yes	No
□		Library Pass-Fail	No	Yes	No



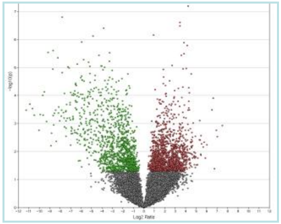


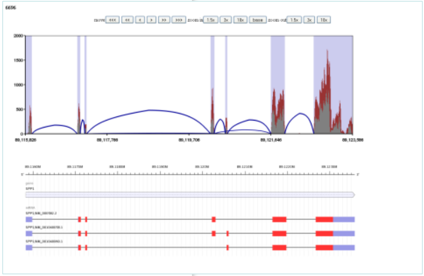
Application-specific Data Analysis

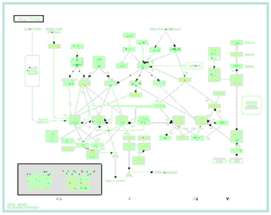
Standard algorithms
Pipelines



Database	Reads In	Matched	% of Total		
				0MM	1MM
Adapter/Primer	11,743,853	428,245	3.65	428,245	0
rRNA, snRNA	11,315,608	118,465	1.01	97,768	18,491
miRBase	11,197,143	707,430	6.02	608,410	98,798
Exon/Intron	10,489,713	1,667,704	14.2	1,309,231	318,691
Intergenic	8,822,009	7,373,383	62.79	5,742,623	1,425,589
Not Mapped	11,743,853	1,448,626	12.34		







Repeat & Compare Data Between Many Samples

Adoption (Sales)

GeneSifter Lab Edition ~200 of Labs – Enterprise $10K's-$100K's deals

Value propositions –

  Core lab directors - business and scientific data production delivery

  ABRF community, presentations, research groups – still go!

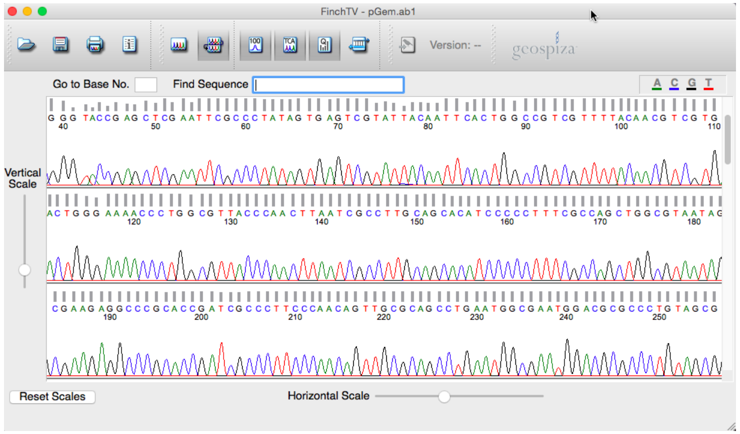
  Biotech (Pharma) scientist – integrated data processing (assembly, BLAST)

GeneSifter Analysis Edition ~500/1000 labs – Researchers, $1K's-$5K's

Value propositions

  Research scientists – ease of use, open source (verifiable) tools, cloud

  Supports common use cases very well – Microarray, RNA-Seq, Exome



```
G G G TACCG AG C T CG AATT CGCCCTATAG TG AGTCGTAT TACAATT CACTGG CCGT CG TTTTACAACG T CG TG
  40        50         60         70         80         90         100         110
ACTGGG AAAACCCTGG CG TTACCCAA CTTAATCGCCTTG CAG CACATCCCCCTTTCGCCAG CTGGCGTAATAG
      120         130         140         150         160         170         180
CGAAG AGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGG ACGCGCCCTGTAGCG
    190         200         210         220         230         240         250
```

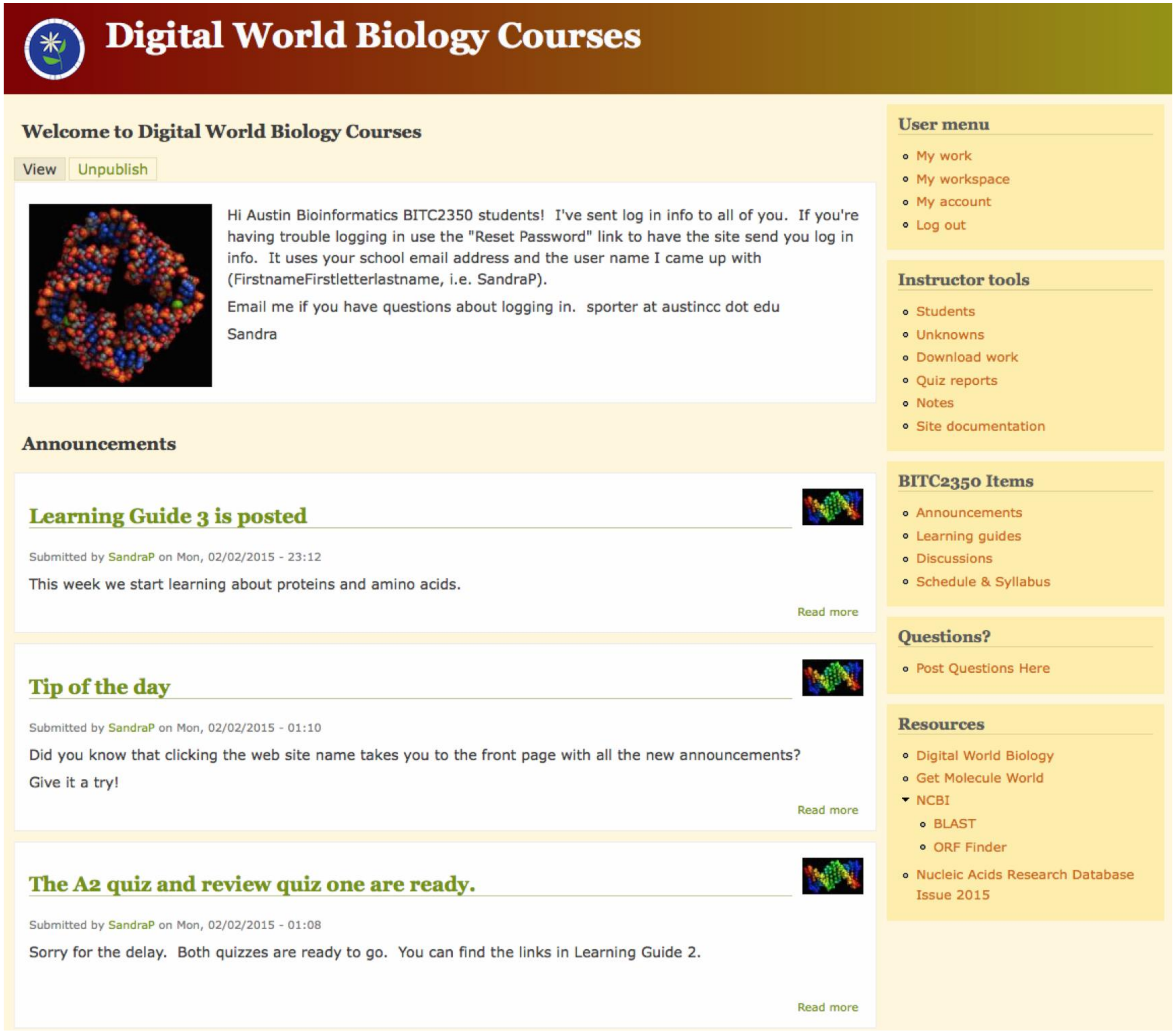
FinchTV >300,000 users

Value propositions

Anyone with a Sanger Sequence file - Full page views, integrated BLAST, drag and drop UI, ease of use

Market builder

Digital World Biology: Education/Consulting



Digital World Biology Courses

Welcome to Digital World Biology Courses

View | Unpublish

Hi Austin Bioinformatics BITC2350 students! I've sent log in info to all of you. If you're having trouble logging in use the "Reset Password" link to have the site send you log in info. It uses your school email address and the user name I came up with (FirstnameFirstletterlastname, i.e. SandraP).

Email me if you have questions about logging in. sporter at austincc dot edu

Sandra

Announcements

Learning Guide 3 is posted

Submitted by SandraP on Mon, 02/02/2015 - 23:12

This week we start learning about proteins and amino acids.

Read more

Tip of the day

Submitted by SandraP on Mon, 02/02/2015 - 01:10

Did you know that clicking the web site name takes you to the front page with all the new announcements?

Give it a try!

Read more

The A2 quiz and review quiz one are ready.

Submitted by SandraP on Mon, 02/02/2015 - 01:08

Sorry for the delay. Both quizzes are ready to go. You can find the links in Learning Guide 2.

Read more

Bioinformatics Education

<u>Software / Databases</u>

Cn3D | Molecule World

Excel

BLAST

Word

ORF finder

Image editing

NCBI resources

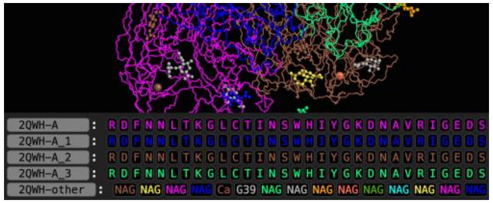
Web browsers!!

15-20 activities

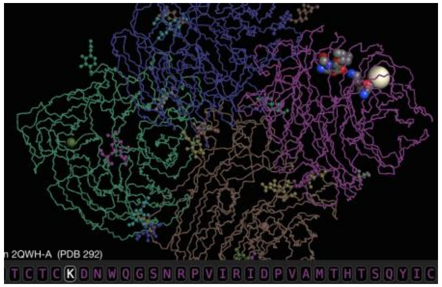
Essential Computer Literacy

1 Sequence

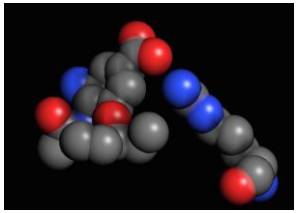
Influenza Virus Neuraminidase


2QWH-A : RDFNNLTKGLCTINSWHIYGKDNAVRIGEDS
2QWH-A_1 :
2QWH-A_2 : RDFNNLTKGLCTINSWHIYGKDNAVRIGEDS
2QWH-A_3 : RDFNNLTKGLCTINSWHIYGKDNAVRIGEDS
2QWH-other : NAG NAG NAG NAG Ca G39 NAG NAG NAG NAG NAG NAG NAG NAG NAG

Structure

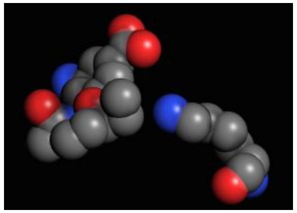


Function



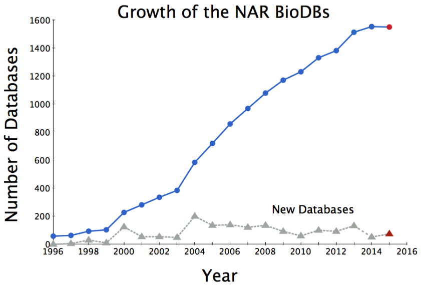
292R>K

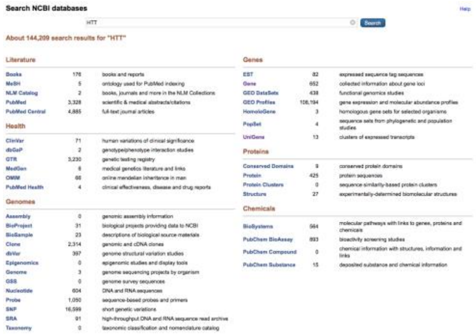
Tamiflu Resistance



2 Understand information resources

Data,
Information,
Knowledge





Search NCBI databases

HTT

About 144,209 search results for "HTT"

Literature

Books	176	books and reports
MeSH	5	ontology used for PubMed indexing
NLM Catalog	2	books, journals and more in the NLM Collections
PubMed	3,328	scientific & medical abstracts/citations
PubMed Central	4,885	full-text journal articles

Health

ClinVar	71	human variations of clinical significance
dbGaP	2	genotype/phenotype interaction studies
GTR	3,230	genetic testing registry
MedGen	6	medical genetics literature and links
OMIM	66	online mendelian inheritance in man
PubMed Health	4	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genomic assembly information
BioProject	31	biological projects providing data to NCBI
BioSample	23	descriptions of biological source materials
Clone	2,314	genomic and cDNA clones
dbVar	397	genome structural variation studies
Epigenomics	0	epigenomic studies and display tools
Genome	3	genome sequencing projects by organism
GSS	0	genome survey sequences
Nucleotide	604	DNA and RNA sequences
Probe	1,050	sequence-based probes and primers
SNP	16,599	short genetic variations
SRA	91	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Genes

EST	82	expressed sequence tag sequences
Gene	652	collected information about gene loci
GEO DataSets	438	functional genomics studies
GEO Profiles	108,194	gene expression and molecular abundance profiles
HomoloGene	3	homologous gene sets for selected organisms
PopSet	4	sequence sets from phylogenetic and population studies
UniGene	13	clusters of expressed transcripts

Proteins

Conserved Domains	9	conserved protein domains
Protein	425	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	27	experimentally-determined biomolecular structures

Chemicals

BioSystems	564	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	893	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	15	deposited substance and chemical information

HTT @ NCBI
144K results over
39 DBs

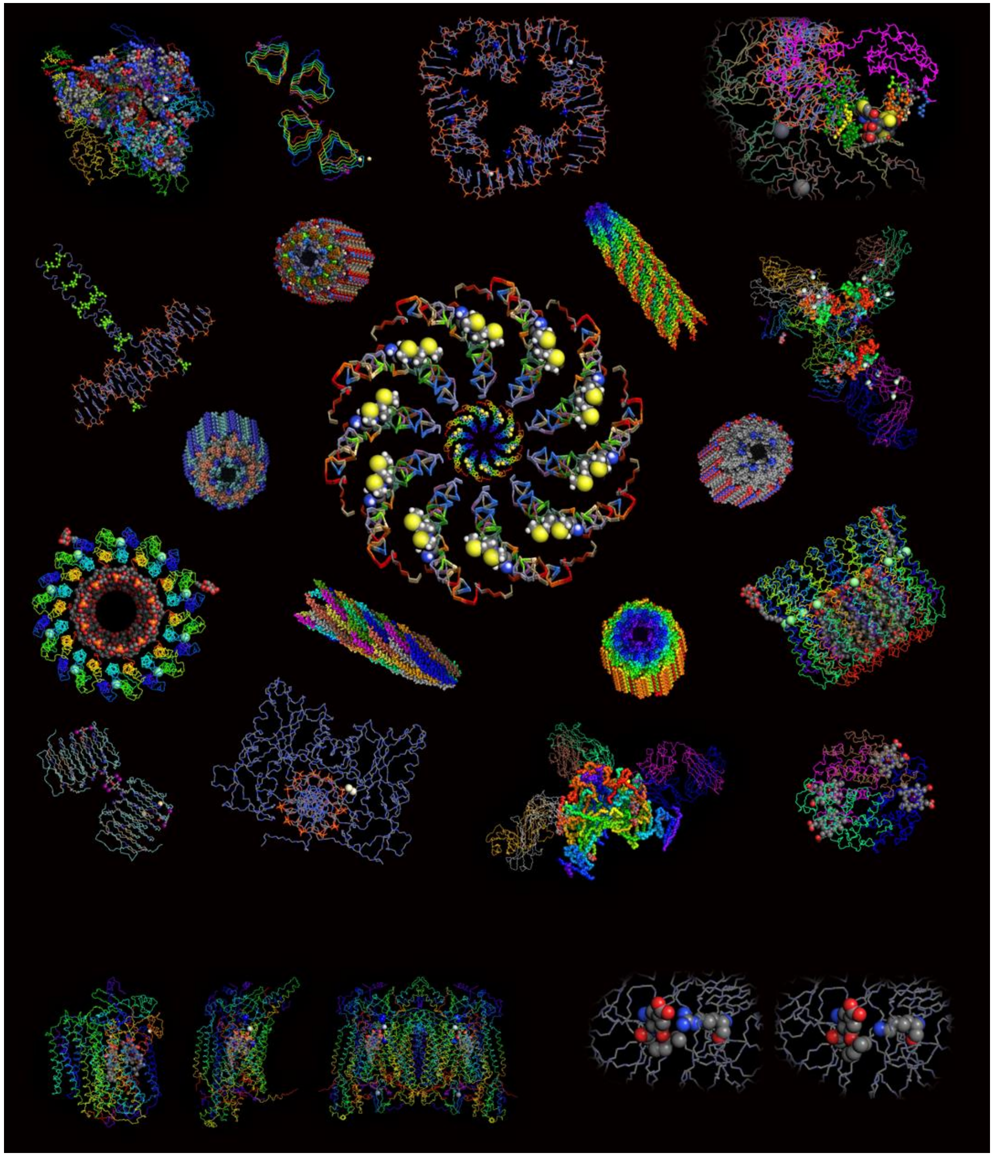
3 Work with data and software

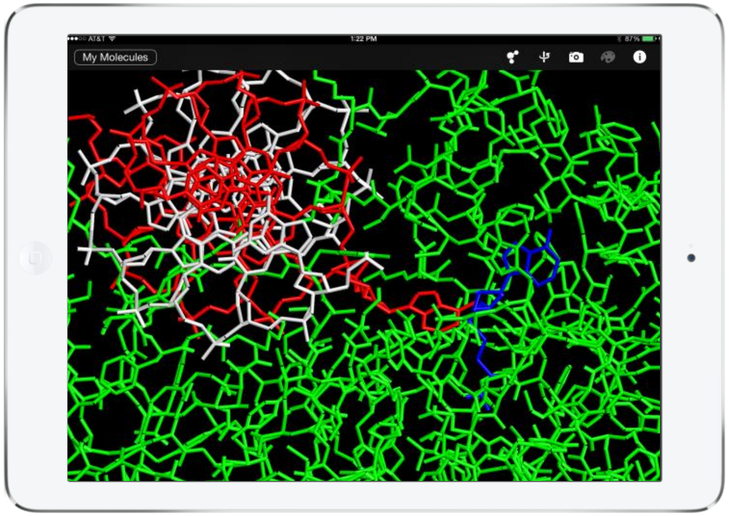
Molecular sequences, data values, structures

Text, graphs, images

Computer programs, software interfaces

Molecule World™





Blending art and science to explore the sequence, structure, function relationship

Lessons

- Implementation is hard
  - Big visions become incremental advances
  - Complex technology, continuous change
  - Data is never analyzed as well/deeply as it could be
- Adoption is harder
  - Informatics, big solutions noise is high
  - Systems and tools are hard to use
  - Every step of the collection, reduction, analysis, discovery phase has 10's/100's of choices

  -----------------------

  - Solutions must be compelling, solve problems
  - Community is critical

GeneLab

Expanding the Impact of Biological Research in Space

Strategic Plan

- Create a biorepository
  - Samples collected in space, over time, experiments
  - Samples collected from terrestrial controls
- Use samples for collecting a variety of “omics” data
- Metadata will be recorded for mining and analysis
- Data and information will be stored in a database
- Analysis will be performed with many tools
- Data and results will be made available for others
- People will be encouraged to utilize the resources to enhance knowledge and discovery

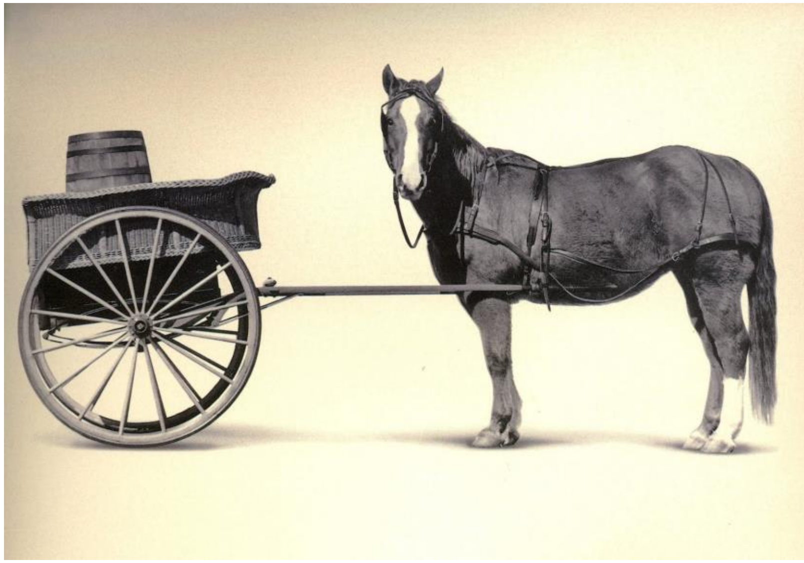
Today’s Questions

- Mission/Vision
  - Value to scientific community?
  - Data or ways to work with the data?

- Implementation
  - Risks of lock in
  - Transitioning data
  - Interfaces
  - Longitudinal data
  - Data integration
  - Metadata (what kinds)
  - Standards
  - Legacy data
  - Experiment scale (research design)
  - …..

All have a Single Answer

- It depends …

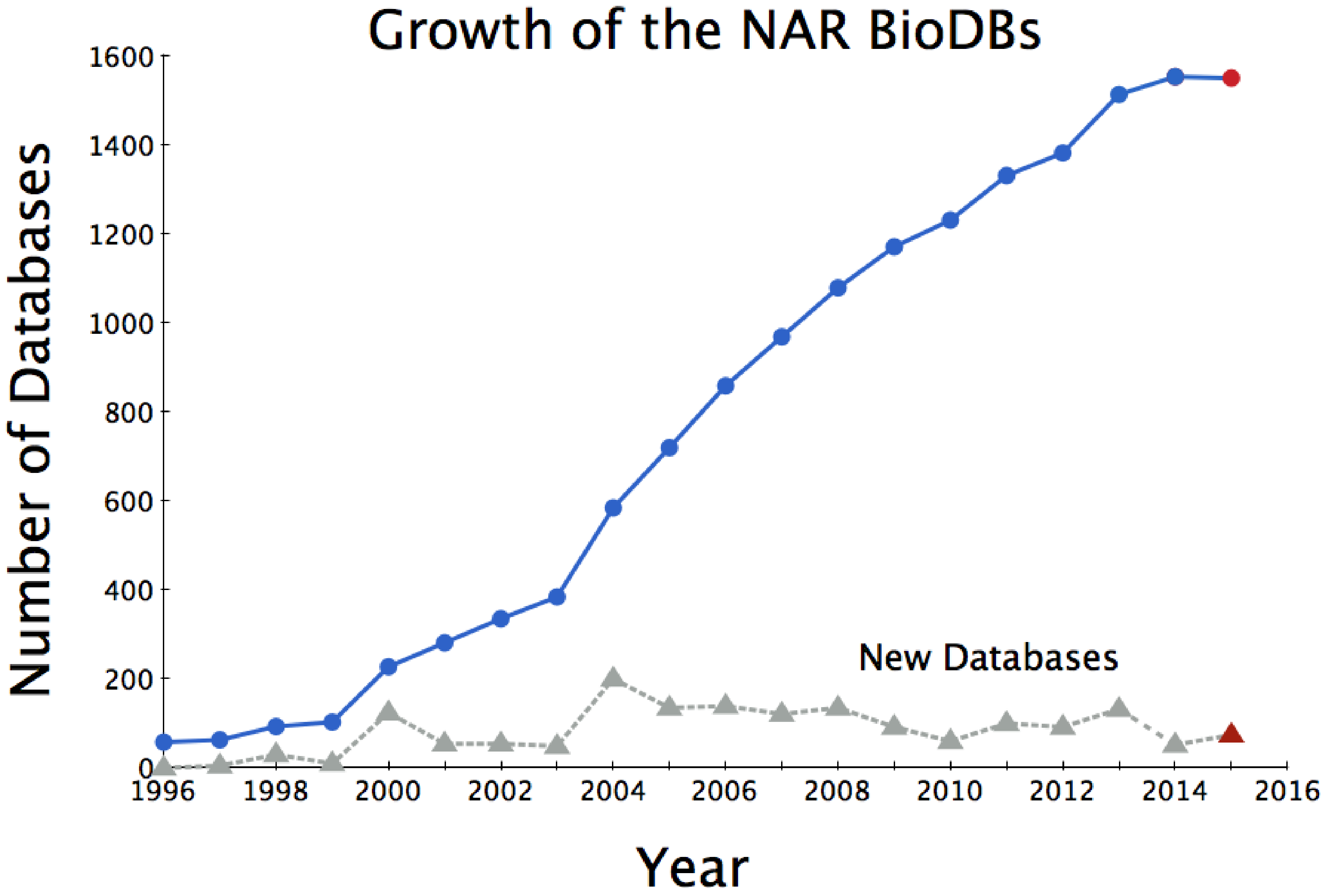


© www.digitalworldbiology.com

Challenges

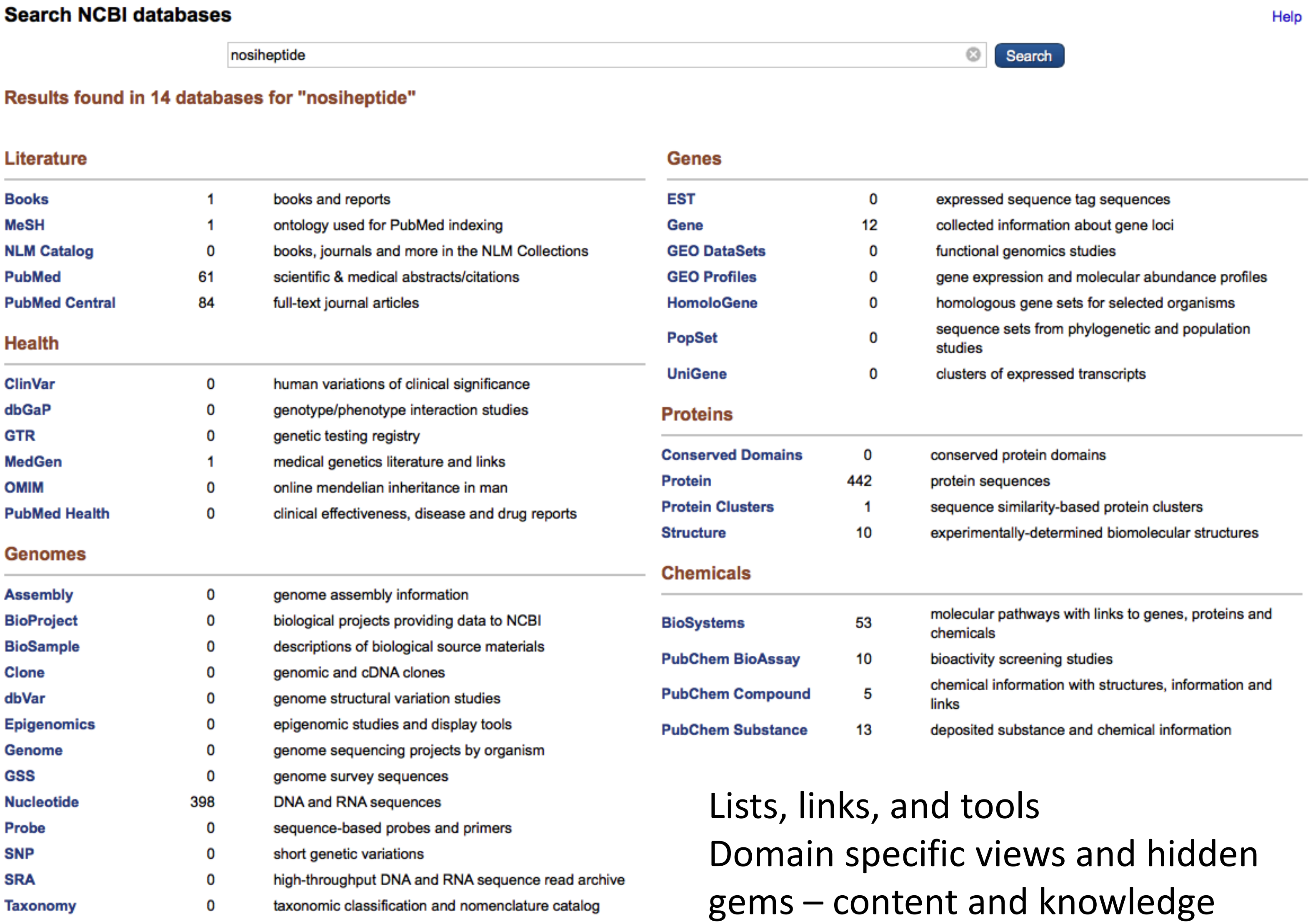
- Clear use cases are lacking
  - What data will be collected?
  - Why is it collected? What is expected?
  - Who is expected to use it?
  - What is the initial market?
- A good example – STEM engagement activities at the K-12 and college levels to engage Americans in the NASA mission, attract and retain students in STEM disciplines, strengthen NASA/Nation's future workforce
- What will compels scientists?

Scientists are Overwhelmed – Why More?



http://scienceblogs.com/digitalbio/2015/01/30/bio-databases-2015/

Databases at NCBI



Search NCBI databases

nosiheptide

Results found in 14 databases for "nosiheptide"

Literature

Books	1	books and reports
MeSH	1	ontology used for PubMed indexing
NLM Catalog	0	books, journals and more in the NLM Collections
PubMed	61	scientific & medical abstracts/citations
PubMed Central	84	full-text journal articles

Health

ClinVar	0	human variations of clinical significance
dbGaP	0	genotype/phenotype interaction studies
GTR	0	genetic testing registry
MedGen	1	medical genetics literature and links
OMIM	0	online mendelian inheritance in man
PubMed Health	0	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioProject	0	biological projects providing data to NCBI
BioSample	0	descriptions of biological source materials
Clone	0	genomic and cDNA clones
dbVar	0	genome structural variation studies
Epigenomics	0	epigenomic studies and display tools
Genome	0	genome sequencing projects by organism
GSS	0	genome survey sequences
Nucleotide	398	DNA and RNA sequences
Probe	0	sequence-based probes and primers
SNP	0	short genetic variations
SRA	0	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Genes

EST	0	expressed sequence tag sequences
Gene	12	collected information about gene loci
GEO DataSets	0	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	0	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and population studies
UniGene	0	clusters of expressed transcripts

Proteins

Conserved Domains	0	conserved protein domains
Protein	442	protein sequences
Protein Clusters	1	sequence similarity-based protein clusters
Structure	10	experimentally-determined biomolecular structures

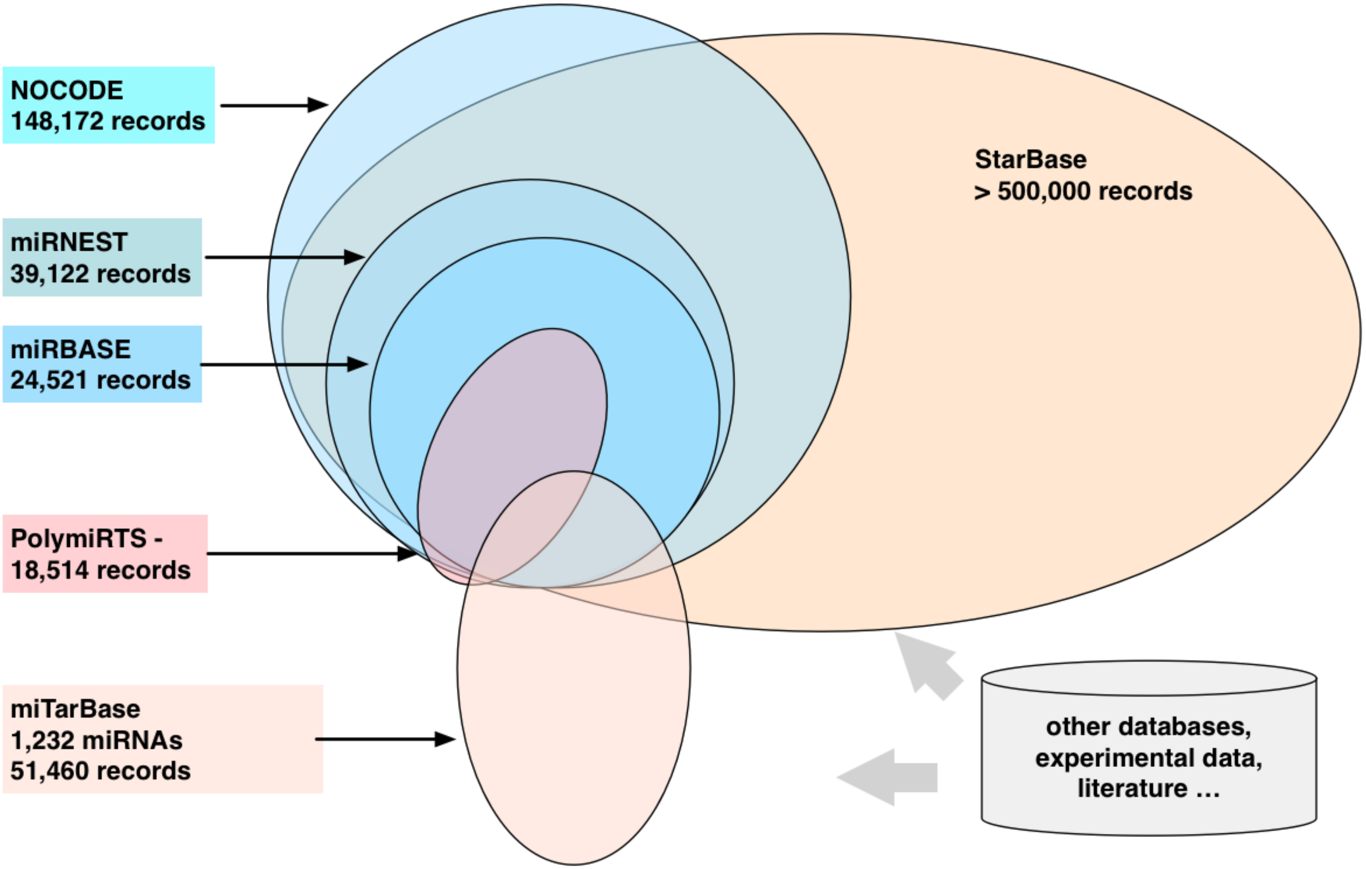
Chemicals

BioSystems	53	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	10	bioactivity screening studies
PubChem Compound	5	chemical information with structures, information and links
PubChem Substance	13	deposited substance and chemical information

Lists, links, and tools
Domain specific views and hidden gems – content and knowledge

Specialized Databases Live in Ecosystems

Relative Content and Conjectured Overlaps



NOCODE
148,172 records

miRNEST
39,122 records

miRBASE
24,521 records

PolymiRTS -
18,514 records

miTarBase
1,232 miRNAs
51,460 records

StarBase
> 500,000 records

other databases,
experimental data,
literature …

Data repositories are indicated by circles and blue tints, and integrative resources indicated by ellipses and orange tint. For each database the number of records is shown and possible (yet unknown) overlaps of information between the databases is suggested using a venn diagram.

Recommendations

- Define applications to drive the requirements
- Stakeholders (people)– users, sponsors, agents … need to be identified and described (business analysis, personas) and engaged
- Determine the initial collaborators, build on examples
- Relevance to the rest of us earthlings – understanding nutrition?