National Research and Education Networks are a platform for data science capacity development in Low- and Middle-Income Countries

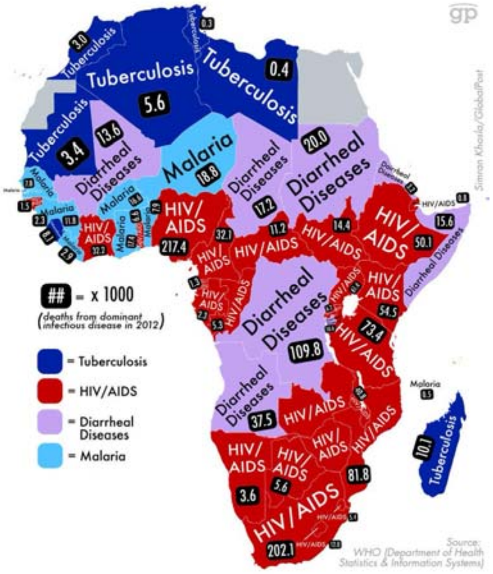
Lloyd Ssentongo[1,2] , Christopher J Whalen [1,3] , **Michael Tartakovsky** [3]

[1] Research Data and Communication Technologies, Inc., Garrett Park, MD, USA
[2] NIH ICER Uganda, Uganda Virus Research Institute, Entebbe, Uganda
[3] Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Lane, Rockville, MD 20852

Two of the most significant obstacles facing the training and development of data scientists and bioinformaticians in low- and improve the capacity of the National Research and Education Networks in this region and others provides and middle-income countries such as those in Africa is the bandwidth and reliability of internet access. The recent movement to expand an opportunity to provide access to these essential tools of education and research. The National Institute of Allergy and Infectious Diseases at the NIH is establishing a public-private partnership with private industry, the Research and Education Network of Uganda (RENU), Makerere University and the Infectious Diseases Institute of Uganda to build the second African Center of Excellence in Bioinformatics in Kampala, Uganda. RENU has built a 1 Gigabit backbone that connects many of the R&E institutions in Uganda. But internet access is still a bottleneck. The ACE partnership and center will provide reference databases and compute infrastructure across the RENU backbone without needing to use internet gateways. The combination of local infrastructure, local connectivity, and local services for data science will improve the educational and analytical capacity of researchers in Uganda and, through the regional NREN Ubuntnet, across East Africa.

Data Science

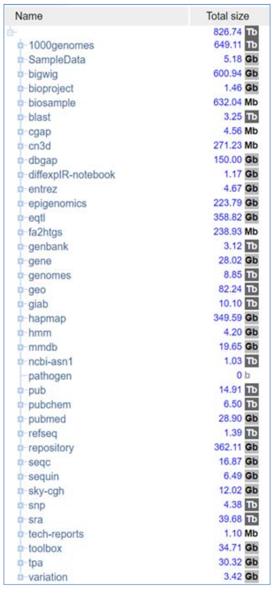
Is a broad field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data. It is geared toward helping individuals and organizations make better decisions from stored, consumed and managed data.



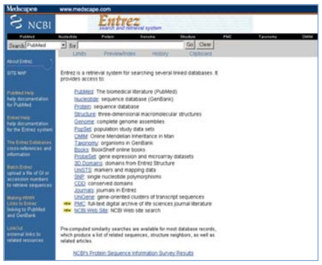
Source: WHO (Department of Health Statistics & Information Systems)

Why is this important?

Data Science Requires Access to Reference Databases



Name	Total size
	826.74 Tb
1000genomes	649.11 Tb
SampleData	5.18 Gb
bigwig	600.94 Gb
bioproject	1.46 Gb
biosample	632.04 Mb
blast	3.25 Tb
cgap	4.56 Mb
cn3d	271.23 Mb
dbgap	150.00 Gb
diffexpIR-notebook	1.17 Gb
entrez	4.67 Gb
epigenomics	223.79 Gb
eqtl	358.82 Gb
fa2htgs	238.93 Mb
genbank	3.12 Tb
gene	28.02 Gb
genomes	8.85 Tb
geo	82.24 Tb
giab	10.10 Tb
hapmap	349.59 Gb
hmm	4.20 Gb
mmdb	19.65 Gb
ncbi-asn1	1.03 Tb
pathogen	0 b
pub	14.91 Tb
pubchem	6.50 Tb
pubmed	28.90 Gb
refseq	1.39 Tb
repository	362.11 Gb
seqc	16.87 Gb
sequin	6.49 Gb
sky-cgh	12.02 Gb
snp	4.38 Tb
sra	39.68 Tb
tech-reports	1.10 Mb
toolbox	34.71 Gb
tpa	30.32 Gb
variation	3.42 Gb


Medscape
www.medscape.com
NCBI
Entrez
search and retrieval system
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
OMIM
Search PubMed
for
Go
Clear
Limits
Preview/Index
History
Clipboard
About Entrez
SITE MAP
PubMed Help
help documentation for PubMed
Entrez Help
help documentation for the Entrez system
The Entrez Databases
cross-references and information
Batch Entrez
upload a file of GI or accession numbers to retrieve sequences
Making WWW Links to Entrez
linking to PubMed and GenBank
LinkOut
external links to related resources

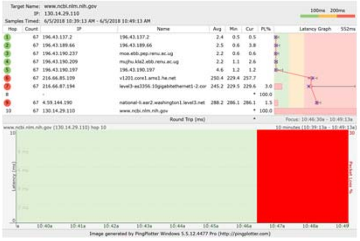
Entrez is a retrieval system for searching several linked databases. It provides access to:

- PubMed: The biomedical literature (PubMed)
- Nucleotide: sequence database (GenBank)
- Protein: sequence database
- Structure: three-dimensional macromolecular structures
- Genome: complete genome assemblies
- PopSet: population study data sets
- OMIM: Online Mendelian Inheritance in Man
- Taxonomy: organisms in GenBank
- Books: BookShelf online books
- ProbeSet: gene expression and microarray datasets
- 3D Domains: domains from Entrez Structure
- UniSTS: markers and mapping data
- SNP: single nucleotide polymorphisms
- CDD: conserved domains
- Journals: journals in Entrez
- UniGene: gene-oriented clusters of transcript sequences
- NEW PMC: full-text digital archive of life sciences journal literature
- NEW NCBI Web Site: NCBI Web site search

Pre-computed similarity searches are available for most database records, which produce a list of related sequences, structure neighbors, as well as related articles.

NCBI's Protein Sequence Information Survey Results

Size can be a problem



Target Name: www.ncbi.nlm.nih.gov
IP: 130.14.29.110
Samples Timed: 6/5/2018 10:39:13 AM - 6/5/2018 10:49:13 AM

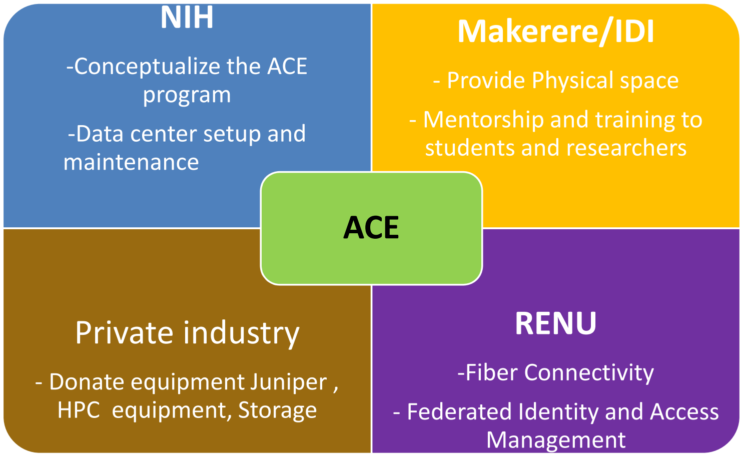
Hop	Count	IP	Name	Avg	Min	Cur	PL%
1	67	196.43.137.2	196.43.137.2	2.4	0.5	0.5	
2	67	196.43.189.66	196.43.189.66	2.5	0.6	3.8	
3	67	196.43.190.237	moa.ebb.pep.renu.ac.ug	2.2	0.6	0.6	
4	67	196.43.190.209	mujhu.kla2.ebb.renu.ac.ug	2.2	1.1	2.6	
5	67	196.43.190.197	196.43.190.197	4.6	1.2	1.2	
6	67	216.66.85.109	v1201.core1.ams1.he.net	250.4	229.4	257.7	
7	67	216.66.87.194	level3-as3356.10gigabitethernet1-2.cor	245.2	229.5	229.6	3.0
8		*				*	100.0
9	67	4.59.144.190	national-li.aar2.washington1.level3.net	288.2	286.1	286.1	1.5
10	67	130.14.29.110	www.ncbi.nlm.nih.gov			*	100.0

Round Trip (ms) *

www.ncbi.nlm.nih.gov (130.14.29.110) hop 10
10 minutes (10:39:13a - 10:49:13a)

Image generated by PingPlotter Windows 5.5.12.4477 Pro (http://pingplotter.com)

Current Access is limited by distance and bandwidth



NIH
-Conceptualize the ACE program
-Data center setup and maintenance

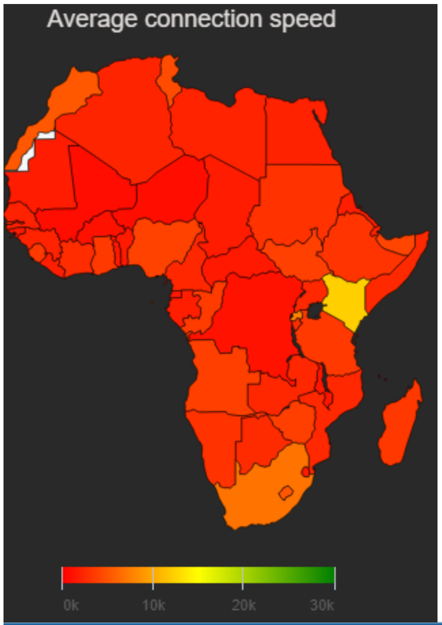
Makerere/IDI
- Provide Physical space
- Mentorship and training to students and researchers

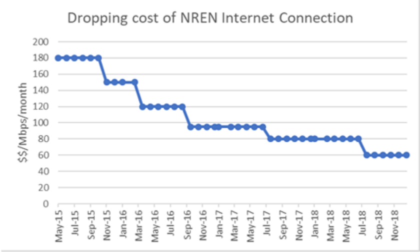
ACE

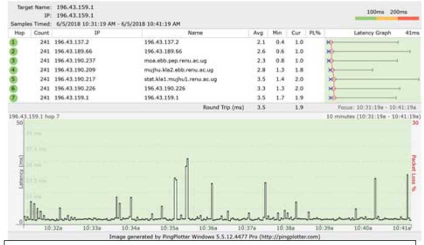
Private industry
- Donate equipment Juniper , HPC equipment, Storage

RENU
-Fiber Connectivity
- Federated Identity and Access Management

The African Centers for Excellence in Bioinformatics demonstrate a way to provision regional research collaborations through the NREN infrastructure







Target Name: 196.43.159.1
IP: 196.43.159.1
Samples Timed: 6/5/2018 10:31:19 AM - 6/5/2018 10:41:19 AM

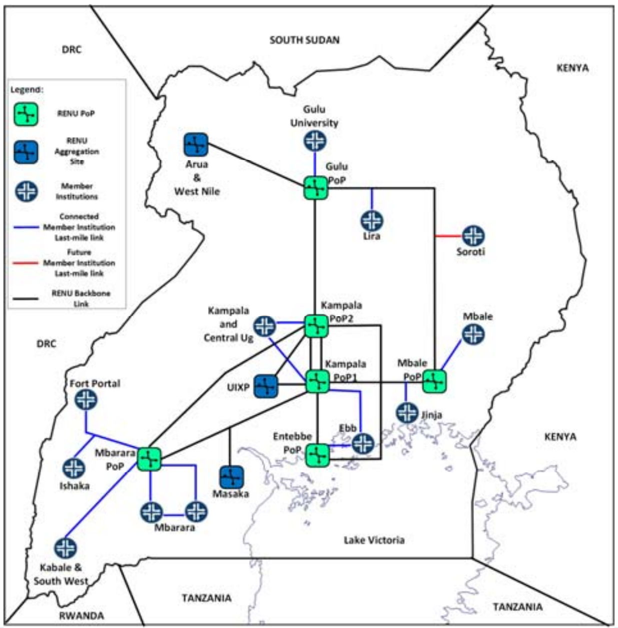
Hop	Count	IP	Name	Avg	Min	Cur	PL%
1	241	196.43.137.2	196.43.137.2	2.1	0.4	1.0	
2	241	196.43.189.66	196.43.189.66	2.6	0.6	1.0	
3	241	196.43.190.237	moa.ebb.pep.renu.ac.ug	2.3	0.8	1.0	
4	241	196.43.190.209	mujhu.kla2.ebb.renu.ac.ug	2.8	1.3	1.8	
5	241	196.43.190.217	stat.kla1.mujhu1.renu.ac.ug	3.5	1.4	2.0	
6	241	196.43.190.226	196.43.190.226	3.3	1.3	2.0	
7	241	196.43.159.1	196.43.159.1	3.5	1.7	1.9	

Round Trip (ms) 3.5 1.9

196.43.159.1 hop 7
10 minutes (10:31:19a - 10:41:19a)

Image generated by PingPlotter Windows 5.5.12.4477 Pro (http://pingplotter.com)

Locating Mirrors or subsets in NRENs will make data science possible in regions with high rates of infection and emerging pathogens



NIAID Office of Cyber Infrastructure and Computational Biology